A Benchmark for Neural Readability Assessment of Texts in Spanish

Laura Vásquez-Rodríguez[1], **Pedro-Manuel Cuenca-Jiménez**[2,3],
Sergio Esteban Morales-Esquivel[4], **Fernando Alva-Manchego**[5]
[1]National Centre for Text Mining, The University of Manchester, UK
[2]Universidad Rey Juan Carlos, Fuenlabrada, Madrid, Spain
[3]Hugging Face SAS, Paris, France
[4]School of Software Enigeering, Universidad Cenfotec, San José, Costa Rica
[5]School of Computer Science and Informatics, Cardiff University, UK
laura.vasquezrodriguez@manchester.ac.uk, pedro.cuenca@urjc.es
smorales@ucenfotec.ac.cr, alvamanchegof@cardiff.ac.uk

Abstract

We release a new benchmark for Automated Readability Assessment (ARA) of texts in Spanish. We combined existing corpora with suitable texts collected from the Web, thus creating the largest available dataset for ARA of Spanish texts. All data was pre-processed and categorised to allow experimenting with ARA models that make predictions at two (simple and complex) or three (basic, intermediate, and advanced) readability levels, and at two text granularities (paragraphs and sentences). An analysis based on readability indices shows that our proposed datasets groupings are suitable for their designated readability level. We use our benchmark to train neural ARA models based on BERT in zero-shot, few-shot, and cross-lingual settings. Results show that either a monolingual or multilingual pre-trained model can achieve good results when fine-tuned in language-specific data. In addition, all models decrease their performance when predicting three classes instead of two, showing opportunities for the development of better ARA models for Spanish with existing resources.

1 Introduction

The readability of a text refers to the aggregation of all its elements that affect the reader's understanding, reading speed, and interest in the content (Dale and Chall, 1949). Some of these elements are the words in the text, the grammatical structure of its sentences, and its writing style (Xia et al., 2016). For example, a newspaper article may be more readable than a scientific paper or a novel. Automated Readability Assessment (ARA) aims to exploit these textual elements to predict how "difficult" or comprehensible a text is (Collins-Thompson, 2014). For texts in English, several techniques have been developed, ranging from formulae that relies on surface characteristics such as average word and sentence lengths (Gunning et al., 1952; Kincaid et al., 1975) to machine learning approaches based on feature engineering (François and Miltsakaki, 2012; Vajjala and Meurers, 2012; Howcroft and Demberg, 2017) and, more recently, neural networks and deep learning models (Martinc et al., 2021; Imperial, 2021; Qiu et al., 2021).[1]

Similar to English, some work on ARA for texts in Spanish has developed methods that rely on surface features (Fernández-Huerta, 1959; Szigriszt Pazos, 2001). Others have implemented tools that extract readability indices (e.g. lexical diversity, word information, syntactic complexity) and used them as features to train standard machine learning classifiers to estimate a text's readability (Quispesaravia et al., 2016; López-Anguita et al., 2018; Bengoetxea and Gonzalez-Dios, 2021).

However, these studies were performed on small corpora of at most 300 texts (for both training and testing), limiting its generalisability. In addition, it is unknown to what extend modern neural models are able perform the task for texts in Spanish.

To mitigate the aforementioned issues, we introduce a new benchmark for training and evaluating models for ARA of texts in Spanish. Our benchmark includes the following contributions:[2]

- A collection of 6 datasets aimed to different audiences (e.g. children, Spanish learners as a second language, or people with learning disabilities) and with several "natural" levels of readability. With a total of 31,894 documents, this is the largest collection of texts in Spanish that has been used for ARA research.
- A simple baseline based on TF-IDF and Logistic Regression.

[1]See (Vajjala, 2022) for an up-to-date survey.
[2]Our datasets, models and code are available at: https://github.com/lmvasque/readability-es-benchmark

Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022), pages 188 - 198

- Neural models resulting from fine-tuning BERT (Devlin et al., 2019)-based pre-trained language models in monolingual and multilingual settings. We experimented with classifying texts at two ("simple" and "complex") and three ("basic", "intermediate" and "advanced") readability levels, as well as considering two text granularities (sentence-level and paragraph-level). We have demonstrated a better performance in a 2-class setting and also explain the limitations of working with a 3-class readability.

- An analysis of the performance of the neural models in different settings including zero-shot (no training, test with Spanish), cross-lingual zero-shot (training with English data, test with Spanish), monolingual few-shot (training with Spanish data) and cross-lingual few-shot (training with English and Spanish data, test with Spanish). Our study shows that multilingual models perform better at the paragraph level, while Spanish-specific models are the best at sentence level.

We expect to contribute to the development of ARA models that can help tailor relevant content for wider populations, and even benefit downstream NLP tasks, such as Text Simplification.

2 Related Work

Earlier readability studies focused on readers' background since audiences may have specific needs, and hence individual difficulties when reading a text. These audiences included people with dyslexia who struggle with long and uncommon words (Rello et al., 2013); or second-language learners, who are more affected by grammatical aspects than the content itself (Xia et al., 2016). Other studies focused on methods for readability assessment that relied on surface features, such as character and sentence counts (Dale and Chall, 1949; Collins-Thompson, 2014).

In recent years, researchers have explored alternative methods based on user-oriented studies where scroll interactions are captured to determine a document's easiness to read (Gooding et al., 2021). ARA has also been used in the evaluation of downstream NLP tasks, such as text simplification (Dell'Orletta et al., 2011) and word complexity analysis (Maddela and Xu, 2018).

Readability assessment itself has been approached in multiple ways, including through supervised and unsupervised methodologies (Martinc et al., 2021). The simplest approach is to use traditional metrics such as Gunning Fog Index (GFI, Gunning et al., 1952), Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL, Kincaid et al., 1975), which evaluated the readability of a document based on its characters, words, syllables, and sentences.

While most ARA work is done for texts in English, there is research for other languages such as Portuguese (Evaldo Leal et al., 2020; Scarton and Aluísio, 2010; Scarton et al., 2010), German (Hancke et al., 2012), French (François and Fairon, 2012), Italian (Dell'Orletta et al., 2011; Miliani et al., 2022), Russian (Reynolds, 2016), Vietnamise (Luong et al., 2017) and Swedish (Luong et al., 2017). However, these studies tend to be language and/or domain specific and thus, sparse without benchmarking multiple models.

Recent readability studies in Spanish are focused on specific audiences. These applications include the evaluation of the readability of e-government websites (Morato et al., 2021), the evaluation of the suitability of hearing aid user guides in Spanish (Gaeta et al., 2021) or in more specialised domain such as medical (Rodriguez and Singh, 2018). These studies mostly use traditional metrics (e.g., number of syllables, words, sentences), rather than neural approaches.

Finally, there are limited resources for readability in Spanish and most of them are shared within the domain of Text Simplification (Xu et al., 2015; Saggion et al., 2011; Štajner and Saggion, 2013) and Text Complexity analysis (Quispesaravia et al., 2016). We contribute with the collection of the proposed readability datasets (Section 3) and a benchmark of neural models (Section 4) to this growing field of research in Spanish.

3 Dataset Collection

We describe the data sources and characteristics of each dataset (Section 3.1) in our benchmark, as well as the standardisation process applied to each that allows for ARA experimentation (Section 3.2). We also analyse the readability of the documents and text groupings in our benchmark using readability metrics (Section 3.3), and comment on the datasets limitations (Section 3.4).

Dataset	Documents	Paragraphs	Paragraph/Doc	Sent.	Sent./Paragraph	Words	Words/Sent
CAES	30,935	30,935	1	325,135	11	5,154,567	15.85
Coh-Metrix-Esp	100	100	1	3,066	31	57,459	18.74
Hablacultura.com	217	713	3	2,607	4	62,582	24.01
Kwiziq	206	206	1	3,172	15	61,364	19.35
Newsela-es	243	5,444	22	53,470	10	1,079,921	20.20
Simplext	193	386	2	2,733	7	64,383	23.56
Total	31,894	37,784	31	390,183	77	6,480,276	121.70

Table 1: Datasets statistics including paragraphs and sentences.

3.1 Data Sources

Our benchmark includes resources scraped from the web, as well as datasets previously used for research in ARA and Text Simplification. Table 1 presents some statistics of the datasets, with more detailed descriptions below.

- **Newsela** (Xu et al., 2015): professional translators rewrote news articles (called version 0) to comply with multiple school grade levels (called versions 1 to 4, with higher versions being more readable). Our benchmark considers the Spanish portion of this dataset.
- **Simplext** (Saggion et al., 2011): collection of 200 short news articles that were rewritten following easy-to-read guidelines for wider audiences. While this corpus has been mostly used for Text Simplification research, it naturally provides documents in two levels ("complex" and "simple"), making it suitable for ARA studies.
- **Coh-Metrix-Esp (Cuentos)** (Quispesaravia et al., 2016): collection of 100 documents consisting of 50 children fables ("simple" texts) and 50 stories for adults ("complex" texts) scrapped from the web.
- **CAES**[3] (Parodi, 2015): the "Corpus de Aprendices del Español" (CAES) is a collection of texts created by Spanish L2 learners from Spanish learning centres and universities. Students had different learning levels, different backgrounds (11 native languages) and various levels of experience with the language. We used web scraping techniques to download a portion of the full dataset since its current website only provides content filtered by categories that have to be manually selected. The readability level of each text in CAES follows the Common European Framework of Reference for Languages (CEFR, Uchida et al., 2018). The corpus also includes information about the learners and the type of assignments with which they were assigned to create each text.
- **Other Language Learners Resources**: we collected articles from kwiziq,[4] a website dedicated to aid Spanish learning through automated methods. It also provides articles in different CEFR-based levels. We also collected texts from HablaCultura,[5] a website with resources for Spanish students, labeled by instructors following the CEFR. We scraped the freely available articles from both websites for our benchmark.

These datasets were selected since they inherently provide information about the readability levels of their texts. Although other resources exist (especially aimed at learners of Spanish L2), they have strict data-agreement licenses that prevent their use, or they are not publicly available.

3.2 Data Preprocessing

We used most of the documents from the datasets described in Section 3.1, without discarding any content. Since the documents have different types of readability labels ("complex" and "simple", school grade levels, or CEFR levels), we mapped them into two groups to allow easier and more standardised experimentation. Table 2 summarises this mapping, with further details given below.[6]

- **2-class (simple, complex):** when CEFR information was available, we split texts into "simple" for levels [A1, A2, B1], and "complex" for levels [B2, C1, C2]. For Newsela

[3] http://galvan.usc.es/caes/

[4] https://www.kwiziq.com/

[5] https://hablacultura.com/

[6] In Table 4 we show an example for each of our proposed classifications (2-class and 3-class.)

Group	Readability Label	Newsela	CAES	kwiziq	HablaCultura	Coh Cuentos	Simplext
2-class	simple	versions 3-4	A1, A2, B1			simple	
	complex	versions 0-1	B2, C1, C2			complex	
3-class	basic	grades 2-5	A1, A2			simple	
	intermediate	grades 6-8	B1, B2			-	
	advanced	grades 9-12	C1, C2			complex	

Table 2: Mapping between the original readability labels of each dataset in the benchmark to 2-class and 3-class groups for ARA experimentation.

Text Granularity	Group	Readability Labels	fernandez-huerta↑	szigriszt-pazos ↑	gutierrez-polini↑	crawford ↓
paragraph	2-class	simple	98.049	94.682	43.614	2.647
		complex	83.959	80.698	38.927	3.686
	3-class	basic	99.971	96.588	44.270	2.491
		intermediate	89.273	85.949	40.759	3.420
		advanced	82.909	79.673	38.545	3.746
sentence	2-class	simple	98.700	95.228	43.628	2.542
		complex	81.969	78.495	37.884	3.650
	3-class	basic	99.180	95.729	43.810	2.481
		intermediate	88.527	84.955	40.008	3.417
		advanced	80.953	77.495	37.460	3.689

Table 3: Readability indices for texts in each proposed readability level (2-class or 3-class) and granularity (paragraph or sentence). Arrows indicate if higher (↑) or lower (↓) values can be interpreted as more readable texts.

(with school grade levels), we classified entries as "simple" for simplification degrees [3-4], and "complex" for [0-1]. We skipped level 2 due to its close similarity with texts from versions 1 and 3. Datasets that already had binary labels (i.e. "simple" or "complex") were not modified.

- **3-class (basic, intermediate and advanced):** when school grade levels were available, grades [2-5] were considered as "basic", levels [6-8] as "intermediate", and levels [9-12] as "advanced". For datasets with CEFR information, we considered [A1, A2] as "basic", [B1, B2] as "intermediate", and [C1, C2] as "advanced". We only considered levels "basic" and "advanced" for datasets with only "simple" and "complex" labels, respectively.

We expect our benchmark to be used to develop neural-based ARA models, which are mostly based on BERT (Martinc et al., 2021; Imperial, 2021). As such, due to the input size limitations of BERT-based models, it would be difficult for them to handle full documents from some datasets in the benchmark. Previous work in English dealt with this by chunking documents by a certain number of sentences (Martinc et al., 2021). Instead, we rely on the natural boundaries or structure of documents to split them into paragraphs and sentences. This allows us to implement ARA models at different granularities. For Newsela and HablaCultura, paragraphs could be easily identified, since each one appears as a single line in the files. Documents with no clear paragraph-level divisions (e.g. from CAES, kwiziq and Coh Cuentos) were treated as having a single paragraph. Paragraphs were later split into sentences using NLTK (Bird et al., 2009).

3.3 Readability Assessment

We computed multiple readability indices for Spanish texts in order to validate the splitting of the data into the proposed 2-class and 3-class groups. We used `textstat` to calculate the following indices:[7]

Fernandez-Huerta (Fernández-Huerta, 1959): proposes the implementation of the Flesch Reading Ease (FRE) score for Spanish.[8] This score is given by Equation 1 where P is the number of syllables and F the number of sentences. The values range from 0 to 100, where the lower values correspond to university-level texts.

$$Score = 206.84 - (0.60 * P) - (1.02 * F) \quad (1)$$

[7] https://github.com/textstat/textstat

[8] We have used the corrected formula as proposed in https://linguistlist.org/issues/22/22-2332/#1.

Granularity	Text	Readability
Paragraph (2-class)	Sevilla es una ciudad de tradiciones, que las celebra con gran devoción y orgullo. Una de estas tradiciones es la ronda de las tunas a la “Inmaculada". Cada siete de diciembre por la noche, diferentes tunas se reúnen en la Plaza del Triunfo, en el centro de la ciudad, para entonar canciones tradicionales que se han cantado durante décadas. [..]	simple
Paragraph (2-class)	Voz de la guitarra mía, al despertar la mañana, quiere cantar su alegría a mi tierra mexicana. Yo le canto a sus volcanes, a sus praderas y flores, que son como talismanes del amor de mis amores. [..]	complex
Paragraph (3-class)	En Nochebuena, 24 de diciembre, cenamos en familia. En la cena típica hay gambas, langostinos, cordero o pavo, vino y champán o cava. Pero lo más típico son los dulces: el turrón, los polvorones, los mantecados y el mazapán.	basic
Paragraph (3-class)	Unos 15 kilómetros al sur de Sidi Ifni, en una de las playas vírgenes que baña el Océano Atlántico en esta parte de la costa, hay un viejo barco encallado. Se trata de una enorme mole oxidada de origen incierto, abandonada en este despoblado punto de la costa. [..]	intermediate
Paragraph (3-class)	Existen artistas con un don, seres únicos elegidos para trasmitir emociones e inquietudes de una manera diferente y a la vez familiar. De aquel niño que escudriñaba a su madre mientras ella interpretaba cartas de amor y de muerte a las vecinas del pueblo, queda la mirada pícara y luminosa del visionario, de aquel que antes de inventar la fábula ya ha imaginado el final. [..]	complex

Table 4: Examples from HablaCultura and Kwiziq paragraph datasets.

Szigriszt-Pazos (Szigriszt Pazos, 1993): measures the “perspicuity” (i.e. intelligibility) of texts using Equation 2, where S is the total of syllables, P is the total of words, and F is the number of sentences.

$$Score = 206.835 - \frac{62.3 * S}{P} - \frac{P}{F} \quad (2)$$

Gutierrez-Polini (Gutiérrez de Polini, 1972): a readability metric designed directly for Spanish, without adapting existing English readability measures. Its value is given by Equation 3, where L is the number of characters, P the number of words, and F the number of sentences.

$$Score = 92.5 - \frac{9.7 * L}{P} - \frac{0.35P}{F} \quad (3)$$

Crawford (Crawford, 1989): this index is limited to measure the difficulty for children at primary school to learn a text. Its value is given by Equation 4, where OP is the number of sentences for every 100 words, and SP the number of syllables for every 100 words. The output refers to the years in primary school needed to understand a text. Therefore, the higher the number of years at school, the less readable the text will be. We are interested in lower values for more legible texts.

$$Score = -0.205OP + 0.049SP - 3.407 \quad (4)$$

Table 3 shows these readability indices for each of the proposed dataset splits for sentences and paragraphs. For both granularities, all scores for texts in each group differ significantly between readability levels. For example, for paragraphs, in the 3-class group, the corresponding fernandez-huerta index for “basic” texts is more than 10 points higher than for “intermerdiate” texts, which in turn is around 7 points higher than for “advanced” texts. Since for this index higher scores indicate more readable texts, this indicates that the split is adequate. In general, these results support our proposed mapping summarised in Table 2.

3.4 Datasets Limitations

Texts from the the Spanish portion of the Newsela dataset are translations from the original English articles.[9] This may impact the quality and generalisation capabilities of the models we created compared to Newsela-based studies in English.

Texts in CAES were written by learners of Spanish with different backgrounds and levels of experience. Therefore, there are grammatical and syntactical errors in their construction. Also, the topics of each text depend on the CEFR levels of the students. For example, A1 students mostly write emails, while B1 students write essays. This could bias the ARA classifiers to learn to identify topics rather than readability levels. For this reason, we did not include CAES in our experiments (Sec. 6). However, this dataset will still be available for further studies where these limitations are not relevant or actually want to be explored.

[9] https://newsela.com/about/blog/how-to-use-spanish-texts-on-newsela/

Group	Subset	Readability Labels	OneStopEnglish	Paragraphs (ES)	Sentences (ES)
2-class	train	complex	-	2,470	9,532
		simple	-	2,096	7,708
	valid	complex	-	313	1,181
		simple	-	258	974
	test	complex	-	317	1,249
		simple	-	254	908
3-class	train	basic	145	1,603	6,147
		intermediate	150	1,975	6,187
		advanced	158	1,512	6,424
	valid	basic	24	201	804
		intermediate	17	256	770
		advanced	16	179	770
	test	basic	20	199	748
		intermediate	22	271	818
		advanced	15	167	779

Table 5: Number of samples for each dataset, stratified by split and readability labels and levels

For our experiments in the next section, we used all the datasets in the benchmark. However, our final release will include only those that are freely available. Researchers would need to request the specific licenses for Newsela and Simplext before we could share with them our specific data splits.

4 Neural ARA Models

Considering the characteristics of our dataset, we treated ARA as a classification task, and used the datasets in our benchmark to implement neural supervised models. Following previous work (Martinc et al., 2021; Lee and Vajjala, 2022), we used BERT (Devlin et al., 2019) models with fully connected layers and softmax outputs. As base BERT models, we experimented with:

- **BERTIN** (De la Rosa et al., 2022): a RoBERTa-based (Liu et al., 2019) pretrained model using Spanish corpora and a perplexity sampling that allowed its fine-tuning with a reduced training time and data.

- **mBERT** (Devlin et al., 2019): a BERT-based model, trained over 102 languages, including Spanish. This model will determine whether a multilingual, general-purpose model is appropriate for this task, in comparison to a dedicated model for Spanish. Also, previous work has relied on these models for multilingual readability assessment in English, French and Spanish (Lee and Vajjala, 2022).

We used the BERTIN[10] and mBERT[11] checkpoints available in HuggingFace (Wolf et al., 2020) to implement six models in the following settings:

Zero-shot. Models based on BERTIN and mBERT are not trained in any task-specific data and are used directly to make predictions in the test set. This aims to explore if pre-trained models (monolingual or multilingual) are by default capable of performing ARA without any fine-tuning. We refer to these models as BERTIN (Zero) and mBERT (Zero).

Cross-lingual Zero-shot. We study if a multilingual model is able to perform ARA after being fine-tuned using task specific data but from a different language. In particular, we fine-tune mBERT using the OneStopEnglish corpus (Vajjala and Lučić, 2018), which includes articles from newspapers rewritten by teachers of English as a second language. Texts in this dataset are divided into elementary, intermediate, and advanced levels. As such, we limit our experiments to the evaluation of the 3-class groups since OneStopEnglish does not have a predefined alignment for 2 levels. Using the intermediate corpus in any of the other categories could be unreliable. In addition, removing the intermediate level to evaluate the 2-class groups would result in a very small dataset. We refer to this model as mBERT (EN).

[10] https://huggingface.co/bertin-project/bertin-roberta-base-spanish

[11] https://huggingface.co/bert-base-multilingual-uncased

Granularity	Model	2-class			3-class		
		F1–Score	Precision	Recall	F1–Score	Precision	Recall
Paragraph	Baseline (TF-IDF+LR)	0.829	0.832	0.827	0.556	0.563	0.550
	BERTIN (Zero)	0.308	0.222	0.500	0.227	0.284	0.338
	BERTIN (ES)	0.924	0.923	0.925	0.772	0.776	0.768
	mBERT (Zero)	0.308	0.222	0.500	0.253	0.312	0.368
	mBERT (EN)	-	-	-	0.505	0.560	0.552
	mBERT (ES)	**0.933**	**0.932**	**0.936**	0.776	0.777	0.778
	mBERT (EN+ES)	-	-	-	**0.779**	**0.783**	**0.779**
Sentence	Baseline (TF-IDF+LR)	0.811	0.814	0.808	0.525	0.531	0.521
	BERTIN (Zero)	0.367	0.290	0.500	0.188	0.232	0.335
	BERTIN (ES)	**0.900**	**0.900**	**0.900**	**0.699**	**0.701**	**0.698**
	mBERT (Zero)	0.367	0.290	0.500	0.278	0.329	0.351
	mBERT (EN)	-	-	-	0.521	0.565	0.539
	mBERT (ES)	0.893	0.891	0.896	0.688	0.686	0.691
	mBERT (EN+ES)	-	-	-	0.679	0.676	0.682

Table 6: F1–score, precision, and recall scores for readability baselines. In **bold** we select the best model for each combination of granularity in a group of readability levels.

Monolingual Few-shot. This is the standard supervised setting where BERTIN and mBERT are fine-tuned using the training data from our benchmark. We consider this setting as few-show since the Spanish corpora is not large. We refer to these models as BERTIN (ES) and mBERT (ES).

Cross-lingual Few-shot. We experiment with further fine-tuning the cross-lingual mBERT (EN) model with the language-specific training data from our benchmark (few-shot). We refer to this model as mBERT (EN+ES).

Each of these settings is applied to the two text granularities (paragraph and sentence) and the two groups of readability labels (2-class and 3-class).

5 Experimental Setting

Baseline. We implemented a simple approach based on Logistic Regression and TF-IDF.[12] We extracted the features for each text using TF-IDF algorithm.[13]. Then, we trained a Linear Regression classifier[14] using these features in splits (train/dev).

Data Splits. We randomly split all data into 80% for training, 10% for validation and 10% for testing, consistently across all experiments. We show the data distribution in Table 5.

[12] https://www.kaggle.com/code/kashnitsky/logistic-regression-tf-idf-baseline/notebook

[13] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[14] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

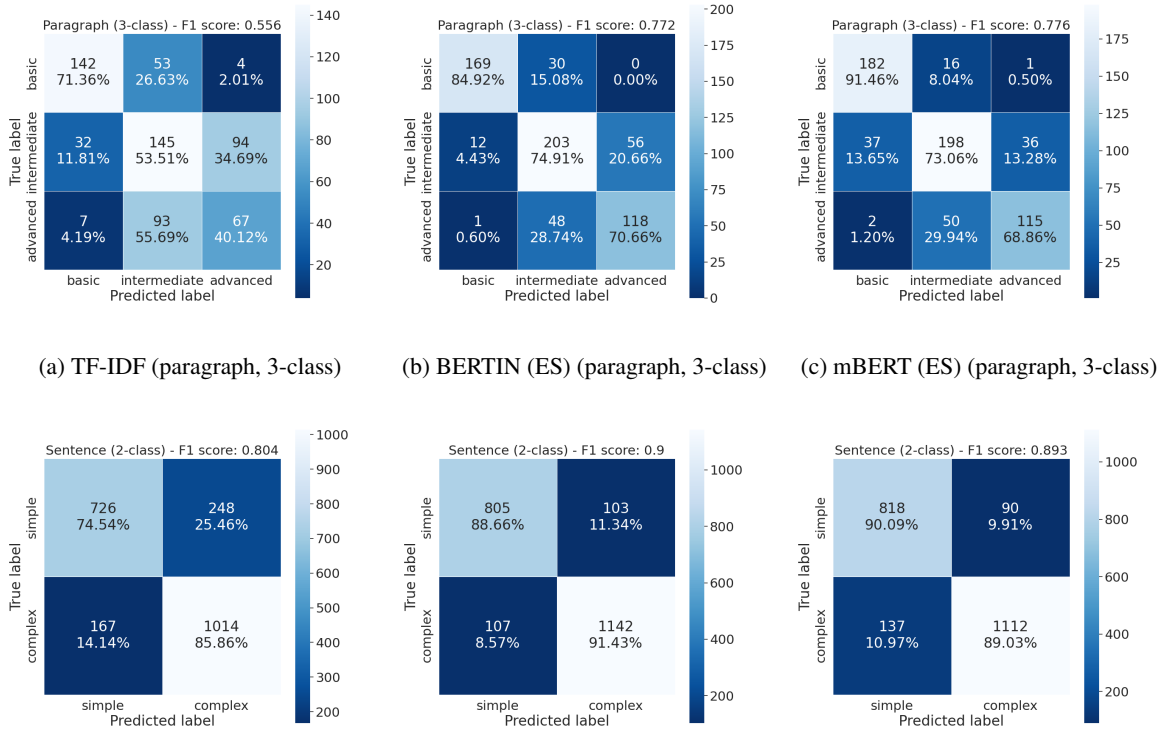
Training Details. We performed hyperparameter optimisation and observed training behaviour to select the most stable models (i.e. less variability of the validation loss) as the best for the task. We selected AdamW optimizer, using a beta value of 0.9. For our 2-class and 3-class experiments we used a learning rate of 3e-6, weight_decay of 0.02, batch size of 16 and a number of epochs equal to 10. Once the models were trained, we evaluated the models in the held-out test set. We trained with a 1 Nvidia v100 GPUs (16GB GPU RAM) and training time was about 4 hours for the biggest dataset (sentence-level).

6 Results

Table 6 shows the performance in the test set of the baseline and neural models in all the training settings previously described. In addition, Figure 1 presents representative confusion matrices for our baseline and best performing models. Due to space constraints, we only include matrices for the paragraph-based corpus in the 3-class group and the sentence-based corpus in the 2-class group.

Most BERT-based models are consistently better than the TF-IDF baseline, for all text granularities and readability labels. An exception are mBERT (Zero) and BERTIN (Zero) who had the lowest performance in all cases. This implies that pretrained models by themselves are unable to perform ARA for Spanish texts in our benchmark.

All models dropped their performance when trained in the 3-class setting compared to the



(a) TF-IDF (paragraph, 3-class) (b) BERTIN (ES) (paragraph, 3-class) (c) mBERT (ES) (paragraph, 3-class)

(d) TF-IDF (sentence, 2-class) (e) BERTIN (ES) (sentence, 2-class) (f) mBERT (ES) (sentence, 2-class)

Figure 1: Confusion matrices for LR-TF-IDF, BERTIN (ES) and mBERT (ES) in 2-class and 3-class task setting (paragraphs and sentences).

2-class one, including in the zero-shot models. As shown in Figure 1a, Figure 1b and Figure 1c, an "intermediate" class makes it more difficult to classify the samples, especially in the boundaries ("basic"/"intermediate" and "intermediate"/"advanced"). We can observe that it is easier to distinguish between "basic" and "advanced", evidenced by just having a few misclassified samples (see the matrices corner values). In contrast, there is a significant number of incorrect samples between the "intermediate"-"advanced" boundaries in the 3 models.

The cross-lingual mBERT (EN) models performed comparably to or worse than the baseline in the 3-class group for text granularities. Adding language-specific data, makes the model perform better, making mBERT (EN+ES) comparable to the mBERT (ES) model in all cases. While BERTIN (ES) is still the best in the 3-class paragraph setting, these results suggest that a multilingual pre-trained model can be leveraged for ARA in Spanish is no language-specific model is available.

Overall, we can observe that fine-tuning either BERTIN or mBERT with language-specific data would result in a good performing model for the task, for all the settings we considered. As such, these serve as strong baselines for future research.

7 Discussion

Our zero-shot (cross-lingual) experiments demonstrate that the readability task is not trivial to learn to perform ARA, and that it is directly transferable between languages in the settings we studied. For BERTIN (Zero) and mBERT (Zero), the models were not previously trained for the readability classification task resulting in poor performance.

The decrease in performance between models in the paragraph-based and sentence-based datasets could be attributed to multiple reasons. First, not all texts in the datasets could be mapped to three classes, resulting in fewer instances for training and evaluation. In addition, it may be easier for a model to distinguish between extremes ("simple" or "complex") than to also consider an "intermediate" class. This effect is clearer in the analysis of the confusion matrices in Figure 1.

Regarding our best models, we benefited from the fact that the BERTIN model was trained on

Spanish texts, which contributes for a better "understanding" of readability in this specific language. The multilingual model (mBERT) was trained in multiple languages beside Spanish, which could have contributed to the improvement of its results.

While our results may be encouraging, we state the limitations of our experiments. When short and simple texts are used for training, readability results can easily be related to short sentences and words. However, texts can also be readable in other scenarios, such as using active voice, instead of passive voice, being consistent in the narrative (e.g. following on the same topic) and the use of simpler words, which are not necessarily shorter. These features are harder to learn, as shown in our 3-class experiments, but with the use of more corpora from multiple domains, it may be possible to obtain more robust ARA models. Regarding the models, we could consider that BERTIN model is not uncased, whereas the multilingual model is; this could also be a limitation and a variability factor in the models performance. Overall, current datasets are scarce, and it is advisable to train in wider corpora for the generalisation in multiple domains.

8 Conclusion and Future Work

In this paper, we have introduced a new benchmark for ARA of texts in Spanish. We combined existing datasets for research in ARA and Text Simplification, with other resources scraped from the web. With these data, we trained neural ARA models based on BERT to classify texts into "simple" and "complex" (2-class), or "basic", "intermediate" and "advanced" (3-class), at two levels of text granularities (paragraph and sentence). The neural models proved to be better than simple baselines.

In the future, we plan to include more datasets to the benchmark, such the one used in (López-Anguita et al., 2018). In addition, we plan on training feature-based models for a more comprehensive evaluation of our neural models. Finally, it would be interesting to study the effect that larger multilingual pre-trained models, like XLM-R (Conneau et al., 2020), could have on the performance of neural models.

All of our models are publicly available, as well as demo that showcases their performances. We expect that research communities in Spanish speaking countries will benefit from this effort towards the further development of the field.

References

Kepa Bengoetxea and Itziar Gonzalez-Dios. 2021. Multiaztertest: a multilingual analyzer on multiple levels of language for readability assessment.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly, Beijing.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alan N Crawford. 1989. *Fórmula y gráfico para determinar la comprensibilidad de textos de nivel primario en castellano*. Lectura y Vida. Revista Latinoamericana de Lectura.

Edgar Dale and Jeanne S. Chall. 1949. The concept of readability. *Elementary English*, 26(1):19–26.

Javier De la Rosa, Eduardo G Ponferrada, Manu Romero, Paulo Villegas, Pablo González de Prado Salas, and María Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia 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*, pages 73–83, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sidney Evaldo Leal, João Marcos Munguba Vieira, Erica dos Santos Rodrigues, Elisângela Nogueira Teixeira, and Sandra Aluísio. 2020. Using eye-tracking data to predict the readability of Brazilian Portuguese sentences in single-task, multi-task and sequential transfer learning approaches. In *Proceedings of the*

28th International Conference on Computational Linguistics, pages 5821–5831, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Thomas François and Eleni Miltsakaki. 2012. Do NLP and machine learning improve traditional readability formulas? In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 49–57, Montréal, Canada. Association for Computational Linguistics.

Thomas François and Cedric Fairon. 2012. An "ai readability" formula for french as a foreign language. In *EMNLP*.

Laura Gaeta, Edward Garcia, and Valeria Gonzalez. 2021. Readability and suitability of spanish-language hearing aid user guides. *American Journal of Audiology*, 30(2):452–457.

Sian Gooding, Yevgeni Berzak, Tony Mak, and Matt Sharifi. 2021. Predicting text readability from scrolling interactions. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 380–390, Online. Association for Computational Linguistics.

Robert Gunning et al. 1952. Technique of clear writing.

L.E. Gutiérrez de Polini. 1972. *Investigación sobre lectura en Venezuela*. Documento presentado a las Primeras Jornadas de Educación Primaria, Ministerio de Educación, Caracas.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of COLING 2012*, pages 1063–1080, Mumbai, India. The COLING 2012 Organizing Committee.

David M. Howcroft and Vera Demberg. 2017. Psycholinguistic models of sentence processing improve sentence readability ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 958–968, Valencia, Spain. Association for Computational Linguistics.

Joseph Marvin Imperial. 2021. BERT embeddings for automatic readability assessment. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 611–618, Held Online. INCOMA Ltd.

J. Peter Kincaid, Robert P. Fishburne, R L Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. In *Institute for Simulation and Training*.

Justin Lee and Sowmya Vajjala. 2022. A neural pairwise ranking model for readability assessment. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3802–3813, Dublin, Ireland. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

An-Vinh Luong, Diep Nguyen, and Dinh Dien. 2017. Examining the text-length factor in evaluating the readability of literary texts in vietnamese textbooks. *2017 9th International Conference on Knowledge and Systems Engineering (KSE)*, pages 36–41.

Rocío López-Anguita, Arturo Montejo-Ráez, Fernando J. Martínez-Santiago, and Manuel Carlos Díaz-Galiano. 2018. Legibilidad del texto, métricas de complejidad y la importancia de las palabras. *Procesamiento del Lenguaje Natural*, 61(0):101–108.

Mounica Maddela and Wei Xu. 2018. A word-complexity lexicon and a neural readability ranking model for lexical simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3749–3760, Brussels, Belgium. Association for Computational Linguistics.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Martina Miliani, Serena Auriemma, Fernando Alva-Manchego, and Alessandro Lenci. 2022. Neural readability pairwise ranking for sentences in italian administrative language. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, Online. Association for Computational Linguistics.

Jorge Morato, Ana Iglesias, Adrián Campillo, and Sonia Sanchez-Cuadrado. 2021. Automated readability assessment for spanish e-government information. *Journal of Information Systems Engineering and Management*, 6:em0137.

Giovanni Parodi. 2015. Corpus de aprendices de español (caes). *Journal of Spanish Language Teaching*, 2(2):194–200.

Xinying Qiu, Yuan Chen, Hanwu Chen, Jian-Yun Nie, Yuming Shen, and Dawei Lu. 2021. Learning syntactic dense embedding with correlation graph for automatic readability assessment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3013–3025, Online. Association for Computational Linguistics.

Andre Quispesaravia, Walter Perez, Marco Sobrevilla Cabezudo, and Fernando Alva-Manchego. 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'16)*, pages 4694–4698, Portorož, Slovenia. European Language Resources Association (ELRA).

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. 2013. Frequent words improve readability and short words improve understandability for people with dyslexia. In *INTERACT*.

Robert Joshua Reynolds. 2016. Insights from russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories. In *BEA@NAACL-HLT*.

Jorge A. Rodriguez and Karandeep Singh. 2018. The spanish availability and readability of diabetes apps. *Journal of Diabetes Science and Technology*, 12(3):719–724. PMID: 29291639.

Horacio Saggion, Elena Gómez-Martínez, Esteban Etayo, Alberto Anula, and Lorena Bourg. 2011. Text simplification in simplext. making text more accessible. *Proces. del Leng. Natural*, 47:341–342.

Carolina Scarton, Caroline Gasperin, and Sandra Aluisio. 2010. Revisiting the readability assessment of texts in portuguese. In *Advances in Artificial Intelligence – IBERAMIA 2010*, pages 306–315, Berlin, Heidelberg. Springer Berlin Heidelberg.

Carolina Evaristo Scarton and Sandra Maria Aluísio. 2010. Coh-metrix-port: a readability assessment tool for texts in brazilian portuguese. In *International Conference on Computational Processing of the Portuguese Language - Propor*. SBC.

Sanja Štajner and Horacio 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. Asian Federation of Natural Language Processing.

Francisco Szigriszt Pazos. 1993. *Sistemas Predictivos de Legibilidad del Mensaje Escrito: Formula de Perspicuidad*. Universidad Complutense de Madrid.

Francisco Szigriszt Pazos. 2001. *Sistemas predictivos de legilibilidad del mensaje escrito: fórmula de perspicuidad*. Universidad Complutense de Madrid, Servicio de Publicaciones.

Satoru Uchida, Shohei Takada, and Yuki Arase. 2018. Cefr-based lexical simplification dataset. In *Proceedings of International Conference on Language Resources and Evaluation*, volume 11, pages 3254–3258. European Language Resources Association.

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the 13th Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173, Montréal, Canada. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.