Procesamiento del Lenguaje Natural, Revista nº 64, marzo de 2020, pp. 61-68 recibido 22-11-2019 revisado 10-01-2020 aceptado 15-01-2020

AzterTest: Open Source Linguistic and Stylistic Analysis Tool

AzterTest: Herramienta de Análisis Lingüístico y Estilístico de Código Abierto

Kepa Bengoetxea, Itziar Gonzalez-Dios, Amaia Aguirregoitia
University of the Basque Country (UPV/EHU)
{kepa.bengoetxea,itziar.gonzalezd,amaia.aguirregoitia}@ehu.eus

Abstract: Text Analysis is a useful process to assist teachers in the selection of the most suitable texts for their students. This task demands the analysis of several text features, which is done mostly manually (e.g. syntactic complexity, words variety, etc.). In this paper, we present an open source tool useful for linguistic and stylistic analysis, called AzterTest. AzterTest calculates 153 features and obtains 90.09 % in accuracy when classifying into three reading levels (elementary, intermediate, and advanced). AzterTest is available also as web tool.
Keywords: text analysis, readability assessment, web application

Resumen: El análisis de texto es un procedimiento útil para ayudar a los profesionales de la educación en la selección de los textos más adecuados para sus alumnos. Esta tarea exige el análisis de varias características de texto (por ejemplo, complejidad sintáctica, variedad de palabras, etc.), que se realiza principalmente de forma manual. En este artículo, presentamos AzterTest, una herramienta de código abierto para el análisis lingüístico y estilístico. AzterTest calcula 153 características y obtiene una exactitud de 90.09 % al distinguir tres niveles de lectura (elemental, intermedio y avanzado). AzterTest también se encuentra disponible como herramienta web.
Palabras clave: análisis de texto, lecturabilidad, aplicación web

1 Introduction

According to the results of PISA 2015 (OECD, 2016), around % 20 of students in the OECD countries do not attain the baseline level of proficiency in reading. Reading is one of the most effective mechanisms in the learning process, since it is our reading capability what let us access and interpret all the available information.

Classifying and adapting reading educational contents manually for students with special needs is an expensive and a hard task for teachers and educators. However, language technologies and Natural Languages Processing (NLP) tools can play a relevant role in the classification and adaptation of available resources, as proved in the works presented in the series of BEA workshops organized by ACL SIGEDU[1]. These technologies have been proved to ease the burden of the educational professionals when selecting texts with a specific level and specific characteristics.

In this paper we present AzterTest, an open-source NLP based tool and web service. AzterTest analyzes 153 linguistic and stylistic features of texts, such as word frequency, sentence length, vocabulary level, argument overlap or use of connective devices. The aim of AzterTest is to provide a detailed linguistic and stylistic analysis of the text to help teachers find the most appropriate reading materials. To evaluate our tool, we test AzterTest in a readability assessment scenario and we compare AzterTest to Coh-Metrix (Graesser, McNamara, and Kulikowich, 2011), a well-known computational tool which analyzes the linguistic and discourse indices. However, its use is not limited to readability assessment, since it can also be used for other purposes such as stylometry or to assess differences among genres or varieties.

This paper is structured as follows: in Section 2 we introduce the related work, in Section 3 we present AzterTest, which we evaluate in Section 4. Later, we describe the web version in Section 5. Finally, we conclude and outline the future work in Section 6.

2 Related work

Traditionally, reading materials have been assessed with conventional readability formulae such as Flesch (Flesch, 1948), Dale-Chall

[1] `https://sig-edu.org/`

ISSN 1135-5948. DOI 10.26342/2020-64-7

(Chall and Dale, 1995), the indexes Gunning FOG (Gunning, 1968) or Simple Measure Of Gobbledygook (SMOG) grade (Mc Laughlin, 1969). In general, these formulae are based on raw features such as word and sentence length, vocabulary lists and frequencies and give a score to classify texts. NLP based tools have proved that these formulae are not reliable when assessing the levels of the texts (Si and Callan, 2001; Petersen and Ostendorf, 2009; Feng et al., 2010). Moreover, the information offered by these traditional formulae is insufficient, since they do not detect slight changes in aspects such as coherence and cohesion of the texts (Graesser, McNamara, and Kulikowich, 2011).

Computational tools for linguistic analysis, generally, focus on the quantitative dimension of text complexity, where features related to quantitative aspects of the texts (word length, frequency, incidence of grammar structures, etc.) are used to assess linguistic complexity. Concerning research for English, Coh-Metrix 3.0 (Graesser, McNamara, and Kulikowich, 2011) analyzes texts providing 110 measures in its free version, which are classified in 11 groups: descriptive, text easability principal components scores, referential cohesion, latent semantic analysis, lexical diversity, connectives, situation model, syntactic complexity, syntactic pattern density, word information and readability. Coh-Metrix[2] has also been partially adapted to Brazilian Portuguese (Scarton and Alusio, 2010) and Spanish (Quispersaravia et al., 2016). In the case of Spanish, the tool *El Manchador de Textos* analyses some linguistic features (Venegas, 2008).

These tools are mainly used for Readability assessment, which aims to grade the ease or the difficulty of written texts. Most of research in this line has focused on classifying texts as simple or complex e.g. for English (Feng et al., 2010), Italian (Dell'Orletta, Montemagni, and Venturi, 2011), German (Hancke, Vajjala, and Meurers, 2012), Spanish (Štajner and Saggion, 2013), or Basque (Gonzalez-Dios et al., 2014). However, only a few of them have also tried to assess three levels e.g. for Brazilian Portuguese (Aluísio et al., 2010), the CEFR levels for French (François and Fairon, 2012) or have developed a multilingual proposal (Madrazo and Pera, 2019). Besides, getting an open licensed corpora and data is the main problem when training these systems.

3 AzterTest: Tools and Features

In this section, we describe AzterTest which is freely available from a public GitHub repository[3] and is licensed under GNU General Public License v3.0. First of all, we outline the resources and tools used in the implementation process and later, we introduce the features computed by the tool.

3.1 Preprocessing tools and resources

AzterTest uses NLP-Cube (0.1.0.7) (Boroș, Dumitrescu, and Burtica, 2018) tool for automatic analysis, which is the state-of-the-art in segmentation (tokenization and sentence-splitting), lemmatization, POS tagging and dependency parsing task for over 50 languages. NLP-Cube was one of the best systems in English on CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies (Zeman and Hajič, 2018). We have also tested StanfordNLP (0.2.0) (Qi et al., 2019) because AzterTest is based on Universal Dependencies (UD) and it allows to use others UD parsers. Hence, AzterTest can be easily adapted to other languages in the near future.

Moreover, for the analysis of English text, we have implemented a syllable splitter based on CMUdict (Carnegie Mellon University Pronouncing Dictionary) (Weide, 2005). This splitter is used to count the number of syllables in a text.

To obtain lexical information we have used the stop words from the NLTK stopwords Corpus, the list of irregular verbs which are freely available[4], the Dale-Chall word list[5], a list of the 3,000 core words that every learner of English needs to know at A1-B2 level and an additional 2,000 word list including the most useful high-level words guiding advanced learners at B2-C1 level to learn to expand their vocabulary[6].

For word frequency, we have used wordfreq (Speer et al., 2018), which provides access to estimates of how often a word is used in 36 languages. wordfreq detects the word frequency

[2]The versions for Brazilian Portuguese and Spanish are based on Coh-Metrix 2.0 and the authors adapted around 40 Coh-Metrix indices related to cohesion, coherence and the difficulty of text comprehension, according to specific characteristics of each language.

[3]https://github.com/kepaxabier/AzterTest

[4]https://github.com/Bryan-Legend/babel-lang/blob/master/Babel.EnglishEmitter/Resources/Irregular

[5]http://www.readabilityformulas.com/articles/dale-chall-readability-word-list.php

[6]https://www.oxfordlearnersdictionaries.com/wordlists/oxford3000-5000

of a word as the logarithm in base 10 of the number of times a word appears per one billion words. A word rated as 3 appears 10^3 times for every 10^9 words, that is, once per million words. Using wordfreq and after testing different values, for our educational purposes, we have decided to consider words with a value below 4 as rare words.

For semantic information, we have used WordNet (Miller, 1995), which groups nouns, verbs, adjectives and adverbs into sets of cognitive synonyms (synsets), each expressing a distinct concept. Moreover, synsets are interlinked by means of conceptual-semantic and lexical relations.

For semantic similarity, we have used the Universal Sentence Encoder. The Universal Sentence Encoder encodes text into high dimensional vectors that can be used for text semantic similarity. The pre-trained Google's Universal Sentence Encoder (Cer et al., 2018) is publicly available in Tensorflow-hub[7].

3.2 Linguistic and Stylistic Features

Linguistic features are those related to morphology, syntax and semantics while the stylistic features are related to cohesion, vocabulary knowledge etc. In order to decide which features to implement in AzterTest, we have analyzed in detail the features provided by the works presented in Section 2. After a deep analysis, we have implemented a set of 153 AzterTest features. Following, we present the list of the features included in AzterTest organized by type:

- **Descriptive:** numbers and incidences of paragraphs, sentences, words, distinct words, words with punctuation; sentences per paragraph, words per sentence with and without stopwords; syllables per word, letters per lemma, and letters per word with and without stopwords.

- **Classical Readability formulae:** Flesch-Kincaid grade level, Flesch readability ease, Dale-Chall and SMOG.

- **Lexical Diversity:** lexical density, densities of nouns, verbs, adjectives and adverbs, simple type-token ratio, content type-token ratio, type-token ratio of nouns, verbs, adjectives and adverbs, lemma simple type-token ratio, lemma content type-token ratio, lemma noun, verb, adjective, and adverb type-token ratio, Honoré and Maas lexical density measures and Measure of Textual Lexical Diversity (MTLD).

- **Word Frequency:** words with a value below 4 in wordfreq as rare words.

- **Vocabulary knowledge:** numbers and incidences of A1, A2, B1, B2 and C1 level vocabulary in the text, and number and incidence of content words not in A1-C1 vocabulary.

- **Word Semantic information:** the average values of polysemy of words, hypernym values of verbs, hypernym values of nouns, hypernym values of nouns and verbs.

- **Word Morphological information:** incidences and numbers of nouns, adjectives, adverbs, pronouns and each type of pronoun (first, second and third person; singular and plural), verbs and all variations for verbs (tense, mood, regularity, etc).

- **Syntactic Complexity:** left embeddedness (mean of number of words before the main verb), means of the number of propositions, levels of dependency tree, subordinate clauses and relative subordinate clauses, verbs in gerund form, verbs in infinitive form; descendants and modifiers per noun phrase; verb and noun phrases per sentence, and punctuation marks per sentence.

- **Syntactic Pattern Density:** noun and verbal phrase density, passive voice agentless passive voice verbs density, negation density, infinitive and gerund form density.

- **Referential Cohesion:** noun, argument, stem and content word overlap, mean and standard deviation of semantic similarity.

- **Connectives (logical cohesion):** incidences of causal, logical, adversative/contrastive, temporal and conditional connectives.

The main contributions of this work to the field of feature analysis are the ones concerning word frequency and vocabulary knowledge. For a detailed explanation of the rationale behind each of these metrics we refer the interested reader to the AzterTest documentation[8].

[7] https://tfhub.dev/

[8] http://178.128.198.190/information.html

4 *Extrinsic Evaluation: Readability Assessment of English Texts*

In this Section we present an extrinsic evaluation of AzterTest in a readability assessment scenario for English texts. In this evaluation, we have tested various classifiers to detect three reading levels (elementary, intermediate, advanced) based on Coh-Metrix and AzterTest's output on an open licensed corpora. We compare our results to other systems, perform an error analysis and discuss the best features.

4.1 Corpus

In order to train and validate AzterTest, we have used the corpus OneStopEnglish corpus (Vajjala and Lucic, 2018). This corpus compiles newspaper articles aligned at text and sentence level across three reading levels (elementary, intermediate, advanced), targeting English as Second Language (ESL) learners. The corpus consists of 189 texts, each of them in three versions (567 in total). We have decided to use this corpus because it is one of the few available[9] and it is licensed under CC BY-SA 4.0. Moreover, this corpus demonstrates its usefulness for automatic readability assessment among others. Namely, Vajjala and Lucic (2018) obtained an accuracy of 78.13 % using features based on readability classification research with the Sequential Minimal Optimization (SMO) (Platt, 1998) classifier.

For our experimental purposes we have randomly divided the corpus (in total 567 texts) into 2 non-overlapping datasets: 456 texts (152 texts for each class) as the training set and 111 texts (37 texts for each class) as the test set.

4.2 Classifying Experiments and Results

In order to classify the texts according to their complexity level, we have trained several classifiers that are included in WEKA (Hall et al., 2009). To evaluate the classifiers we have used the 10-fold cross-validation and the test set.

In these experiments we have tested three tools: Coh-Metrix and two configurations of AzterTest. In the first configuration of AzterTest we have taken into account all the features (absolute numbers and ratios) while in the second, AzterTest-ratios, we have only selected features based on incidents, means or typical deviations. Regarding the features, first, we have tested all the Coh-Metrix, AzterTest and AzterTest-ratios features.

[9] https://zenodo.org/record/1219041

Secondly, in order to detect the best features to tag and automatically remove the noise ones, we have tested different sets of attributes (25, 50, 75 and 100). In this experiment, we have tested chi square using different sets of attributes: 25, 50, 75 and 100. We have used Flesch readability ease as baseline.

In Table 1 we present the accuracy of the classifiers (Class. column) for each tool [Baseline, Coh-Metrix (Coh), AzterTest (Azt) and AzterTest-ratios (Azt-r)] and using different features (Feat. column). For brevity, we only show the classifiers [*i*) Sequential Minimal Optimization (SMO) (Platt, 1998) and *ii*) Simple Logistic (SL) (Landwehr, Hall, and Frank, 2005)] and the feature sets (all, 50 and 25) that have obtained the best results. We have tested the classifiers with the defaults hyperparameters.

Tool	Class.	Feat.	Data	Accu.
Baseline	SMO	Flesch	Cross Test	49.78 54.95
Coh	SL	All	Cross Test	77.19 81.98
Coh	SL	50	Cross Test	**77.85** **81.98**
Coh	SL	25	Cross Test	75.65 85.58
Azt	SMO	All	Cross Test	82.01 84.68
Azt	SMO	50	Cross Test	**82.01** **90.09**
Azt	SMO	25	Cross Test	82.01 88.28
Azt-r	SMO	All	Cross Test	80.92 84.68
Azt-r	SMO	50	Cross Test	80.04 85.58
Azt-r	SMO	25	Cross Test	**81.35** **84.68**

Table 1: Classification results of the three level readability assessment experiment

Respecting the classification results, the best classifier for Coh-Metrix is Simple Logistic, while it is SMO for AzterTest's two configurations and the baseline. The best results are obtained with the 50 most predictive features in Coh-Metrix and AzterTest, but with 25 in AzterTest-ratios. Moreover, all the results are lower when evaluating with 10 fold cross-validation.

In sum, looking at these results, the best model is the SMO classifier with 50 features of

AzterTest, namely 90.09 %. It is 4.16 points better than the best Coh-Metrix results using the cross-validation and 8.11 points using the test set. AzterTest is also better than the baseline in 32.23 points with the cross-validation and in 35.14 points with the test set. Therefore, in this scenario, AzterTest outperforms Coh-Metrix, the classical readability formula used as baseline and the results reported by Vajjala and Lucic (2018). However, AzterTest-ratios is not far from AzterTest, and outperforms Coh-Metrix when evaluating with cross-validation.

In addition to the all and selected features, we have also trained the classifiers with each type of linguistic/stylistic features that we described in section 3.2. In Table 2 we rank these results by type.

Feature Group	**Data**	**Accu.**
Syntactic	Cross	70.61
	Test	75.67
Lexical Density	Cross	65.13
	Test	72.07
Descriptive	Cross	65.13
	Test	67.56
Word Frequency	Cross	60.96
	Test	68.46
Vocabulary	Cross	55.92
	Test	60.36
Readability	Cross	53.50
	Test	59.45
Word Morphological	Cross	50.21
	Test	55.85
Word Semantic	Cross	50.21
	Test	49.54
Referential Cohesion	Cross	46.05
	Test	44.14
Discourse Connectives	Cross	38.59
	Test	34.23

Table 2: The results of the SMO classifier with specific linguistic features (only AzterTest's ratios)

Taking into account the different feature types, the syntactic features (complexity and pattern density together) performed best with an accuracy of 70.61 %; lexical features and descriptive features (65.13 %) performed almost equally well; word frequencies performed worse than lexical in cross-validation but similarly in the test and, finally, the accuracy of referential cohesion and discourse connectives is below 50 %.

Finally, we present the F measure for each text level of the best model. The F measure is 0.917 for the elementary level, 0.857 for the intermediate and 0.932 for the advanced. Comparing these results to the work for Brazilian Portuguese (Aluísio et al., 2010) that also classified into three levels, rudimentary (F measure 0.732), basic (F measure 0.483) and advanced level (F measure 0.913), we observe that our model is stronger across classes.

4.3 Error analysis

We have also carried out an error analysis, using the output of the best system, which is the SMO classifier incorporating the best 50 characteristics of AzterTest. We have checked manually the annotation results in the test. The test set comprised 111 instances and only 11 of them are errors. 3 instances have been erroneously classified as "intermediate" out of 37 "advanced" type instances. For the 37 "elementary" type, only 4 have been predicted to be "intermediate" and finally, concerning the 37 "intermediate" instances, the system has classified 2 of them as "advanced" and another 2 as "elementary". Under no circumstances has the system predicted an "advanced" instance to be "elementary" or vice versa.

4.4 Discussion: Best Feature Selection and Corpus Analysis

Using both absolute and ratio features results in a higher score (Table 1). However, we have decided to exclude absolute numbers for stylistic analysis of AzterTest, since they may hinder other linguistic and stylistic features. For example, the raw number of words is usually a predictive feature, but it depends on text length and not on its linguistic characteristics. That is, a short text can be simple or complex even though it is short. Measuring features independently of text length allows the user to compare different texts.

Following, in Table 3 we present the the linguistic/stylistic analysis of the corpus, where we show the average values of the 25 most predictive ratios. The abbreviations we use are: n.=number, i.=incidence (per 1000 words), m.=mean, s.=standard deviation, sent.=sentence, prop.=proposition, sub.=subordinate and w.=word.

We observe in this corpus, which compiles journalist texts adapted for ESL readers, that the key features to discriminate among

Feature	A	I	E
Word Frequency			
i. of rare verbs	18.38	11.63	7.57
m. of distinct rare content w.	18.22	13.99	10.73
m. of rare content w.	15.36	12.16	9.76
i. of rare adj.	13.22	10.09	6.51
Descriptive			
s. of w. per sent.	10.66	8.95	7.46
s. of w. per sent. without stop w.	7.47	6.34	5.31
m. of w. per sent.	21.11	18.83	16.14
i. of n. of sent.	48.50	53.90	62.34
m. of w. per sent. without stop w.	14.61	13.00	11.19
s. of letters in w.	2.55	2.49	2.34
Vocabulary			
i. of B2	32.66	27.39	18.42
i. of C1	10.97	7.37	4.30
Lexical Diversity			
Honoré	984.93	896.14	779.31
Maas	0.0506	0.0546	0.0621
MTLD	119.32	106.98	90.09
Syntactic Complexity			
m. of prop. per sent.	52.22	41.65	31.37
m. NP per sent.	6.82	6.18	5.42
m. of punc. per sent.	2.58	2.33	2.04
m. VP per sent.	3.18	2.88	2.55
m. depth per sent.	5.79	5.51	5.11
Syntactic Pattern Density			
i. gerund density	16.42	12.84	7.94
Readability			
Flesch-Kincaid	11.55	10.27	8.59
Flesch Ease	51.54	56.28	63.43
SMOG	8.64	8.04	7.07
Word Semantic			
m. hypernym of verbs	2.09	1.97	1.83

Table 3: Corpus analysis with the 25 best predictive ratios of AzterTest

the three linguistic levels are, a) concerning word frequency, distinct rare content words, particularly verbs and adjectives; b) regarding descriptive features, words per sentence with and without stopwords and letters per word; c) at vocabulary level, incidence of B2 and C1 words; d) and lexical diversity, Honoré, Maas measures and MTLD; e) at syntactic level, propositions, NPs, VPs and punctuation marks per sentence and sentence depth; f) regarding the classical readability formulae, Flesch-Kincaid, Flesch Ease and SMOG; and, finally, g) at semantic the level,the hypermyn verbs index.

All the values decrease from advanced to elementary level, except for the incidence of number of sentences, MAAS lexical density and Flesch Ease. In the case of the MAAS and Flesch Ease, higher scores indicate that the texts are simpler, which correlates to the rest of the features. The higher number of sentences can be explained because simpler texts have shorter sentences and less clauses, and therefore, more sentences are required to communicate the information in the texts.

5 *AzterTest: Web Tool and Source*

Additionally, AzterTest is a web tool that computes 153 features of the linguistic and discourse representations of a text, including descriptive, lexical diversity, readability, word morphological information, word frequency, vocabulary knowledge, syntactic complexity, syntactic pattern density, word semantic information, referential cohesion and connectives. Furthermore, AzterTest web tool classifies the text under three language levels of difficulty (elementary, intermediate and advanced). The tool can be tested in the following website http://178.128.198.190. In Figures 1 and 2 we show the home page of AzterTest and an excerpt of its analysis respectively.

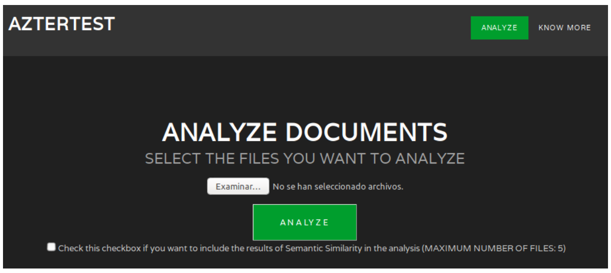

Figure 1: Main Page of AzterTest Web Tool

AzterTest source is implemented in Python and it is freely available from a public GitHub repository https://github.com/kepaxabier/AzterTest.

6 *Conclusions and future work*

In this paper, we have introduced AzterTest, an open source linguistic and stylistic analysis tool. AzterTest computes and takes into account 153

File: Thanksgiving.doc	
Level of difficulty	Elementary
Shallow or descriptive measures	
Number of words (total)	56
Number of distinct words (total)	40
Number of words with punctuation (total)	66
Number of paragraphs (total)	6
Number of paragraphs (incidence per 1000 words)	107.1429
Number of sentences (total)	6

Figure 2: Screenshot of AzterTest Result (Readability and Descriptive Features)

features, which are grouped into descriptive incidences, word frequencies, vocabulary knowledge, lexical diversity, morphological information, syntactic phenomena, classical readability formulae, semantic information and cohesion devices. The main contributions concerning features are related to new vocabulary and frequency.

Moreover, we have tested AzterTest in a readability assessment scenario for English texts, and using a set of 50 features and the classifier SMO we have obtained an accuracy of 90.09 %. This model outperforms the results obtained with Coh-Metrix's output in this task.

Furthermore, we have made the web application available to teachers so that they can assess the linguistic, stylistic and readability characteristics of their reading materials.

Considering that AzterTest is based on universal dependency parsers, in the future, we will adapt it for multiple languages, and we also plan to extend it with additional vocabulary related features. Additionally, we intend to perform a more extensive assessment of the tool with a group of potential users in order to gather information and adapt AzterTest at their suggestions. Finally, we also plan to use AzterTest for other textual analysis across genres and domains or specialised discourse (Parodi, 2006).

Acknowledgments

We acknowledge following projects: DL4NLP (KK-2019/00045), DeepReading RTI2018-096846-B-C21 (MCIU/AEI/FEDER, UE) and BigKnowledge for Text Mining, BBVA.

References

Aluísio, S., L. Specia, C. Gasperin, and C. Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9. ACL.

Boroș, T., S. D. Dumitrescu, and R. Burtica. 2018. Nlp-cube: End-to-end raw text processing with neural networks. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 171–179.

Cer, D., Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Chall, J. S. and E. Dale. 1995. *Readability Revisited: The New Dale–Chall Readability Formula*. Brookline Books, Cambridge, MA.

Dell'Orletta, F., S. Montemagni, and G. Venturi. 2011. READ-IT: assessing readability of Italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, SLPAT '11, pages 73–83. ACL.

Feng, L., M. Jansche, M. Huenerfauth, and N. Elhadad. 2010. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 276–284. ACL.

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

François, T. and C. Fairon. 2012. An AI readability formula for French as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477. ACL.

Gonzalez-Dios, I., M. J. Aranzabe, A. Díaz de Ilarraza, and H. Salaberri. 2014. Simple or complex? assessing the readability of basque texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 334–344, Dublin, Ireland, August. DCU and ACL.

Graesser, A. C., D. S. McNamara, and J. M. Kulikowich. 2011. Coh-Metrix Providing Multilevel Analyses of Text Characteristics. *Educational Researcher*, 40(5):223–234.

Gunning, R. 1968. *The technique of clear writing*. McGraw-Hill New York.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The WEKA data mining software: an update.

ACM SIGKDD Explorations Newsletter, 11(1):10–18.

Hancke, J., S. Vajjala, and D. Meurers. 2012. Readability Classification for German using lexical, syntactic, and morphological features. In *COLING 2012: Technical Papers*, page 1063–1080.

Landwehr, N., M. Hall, and E. Frank. 2005. Logistic model trees. 95(1-2):161–205.

Madrazo, I. and M. S. Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

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

Miller, G. A. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

OECD. 2016. *PISA 2015. Results in Focus.* OECD Publishing.

Parodi, G. 2006. Discurso especializado y lengua escrita: Foco y variación. *Estudios filológicos*, (41):165–204.

Petersen, S. E. and M. Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer Speech & Language*, 23(1):89–106.

Platt, J. 1998. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical Report MSR-TR-98-14.

Qi, P., T. Dozat, Y. Zhang, and C. D. Manning. 2019. Universal dependency parsing from scratch. *arXiv preprint arXiv:1901.10457.*

Quispersaravia, A., W. Perez, M. A. S. Cabezudo, and F. Alva-Manchengo. 2016. Coh-Metrix-Esp: A Complexity Analysis Tool for Documents Written in Spanish. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4694–4698.

Scarton, C. and S. M. Alusio. 2010. Coh-metrix-port: a readability assessment tool for texts in brazilian portuguese. In *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language, Extended Activities Proceedings, PROPOR*, volume 10. sn.

Si, L. and J. Callan. 2001. A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 574–576. ACM.

Speer, R., J. Chin, A. Lin, S. Jewett, and L. Nathan. 2018. Luminosoinsight/wordfreq: v2.2, October.

Vajjala, S. and I. Lucic. 2018. Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification.

Venegas, R. 2008. Interfaz computacional de apoyo al análisis textual:"el manchador de textos". *RLA. Revista de lingüística teórica y aplicada*, 46(2):53–79.

Štajner, S. and H. Saggion. 2013. Readability Indices for Automatic Evaluation of Text Simplification Systems: A Feasibility Study for Spanish. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 374–382, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Weide, R. 2005. The carnegie mellon pronouncing dictionary [cmudict. 0.6].

Zeman, D. and J. Hajič, editors. 2018. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies.* ACL, Brussels, Belgium, October.