Int J Artif Intell Educ (2024) 34:1417–1459
https://doi.org/10.1007/s40593-024-00393-2

ARTICLE

Exploring Automatic Readability Assessment for Science Documents within a Multilingual Educational Context

Suna-Şeyma Uçar[1] · Itziar Aldabe[1] · Nora Aranberri[1] · Ana Arruarte[2]

Accepted: 22 January 2024 / Published online: 7 February 2024

Abstract

Current student-centred, multilingual, active teaching methodologies require that teachers have continuous access to texts that are adequate in terms of topic and language competence. However, the task of finding appropriate materials is arduous and time consuming for teachers. To build on automatic readability assessment research that could help to assist teachers, we explore the performance of natural language processing approaches when dealing with educational science documents for secondary education. Currently, readability assessment is mainly explored in English. In this work we extend our research to Basque and Spanish together with English by compiling context-specific corpora and then testing the performance of feature-based machine-learning and deep learning models. Based on the evaluation of our results, we find that our models do not generalize well although deep learning models obtain better accuracy and F1 in all configurations. Further research in this area is still necessary to determine reliable characteristics of training corpora and model parameters to ensure generalizability.

Keywords Natural language processing · Educational text classification · Science domain · Multilingual education · Automatic readability assessment

✉ Suna-Şeyma Uçar
sunaseyma.ucar@ehu.eus

Itziar Aldabe
itziar.aldabe@ehu.eus

Nora Aranberri
nora.aranberri@ehu.eus

Ana Arruarte
a.arruarte@ehu.eus

[1] HiTZ Basque Center for Language Technologies - Ixa NLP Group, University of the Basque Country (UPV/EHU), Donostia-San Sebastián, Spain

[2] Department of Computer Languages and Systems, University of the Basque Country (UPV/EHU), Donostia-San Sebastián, Spain

Introduction

Automatic Readability Assessment (ARA) is a well-established area of research that seeks to automatically determine the level of difficulty a written text might pose to a reader. Given the importance of fully comprehending a text in countless situations, studies have been conducted from multiple perspectives and contexts, such as the health sector (Basch et al., 2020), education (Vajjala & Lučić, 2018) or even computer science (Scalabrino et al., 2018), among others.

Over the years, researchers have proposed diverse approaches to develop ARA models. Traditional readability formulas, such as Flesch-Kincaid and Gunning (see Tekfi, 1987 for a full review), have long been used in education-related areas and beyond (François & Miltsakaki, 2012; Martinc et al., 2021). More recently, in an attempt to dive further into a broader set of characteristics that may have an effect on a reader's ease to understand a text, researchers have explored feature-based Natural Language Processing (NLP) approaches with considerable success (Vajjala & Lučić, 2018; Bengoetxea et al., 2020). In the last couple of years, neural approaches have also entered the picture Azpiazu and Pera (2019); Schicchi et al. (2020). The results of the NLP approaches have shown great potential, but there is yet ample room for exploration, as studies have so far mainly focused on English and the domain-adaptability of the models has not been thoroughly examined.

In this work, we seek to test the applicability of NLP approaches to ARA for educational material in the Obligatory Secondary Education (ESO for its acronym in Spanish) in the Basque Autonomous Community (BAC). ESO encompasses a total of four grades and covers the ages between 12 and 16. In the BAC, two official languages, Basque – the minority language and main language of instruction – and Spanish – the majority language –, and a foreign language, English, coexist in the majority of classrooms.

The Basque education curriculum promotes the implementation of the Integrated Treatment of Languages (Bikandi & Valls, 2008) and Project-based learning (Hung et al., 2008) approaches. In brief, this means that learners address challenges that involve a number of science, technology, engineering and mathematics (STEM) subjects and develop language skills in line with such challenges in all three languages. To properly implement and combine the teaching approaches, teachers are faced with the arduous task of gathering textual material about the topic of the project in three languages.

In this context, our ultimate goal is to create an ARA model for a multilingual context that will help teachers decide on the adequacy of a text for a particular secondary education student group. Specifically, we focus on predicting the readability level of STEM subject texts in Basque, Spanish and English, which remains a largely under-researched area. What is more, we aim to do so for complete documents (see Section 3).

To build a model based on NLP approaches, we require not only efficient learning algorithms but also annotated science text corpora. It is not yet known of any publicly accessible, domain-specific, graded corpus in Basque, Spanish, or English at secondary education level. Therefore, as a first step toward our objective, we present the compilation process of annotated document-level corpora for Basque (BasqueARA), Spanish (Agrega2Es) and English (Agrega2En+), which will be released upon the acceptance of the paper. Given the nature of our context, the first two corpora consist of texts for native speakers while the English corpus comprises texts created for non-native learners. With regards NLP learning techniques, we explore the behaviour of Machine Learning (ML) approaches using the Support Vector Machine (SVM) algorithm and Deep Learning (DL) approaches using transformer architectures (cf. Sections 2 and 4). Previous studies have identified SVM classification as the most effective among supervised machine learning algorithms, showcasing substantial improvements in classification results compared to other methods (Liu et al., 2010). Additionally, SVM demonstrates a strong ability to capture the inherent characteristics of the data (Bhavani & Kumar, 2021). On the other hand, DL approaches have also shown improved accuracy in readability tasks (Azpiazu & Pera, 2019; Imperial, 2021).

The remaining of this paper is structured as follows: Section 2 describes related research starting from early research in ARA and up to the newest methods of DL; Section 3 presents the compilation process and characteristics of our Basque, Spanish and English educational science text corpora; Section 4 presents the main architectures used in the experiments and Section 5 the developed feature-based ML and DL models; Section 6 explores the generalizability of the models and, finally, Section 7 outlines the conclusions.

Related Work

Early research in ARA focused on the use of mathematical equations that calculated the readability of a text based on shallow features such as number of words, sentences and difficult words, among others. The reliability of those formulas is questioned despite their broad use (Davison & Kantor, 1982; Si and Callan, 2001; Vajjala & Meurers, 2012). Authors claim that the formulas ignore an important in many aspects such as word order, content and purpose of the text (Klare, 1963). Another widely-accepted drawback of these formulas is that they are mostly developed for English (Vajjala, 2021), even though there have been a number of initiatives to create similar methods for other languages such as Fernández-Huerta formula for Spanish (Fernández Huerta, 1959). Compared to English and even Spanish, research in Basque in this area emerged rather late, when NLP approaches were being tested. For this reason, to our knowledge, no traditional formula has been developed for Basque.

Thanks to the advances in computational capacity and the development of ML approaches, researchers started exploring data-driven approaches for ARA as well. As the quality of the texts is thought to be one of the key factors determining the performance of a model, in the last years, various corpora have been compiled and made available for further research in the area. Among the various corpora developed for English, WeeBit (Vajjala & Meurers, 2012), OneStopEnglish (Vajjala & Lučić, 2018) and Newsela (Xu et al., 2015) have been commonly used. While WeeBit comprises educational articles targeting different age groups and school grades, OneStopEnglish and Newsela consist of news articles for adult English as a second language learners and children of different grade levels, respectively. Nadeem and Ostendorf (2018) presented a corpus consisting of texts from the Siyavula project[1], an educational

[1] https://www.siyavula.com/read

initiative with the aim of creating natural sciences and mathematics materials for high school students. The compiled English corpus consists of short science texts from 44 science and 11 history and social sciences textbooks developed for native speakers of English. Recently, Crossley et al. (2022) introduced the CommonLit Ease of Readability (CLEAR) corpus, which comprises 5000 English informational and literature texts rated by humans.

The resources available for Spanish are more scarce. Azpiazu and Pera (2019) introduced the VikiWiki dataset, which comprises texts in 6 languages, including Spanish and English. In this case, the texts, which come from Vikidia(.org) and Wikipedia, were annotated as simple or complex rather than according to academic grades. Interestingly, Lee and Vajjala (2022) published the Spanish version of the Newsela corpus. For Basque, we were able to identify a single corpus, that complied by Gonzalez-Dios et al. (2014), consisting of texts classified as simple or complex.

In reference to learning algorithms, feature-based supervised approaches were the first to be explored. ARA has been treated as a form of regression problem (Feng et al., 2010), classification task (Vajjala & Meurers, 2013), or ranking problem (Xia et al., 2016) with a growing tendency to use SVM classifiers for text classification. An experiment on the WeeBit corpus conducted by Xia et al. (2016) showed their SVM classifier obtained 80.3% accuracy using 5-fold cross-validation using traditional lexicosyntactic language modeling and discourse features. Vajjala and Lučić (2018) conducted a classification experiment by training a SVM using Sequential Minimal Optimization (SMO) with OneStopEnglish, and they achieved an accuracy of 78.13%. Similarly, in a 3-level classification scenario also using SMO, Bengoetxea and Gonzalez-Dios (2021) obtained 90.09% accuracy for the OneStopEnglish test set when using the 50 most predictive features. Crossley et al. (2022) reported 0.726 RMSE (Root Mean Square Error) with a linear model using 107 linguistic features obtained from various extraction tools.

Even though feature-based models continue to be researched, the latest neural network-based approaches incorporate language models that obtain higher accuracy rates in ARA tasks (Lee et al., 2021). For example, predictions of textual embeddings such as the HAN and BERT models have been used as additional features in SVM models and evaluated in WeeBit and Newsela (Deutsch et al., 2020). Imperial (2021) explores the concatenation of BERT embeddings and handcraft linguistic features to be used in ML algorithms for English and Filipino. A range of neural architectures have also been explored. Nadeem and Ostendorf (2018) test GRU and hierarchical RNN architectures on the WeeBit corpus, which prove successful in training a paragraph-level ARA model on Siyavula. Azpiazu and Pera (2019) use multi-attentive RNNs on the VikiWiki dataset with an accuracy of 84.7%. Lee and Vajjala (2022) propose a neural pairwise ranking model and obtained a zero-shot cross-lingual ranking accuracy of over 80% for Spanish when trained on English data from Newsela.

Given this background, in this paper we explore the options that feature-based ML and DL approaches offer in our science educational context to work with Basque, Spanish and English corpora.

Scientific Discourse and Readability

From the review on ARA, we infer that readability measurements have been used to assess the suitability of texts for specific academic grades with little to no adaptation to the characteristics of the texts involved in terms of genre, type or topic. Given the differences between general domain texts and scientific discourse, we wonder if the generic approaches studied so far are indeed useful to evaluate the difficulty of science textbooks.

As posited by Franco Aixelá (2015) based on the principles outlined by Castellví (2004), scientific texts have a particular set of cognitive, linguistic and pragmatic characteristics. Just in terms of purpose, and therefore text type, the scientific discourse looks to communicate knowledge by demonstrating theories, proving hypothesis, and explaining objective phenomena (Shishkova & Popok, 1989).

In terms of linguistic features, which, we could argue, are the ones considered by ARA approaches to a greater or lesser extent, the difference lies in that, in contrast to general texts, scientific texts tend to have a more rigid structure, a higher presence of terminology that is less accessible the more specialised the field is, a simpler syntax and are consistently formal Franco Aixelá (2015). Roméu Escobar (2002) presents an elaborate list of linguistic features of scientific texts that encompasses lexical, morphological and syntactic traits.

While counter intuitively at first sight, it is accepted that narrative texts, often found in everyday discourse and literature, are easier to understand than informative-argumentative texts, core text types within scientific texts (Pérez Zorrilla, 2005), even when the latter are syntactically simpler and more repetitive (Muñoz Calvo et al., 2013). This is mainly because the scientific texts refer to and associate complex events that we cannot always relate to our life experiences and require a higher level of abstraction and linguistic competence (Guevara Benítez et al., 2015).

If we consider scientific articles, Plavén-Sigray et al. (2017) claim that the readability of scientific texts is steadily declining, the main reason being the increased use of scientific jargon. Ball (2017) adds that it is not only scientific jargon that makes scientific texts difficult to read, but rather the use of multi-syllable daily words, that are difficult to process by the reader. Interestingly, Ehara (2022) points out that 10%-30% of the scientific texts available are not readable to intermediate ESL learners.

This being the case, authors such as Uribe (2007) have gone as far as proposing that if, as stated by Gutiérrez Rodilla (1998), in contrast to the everyday discourse, understanding scientific discourse requires a high linguistic awareness and therefore specific training, science teachers are also language teachers (Sutton, 2003).

All in all, in STEM, scientific discourse is considered to be an essential part of the learning process and, as such, teachers are encouraged to use texts specific to the discipline in classrooms (Daugherty et al., 2017). Nevertheless, as we mentioned, these texts can be difficult to process for different student profiles (Arfé et al, 2018). Yet, most readability studies carried out so far have focused on evaluating the suitability of science textbooks by assuming that the readability approaches used are accurate (Chiang-Soong & Yager, 1993; Gyasi, 2013; Nwafor et al., 2022; Hu et al., 2021) and little has been done to investigate their performance in this distinct scenario.

Corpora for Training Models

While building readability models, training corpora have been identified as a key factor in ensuring their accuracy. However, to the best of our knowledge, no publicly available, domain-specific, graded corpus exists to work on science texts for secondary education in Basque, Spanish and English. This is not surprising as the difficulty of finding corpora for readability assessment has been mentioned in previous studies (Petersen & Ostendorf, 2009). Therefore, our first aim was to compile an open, context specific corpus that would allow us to test the performance of different NLP approaches to ARA for our intended educational setting.

In our search for adequate material for our experimental context, we encountered the Agrega2 project[2]. It is a Spanish national initiative co-funded by the European Union (Feder) aimed at creating a unified online repository of teaching materials that cover all educational stages and official languages involved in the Spanish education curriculum. The project repository hosts a wide range of materials (learning objects, teaching sequences, learning programs, etc.) developed by the autonomous communities within Spain and covers the different subjects and languages involved.

It is important to note that while classroom documents in Basque and Spanish are primarily directed at native speakers, materials in English are for non-native learners, following the main profile of students. Consequently, it is expected that materials for a particular grade will vary in textual complexity to fit the language competence of the learners (Xia et al., 2016), that is, for a particular grade, English texts are expected to display simpler vocabulary and structure than Spanish and Basque texts.

Interestingly, all materials in Agrega2 are labeled according to the grade for which they are intended, and thus they are a reliable source as a gold-standard for ARA research. The Agrega2 repository served as our main source for texts: we extracted all classroom-ready texts for the four grades of ESO, which covers ages between 12 and 16[3], that belonged to the category of natural sciences and that were accessible between September-November 2021. As we will see in the description for each corpus, however, the available volume of text was not always sufficient for our experiments and therefore additional data was collected to extend the corpora. The vast majority of the material stored in the Agrega2 repository is free.

Regardless of the source of the texts, the procedure to prepare them for inclusion in our corpora was the same. As we are interested in providing teachers with a tool to identify the science texts that can be useful for the classroom, we limited the content of our corpus to materials with science reading passages and discarded the rest (exercises, for example). We define a document as a reading passage that is self-contained, useful to address a specific topic, definition, theory, fact, formula, or similar in a science classroom. The documents were obtained by dividing didactic units into sections guided by each unit's headings and subheadings, that is, we created a document with the text included under heading 1 up until subheading 1.1, then another with the text included within subheading 1.1 up until subheading 1.2, and so on. For deeper

[2] http://agrega.educacion.es/visualizadorcontenidos2/AcercaDeAgrega/AcercaDeAgrega.do

[3] This age range corresponds to Key Stage 3 and 4 in the UK, meanwhile it corresponds from 6th to 10th grade in the US.

subheadings such as 1.1.1 and 1.1.2, the decision to merge or separate their text to create a document was made based on their length. If the sections were considerably short (one paragraph) and continued with a related topic, they were fused into a single document. Exceptions have occurred where the application of these rules were not possible. As a final step, each document was assigned a level 1-4 according to the original ESO grade for which the material was developed. Note that we discarded duplicated documents and created separate files for each document and its metadata (grade, keywords, topic, subject and source).

The Basque Corpus: BasqueARA

It was Basque science texts that were the most difficult to gather. In fact, we were only able to collect 17 documents from Agrega2. Given this highly limited number of documents, we resorted to a website and science textbooks for additional text. The final BasqueARA corpus includes a total of 329 documents across the 4 ESO levels (see Table 1) (see the science topics covered in each level in Appendix A). It is an unbalanced corpus where ESO-1 has 102 documents, ESO-2 has 90 documents, ESO-3 has 58 and ESO-4 79. The corpus includes a total of 8,311 unique lemmas, 52,459 words, 6,269 sentences and 4,238 paragraphs. We observe that, rather counterintuitively, the average number of words per document is highest for ESO-1, it then drops to the lowest for ESO-2 and increases for ESO-3 and ESO-4. A similar pattern is observed in the count of unique lemmas.

To get to know the linguistic content of our corpus better, and more concretely, to examine whether texts with higher ESO grade levels displayed features that indicate more complexity, we analyzed several linguistic features using MultiAzterTest (Bengoetxea & Gonzalez-Dios, 2021). MultiAzterTest is a tool that provides indices of the linguistic and discourse representations of a text. Specifically, it extracts well over 100 features for Basque, Spanish and English (125, 141 and 163, respectively). We automatically obtained all indices for the texts per ESO grade and examined trends. For enhanced comparison and readability, we present values only for the ESO1 and ESO4 levels.

Results did not show clear uniform progressions of all indices from ESO-1 to ESO-4. We did observe that the length of the words increases with grades ESO1: 6.32 -

Table 1 Quantitative information for the BasqueARA where *#docs* refers to the number of documents, *#lemmas* refer to number of unique lemmas, *#words* to the number of *words, w. avg.* to the average number of words per document and *w. st.dev.* to the standard deviation

language	level	#docs	#unique lemmas	#words	w. avg.	w. st.dev.
Basque	1	102	4,394	22,800	223.52	159.74
	2	90	2,406	8,872	98.57	85.97
	3	58	2,315	7,989	137.74	91.50
	4	79	3,342	12,798	162.00	148.89
Total		329	8,311	52,459	159.45	137.73

Table 2 Quantitative information for the Agrega2-Es corpus, where *#docs* refers to the number of documents, *#lemmas* refer to number of unique lemmas, *#words* to the number of *words, w. avg.* to the average number of words per document and *w. st.dev.* to the standard deviation

language	level	#docs	#unique lemmas	#words	w. avg.	w. st.dev.
Spanish	1	102	6,137	50,314	493.27	440.57
	2	78	4,729	35,522	455.41	345.20
	3	94	3,785	31,789	338.18	178.45
	4	127	4,993	54,232	427.02	167.21
Total		401	12,438	171,857	428.57	302.14

ESO4: 6.96), for example, which is a sign of complexity. However, other indices were not in line with this trend. To mention a few, rare adjectives (ESO1: 0.99 - ESO4: 0.79) and adverbs (ESO1: 0.45 - ESO4: 0.25) are mostly present in ESO-1 documents.

The Spanish Corpus: Agrega2-Es

For the Spanish corpus, we were able to create a total of 401 documents from material extracted from the Agrega2 repository (see Table 2). In reference to the number of documents assigned to each level, we see that the corpus is not completely balanced. It contains 102 documents for ESO-1, then drops to 78 for ESO-2, and then increases to 94 for ESO-3 and 127 for ESO-4. The corpus includes a total of 12,438 unique lemmas, 171,857 words, 12,058 sentences and 7,186 paragraphs (see the science topics covered in each level in Appendix A). The average number of words per document is highest for ESO-1 and the lowest for ESO-3. Notably, the distribution of unique lemmas shows a decreasing pattern until ESO-3 and an increasing pattern in the ESO-4 level.

The linguistic analysis carried out based on the features provided by MultiAztertest revealed that our Agrega2-Es corpus has both regular and irregular progressions with respect to difficulty, and the scores do not systematically assign a higher level of complexity to documents in the higher ESO grades, as one would expect. For example, at the syntactic level, the average value of left embeddedness (ESO1: 2.74 - ESO4: 4.12), that is, the number of words before the main verb, increases as the grades progress, which can be taken as an indication that documents in higher grades are more difficult. At the semantic level, the polysemic index (ESO1: 4.90 - ESO4: 5.41), the average of the polysemy values of nouns and verbs calculated according to WordNet entries, also increases as the ESO level increases. However, hypernymy values also show that there are more general terms in higher levels (ESO1: 6.39 - ESO4: 6.71), which would indicate that the concentration of more general words is higher in the upper levels of ESO.

Given that traditional readability formulas are available for Spanish, we examined how they would classify the documents in our corpus. We used the Fernández Huerta (1959) metric, an adaptation of the Flesch-Kincaid formula for Spanish. According to the results, all the documents in the Agrega2-Es corpus belong to the 8th and 9th grades in the USA school system (see Fig. 1). In other words, all the documents can

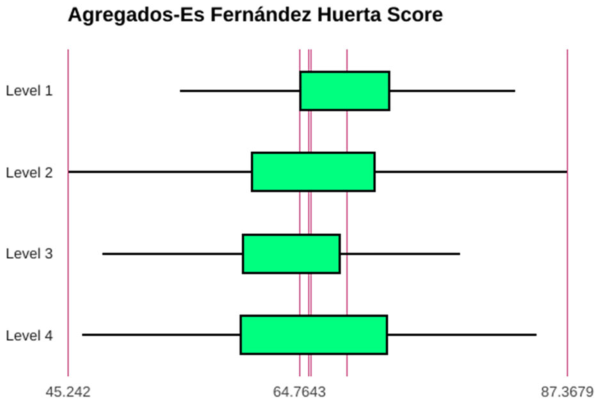

Fig. 1 Box plot of Fernandez Huerta scores for the Agrega2-Es corpus documents per grade

be easily understood by 13 to 15 year-old students. This allocation is in line with the actual grade assignment obtained from our source: Agrega2-Es material is for 12-16 year-old students in Spain but the differences among ESO levels are not clearly differentiate with this metric.

The English Corpus: Agrega2-En+

For English, we gathered a total of 251 documents, 97 of which came from the Agrega2 project. This material was extracted from science units in English. The Agrega2 project is the source of 60 documents in ESO-1, 2 in ESO-2, and 35 documents in ESO-4. In order to conduct the experiments, we incorporated texts from other web and proprietary sources to the pool of documents. We assigned the new documents to the ESO 1-4 levels according to the indications in the books. In total, our Agrega2-En+ corpus consists of 251 documents (see Table 3), 81 documents for ESO-1, 61 for ESO-2, 52 for ESO-3 and 57 for ESO-4. The corpus includes a total of 5,792 unique lemmas, 76,802 words, 7,099 sentences and 4,042 paragraphs (see the science topics covered in each level in Appendix A). ESO-4 level has the highest number of average words

Table 3 Quantitative information for the Agrega2-En+ corpus, where *#docs* refers to the number of documents, *#lemmas* refer to number of unique lemmas, *#words* to the number of *words, w. avg.* to the average number of words per document and *w. st.dev.* to the standard deviation

language	level	#docs	#unique lemmas	#words	w. avg.	w. st.dev.
English	1	81	2,628	24,954	308.02	176.03
	2	61	2,429	17,028	279.14	192.65
	3	52	2,201	14,554	279.88	164.95
	4	57	2,400	20,266	355.54	160.00
Total		251	5,792	76,802	305.98	176.03

meanwhile ESO-2 level has the least. Contrarily, number of unique lemmas are mostly abundant in ESO-1 level and ESO-3 level has the least number of unique lemmas.

We conducted the linguistic analysis also on the Agrega2-En+ corpus. We found even fewer complexity-related patterns across ESO levels in this corpus, and noticed that it was often the documents pertaining to the ESO-3 category that were the ones displaying a different behaviour. There is clearly a higher presence of passive voice verbs phrases as the ESO levels increase (ESO1: 13.29 - ESO4: 13.05). However, no obvious patterns emerge for the remaining syntactic features. For left embeddedness, for example, the highest and lowest values belong to ESO-1 and ESO-4 levels (ESO1: 3.63 - ESO4: 3.46), respectively.

According to Flesch-Kincaid scores, the documents in the Agrega2-En+ corpus contains texts ranging from 8th to 12thth grades in the USA school system (see Fig. 2). That is, there are groups of documents that can be easily understood but others are fairly difficult to read. However, this distinction does not correlate with the ESO levels.

To sum up, BasqueARA, Agregados-Es, and Agregados-En+ are entirely independent collections, each comprising 329, 401, and 251 documents, respectively. The corpora were compiled with material originally written in their respective language and they are not translations of one another. Moreover, Agregados-En+ consists of documents tailored to the proficiency levels and language learning requirements of non-native English speakers. The linguistic analysis carried out revealed few patterns across grades. This suggests that the benefit of these particular linguistic features might be limited for the ARA models for classification.

Model Architecture

In our work, we approached ARA as a document level classification task. While multiple algorithms are available to train the models, we opted to study the behaviour of feature-based ML and DL models. In this section, we present the main architectures used in the experiments for Basque, Spanish and English.

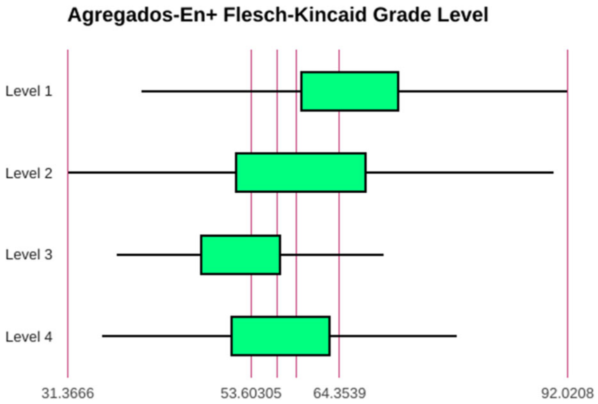

Fig. 2 Box plot of Flesch-Kincaid scores for the Agrega2-En+ corpus documents per grade

Feature-Based Models

SVM is the prevailing approach over all alternatives in the studies carry out in the last few years. Therefore, we follow the same approach in our experiments for classifying our data. Concretely, we used WEKA (Hall et al., 2016) to build an SVM classifier using SMO for each of our languages. SMO is an optimization method used to address quadratic optimization issues that come up during SVM training. We extracted the features to train the models using MultiAzterTest (Bengoetxea & Gonzalez-Dios, 2021), which is a tool that calculates linguistic features from texts in Basque, Spanish and English. The linguistic features are not limited to readability formulas, specifically, Flesch-Kincaid Grade Level and Simple Measure of Gobbledygook (SMOG), but include a wide set of linguistic phenomena that might help the identification of textual complexity. The features capture descriptive characteristics, lexical diversity, readability ability, word frequency, vocabulary knowledge, word information, syntactic complexity, word semantic information, referential cohesion, semantic overlap and connective elements. Linguistic features with no equivalents in Basque and Spanish are not calculated for these languages. These language-specific features include: agentless passive voice incidence[4] and gerund incidence[5] Note that the readability features include the Flesch-Kincaid grade level formula and the SMOG index, and these are also not available for Basque. In total, the tool provided us with 125 features for Basque, 141 features for Spanish and 163 features for English.

Deep Learning Models

Researchers have reported improved accuracy with the advent of transformer-based DL models over feature-based ML methods. These models have shown great capacity to be fine-tuned for particular tasks after being pre-trained as language models. Instead of training a model from scratch, fine-tuning leverages the knowledge gained from the pre-trained language model and applies it to a more specialized task, such as ARA, to make the model more effective in this domain. This transfer learning task has been successfully applied in a variety of NLP tasks with language models such as BERT (Devlin et al., 2019). One of the main limitations of this type of models when working with document-level texts is that the number of input byte-pair tokens cannot exceed 512 tokens. Until recently, the options available to work with longer texts involved either to only use the first 512 tokens of the document or to split it in groups of 512 tokens, classify each of them separately, and then combine the results. However, newer architectures such as Longformer (Beltagy et al., 2020) deal with sequences of up to 16 K tokens.

The average number of words per document in our BasqueARA is 159.44 with a standard deviation between 85 and 148, meanwhile, in the Agrega2-Es corpus, the

[4] In English passive voice is used where there is no agent completing the action, in Spanish and Basque this use and form may differ.

[5] In English gerunds are forms ending with -ing and they function as nouns, even though the concept of gerund exists in Basque and Spanish the form and function is language-specific.

average is 428 with a standard deviations between 167 and 440. The Agrega2-En+ corpus has an average of 305.98 words per document with a standard deviation between 120 and 196. As a result, the high number of words in our documents cannot be fully represented using a BERT-style language model. In this exploratory work, we decided to test both types of architecture and therefore we ran our experiments with both transformed-based pre-trained models, the BERT-based and the Longformer-based models. Overall, our aim is to take a pre-trained language model and fine-tune it with our data for ARA.

BERT-Based Models

Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained language model for NLP tasks such as text classification, question answering or sentiment analysis (Devlin et al., 2019). BERT is based on a transformer architecture and trained on a plain text corpus. Originally trained to represent contextual relations between words in English, new models trained on Basque and Spanish data have also emerged. The English model, BERT, was pre-trained on the BooksCorpus (Zhu et al., 2015) (800 million words) and English Wikipedia (2,500 million words). BERTeus (Agerri et al., 2020), the model for Basque, was trained using the Basque Media Corpus, which contains 189.6 million tokens from crawled news articles and 35 million from the Basque Wikipedia. The Spanish model, BETO (Cañete et al., 2020), includes 3 billion words from different sources: all the data from the Spanish Wikipedia and the OPUS Project (Tiedemann, 2012), which includes journals, TED talks, subtitles, news stories and more. All three models have similar configurations containing 12 transformer encoder layers and 110 million parameters (for further details see Devlin et al. (2019); Agerri et al. (2020); Cañete et al. (2020)). In our experiments, the three models have a maximum length limit and they truncate longer sequences automatically, taking the first 512 tokens of each document in the corpus as input.

Longformer-Based Models

Beltagy et al. (2020) introduced Longformer, a new architecture with an attention mechanism that scales linearly with sequence length, in contrast to previous approaches whose self-attention scaled quadratically. As a consequence, this architecture makes it easy to process long documents. Beltagy et al. (2020) also offer a pre-trained language model that can process English sequence lengths of up to 4,096 tokens. Similarly, XLM-R Longformer (XLM-Long for short) (Sagen, 2021) is an extension of the XLM-R multilingual model (Conneau et al., 2020) which was pre-trained using the Longformer scheme only on the English WikiText-103 corpus. As a result, XLM-Long offers the option to work with long texts in several languages, including Basque and Spanish. In our experiments, Longformer-based models take as input the whole documents from the training corpora.

Science Models

In this section we present the feature-based ML models and the DL models we developed using the BasqueARA, Agrega2-Es and Agrega2-En+ corpora. We first describe the experimental set-up and then report the evaluation results to determine the level of reliability of our models, trained with context-specific corpora, for our multilingual application scenario.

Experimental Set-up

We trained feature-based ML models and DL models for each of the three languages involved in our target educational setting using the model architectures described in Section 4.

A review of the literature showed that the higher the number of classes to be predicted, the less accurate ARA models tend to be (Ma et al., 2012). Our corpora are relatively small and this, in itself, might increase the difficulty of the task. To test for differences in behaviour, we trained models on the four original ESO levels as well as on a modified 2-level setup, where levels 1 and 2, and 3 and 4 were merged, reflecting the two stages comprised by ESO. Additionally, as described in Section 3, our corpora are considerably unbalanced and this might bias the models during training. To avoid this, we also trained the models on the full unbalanced corpora as well as on a balanced sample obtained by randomly selecting the common minimum number of documents available for all levels, that is, 58 for Basque, 78 for Spanish and 52 for English.

As mentioned, we provided the feature-based ML model with linguistic features extracted using MultiAzterTest (Bengoetxea & Gonzalez-Dios, 2021). Next, we used the SVM classifier with 10-fold cross-validation on the balanced and unbalanced configurations of the BasqueARA, Agrega2-Es and Agrega2-En+ corpus for the 4-level and 2-level setups. We report accuracy and F1 score metrics to evaluate the models.

The pre-trained language models (BERT-based and Longformer-based ones) were also fine-tuned to classify the documents in the 4-level and 2-level setups. We fine-tuned for a maximum of 20 epochs with various learning rates (1e-5, 3e-5 and 5e-5) in unbalanced and balanced corpora. As the number of training examples is not high, we followed the cross-validation approach also in this experimental setting. We split the corpora into 10 stratified training and validation folds, created a model for each fold, obtained their accuracy values per epoch and set as best model the one that obtained the best accuracy values on average. It is worth mentioning here that DL models fine-tuned on BasqueARA and Agrega2-Es with XLM-Longformer scored very poorly for all configurations and setups, with the best model achieving an accuracy of 30%, and were therefore discarded for further analysis. We only report the results for the original Longformer model which works for English.

Results

Tables 4 and 5 present the results of our experiments for the 4-level and 2-level setting, respectively. Each row of the tables displays the accuracy and weighted F1 score results

Table 4 Accuracy and weighted F1 score results for the 4-level ARA models trained with BasqueARA, Agrega2-Es and Agrega2-En+, where SVM refers to the ML approach used and BERTeus, BERT, BETO and Longfomer to the approaches used for the DL models. The highest value for each language is marked in bold

corpus	model	balance	accuracy	F1 score
BasqueARA	SVM	No	52.58%	0.51
BasqueARA	SVM	Yes	40.94%	0.41
BasqueARA	BERTeus	No	61.37% (lr 5e-5)	0.60
BasqueARA	BERTeus	Yes	**74.63%** (lr 5e-5)	**0.75**
Agrega2-Es	SVM	No	51.87%	0.51
Agrega2-Es	SVM	Yes	49.35%	0.49
Agrega2-Es	BETO	No	**72.50%** (lr 3e-5)	**0.73**
Agrega2-Es	BETO	Yes	69.77% (lr 5e-5)	0.70
Agrega2-En+	SVM	No	62.94%	0.63
Agrega2-En+	SVM	Yes	61.05%	0.61
Agrega2-En+	BERT	No	**86.01%** (lr 3e-5)	0.84
Agrega2-En+	BERT	Yes	84.02% (lr 5e-5)	0.82
Agrega2-En+	Longformer	No	83.66% (lr 5e-5)	0.84
Agrega2-En+	Longformer	Yes	84.61% (lr 5e-5)	**0.85**

obtained for a specific combination of the settings presented in the experimental set-up (corpus, learning approach, balance). For the DL models, we report the results obtained with the learning rate leading to the highest accuracy only (indicated in parenthesis).

Table 5 Accuracy and weighted F1 score results for the 2-level ARA models trained with BasqueARA, Agrega2-Es and Agrega2-En+, where SVM refers to the ML approach used and BERTeus, BERT, BETO and Longfomer to the approaches used for the DL models. The highest value for each language is marked in bold

corpus	model	balance	accuracy	F1 score
BasqueARA	SVM	No	64.13%	0.64
BasqueARA	SVM	Yes	61.20%	0.61
BasqueARA	BERTeus	No	65.65% (lr 3e-5)	0.65
BasqueARA	BERTeus	Yes	**80.60%** (lr 5e-5)	**0.80**
Agrega2-Es	SVM	No	82.75%	0.83
Agrega2-Es	SVM	Yes	83.79%	0.84
Agrega2-Es	BETO	No	**89%** (lr 1e-5,3e-5)	**0.89**
Agrega2-Es	BETO	Yes	**89%** (lr 1e-5,3e-5)	**0.89**
Agrega2-En+	SVM	No	72.11%	0.72
Agrega2-En+	SVM	Yes	77.06%	0.77
Agrega2-En+	BERT	No	86.07% (lr 1e-5)	0.85
Agrega2-En+	BERT	Yes	**86.95%** (lr 3e-5)	0.85
Agrega2-En+	Longformer	No	84.04% (lr 5e-5)	0.84
Agrega2-En+	Longformer	Yes	86.52% (lr 5e-5)	**0.87**

Please note that the tables with the precision, recall, F1 scores and weighted scores for the 4-level setup are provided in Appendix B.

For the 4-level scenario, results show that our SVM models do not perform well for any of the languages involved, Basque, English and Spanish. Concretely, the best SVM model for Basque obtains an accuracy of 52.58%. For Spanish, we obtain an accuracy of 51.87%. English obtains the models with the highest accuracy with a value of 62.94% and 61.05% for the unbalanced and balanced configurations respectively. As can be observed, F1 scores closely follow those results.

The DL counterparts achieve considerably better results. Using BertEUS on the BasqueARA corpus, we obtain an accuracy of 74.63% for Basque. We obtain a similar result, an accuracy of 73.75%, for Spanish, using BETO on the Agrega2-Es corpus. Again, we obtain the best results for English. The model built using Longformer on the Agrega2-En+ corpus obtains an accuracy of 87.33%. Interestingly, all the DL models showed an improvement of over 20 points with respect to the SVM models.

Let us consider the 4-level configuration in more detail. For the ML model, overall, the results for the unbalanced and balanced settings are similar across languages. It is only the Basque model that obtains a noticeable increase of 0.1 for the F1 score and 12% for accuracy. For this language, in the unbalanced setting, accuracy and F1 score results show better results for categories with a higher number of examples: the difference between ESO-1 and ESO-3 categories is almost 0.5 points. This trend is also present in the balanced setting, where better results (ranging from 0.414-0.521) are obtained for the ESO-1, 2 and 4, but ESO-3 scores are considerably lower with a 0.226 F1 score. Such a trend is more blurred for Spanish. The two categories with the highest number of documents obtain the best results but it is not the category with the lowest number of documents that performs worst. In fact, for both the unbalanced and the balanced settings it is for ESO-3 that performance is the worst. For English the trend is not apparent. The fluctuation of the scores is more limited and no correlation between the number of documents and accuracy and F1 score results is present.

Within the DL models, the effect of balancing the training corpora is different across languages for those trained with the BERT architecture. The outcome is very positive for Basque, but not for Spanish and English. No correlation seems to exist between the number of documents in a category and the results of the models for any of the languages. What we see is that the range of accuracy and F1 score results across categories decreases for the balanced setting, specially for Basque.

Not surprisingly, the models perform substantially better for both the feature-based ML approach and the DL approach for the 2-level setup. SVM models obtain an improvement of over 20 points for Basque and over 30 points for Spanish and English. The accuracy of the DL models increases over 15 points. For this simpler set-up, all approaches score over 60% in accuracy. Yet, DL models prevail over SVM models, the best obtaining 81%, 89% and 86.95% accuracy for Basque, Spanish and English, respectively.

With respect corpus balance, we see that results do not vary considerably for the ML models. For Basque, we obtain slightly better results for the unbalanced setting, while for Spanish and English a balanced scenario proves better. We do observe that for all languages the balance setting results in a smaller fluctuation of results across categories. The DL models follow the same trend for Spanish and English. However, BERTeus perform otherwise: while the unbalanced setting obtains a similar score to the SVM model (considerably lower than that for Spanish and English), the balanced setting performs significantly better, with an improvement of 15 points and much closer to the performance quality obtained for Spanish and English.

Overall, the results show that our feature-based ML models lag behind the DL models regardless of the language, corpus and configuration tested (the balanced and unbalanced corpus, the 2-level and 4-level set-up). The best-scoring feature-based model achieves a relatively low accuracy of 62.94% for the 4-level setup and a more favourable 77.06% for the 2-level setup for English. In turn, all DL models score over 60%.

We further inspected the results of the models to pinpoint the level-pairs with which models struggle the most. A closer look at the confusion matrices for the 4-level models showed interesting results. We observed that, for the three languages in the 4-level setup, both the feature-based ML and the DL models struggle the most when deciding whether a document belongs to level 1 or level 2. Additionally, the Basque model struggles to identify level 1 and level 3 too. No considerable differences or patterns can be observed among the remaining level-pairs for any of the models.

In an attempt to identify the features that might be misleading the feature-based models and make them classify a document as pertaining to a level to which it does not belong, we compared the average values of correctly and incorrectly classified instances for the 25 most relevant features. No specific set of features can be said to be causing the errors across level-pairs. Interestingly, level 3 of Agrega2-En+, which is the class with the highest proportion of documents collected from external proprietary resources, displayed the highest number of statistically different average values to other levels, in particular, to level 1. This leads us into thinking that while most documents allocated to level 3 share similar characteristics to those of Agrega2-En+, inconsistent instances are apparent.

Automatic Evaluation

Having trained various ARA models for Basque, Spanish and English based on context-specific corpora (see Section 5), in this section we aim to gauge their generalizability. To this end, we apply the best performing models for each of the languages to unseen data. We follow the same procedure described in Section 3 to compile the new test corpora and create three additional resources: IkasElkar for Basque, Intef for Spanish and NSC for English. It's important to note that these corpora, except Intef, come from proprietary sources.

In the next paragraphs, we first describe the new test corpora and explain their characteristics in Section 6.1 before we present the results in Section 6.2.

Corpora for Automatic Evaluation

Let us start by presenting the Basque test corpus. IkasElkar is a 4-level Basque corpus that consists of science documents collected from a private source (see details in

Table 6 Quantitative information for the IkasElkar, Intef and NSC corpora, where *#docs* refers to the number of documents, *#words* to the number of *words, w. avg.* to the average number of words per document and *w. st.dev.* to the standard deviation

test corpus	level	#docs	#words	w. avg.	w. st.dev.
IkasElkar	1	45	10,262	228.04	107.41
	2	27	5,444	201.63	82.60
	3	82	16,244	198.09	145.64
	4	105	19,922	189.73	143.05
Intef	1	74	22,705	306.82	180.50
	2	71	19,062	268.47	249.56
	3	148	29.270	197.77	134.17
	4	128	28,668	223.96	174.66
NSC	1	33	5,334	161.63	191.95
	2	74	13,490	182.29	156.57
	3	26	5.899	226.88	180.57
	4	50	10,479	209.58	155.35

Table 6). Specifically, the texts belong to a private publishing company that produces textbooks for the BAC context. We obtained 18 books in total, 3 books for level 1 and 2; and, 6 books for level 3 and 4. In total we were able to extract 259 documents for our Basque test corpus: 45 for level 1, 27 documents for level 2, 82 documents for level 3 and 105 documents for level 4 (see the science topics covered in each level in Appendix A). Note that the average word count per document is higher in the IkasElkar corpus (except for level 1) compared to the BasqueARA corpus. Yet, the standard deviation ranges are quite similar. But, in IkasElkar, the average number of words decreases slightly and rather regularly as the ESO levels go up while in BasqueARA there is no clear order.

For Spanish, we compiled the Intef test corpus. Intef is the unit of the Spanish Ministery of Education and Vocational Training responsible for the integration of information and communications technology and teachers' training in all educational stages (except university education). Intef provides educational materials for different levels of education. The corpus was compiled with material extracted from their website[6] during November 2022. We obtained science education texts for ESO level students and followed the procedure explained in Section 3 to compile our document-level corpus. In total, the Spanish test corpus consists of 421 documents: 74 for ESO-1, 71 for ESO-2, 148 for ESO-3 and 128 for ESO-4 (see the science topics covered in each level in Appendix A). When compared to Agrega2-Es, the average number of words per document is lower, with an average of 236 against 428 for the former. The standard deviation is also smaller, specially for levels 1 and 2.

For English, we compiled the new test corpus, Natural Science Corpus (NSC), from materials from private editorials during November 2022 (see Table 6). In total we collected 183 documents: 33 documents for ESO-1, 74 documents for ESO-2, 26 documents for ESO-3 and 50 documents for ESO-4 (see the science topics covered in each level in Appendix A). More regular than the Agrega2-En+ corpus, in NSC, the

[6] https://intef.es/

word average increases as the ESO level goes up. The standard deviation, however, is similar for both corpora.

In addition to confirming that the topics covered by the training and test corpora per level overlapped in a way that at least almost all topics present in the test corpora were present in the training corpora, we further inspected their differences and similarities in terms of linguistic features. We extracted the 25 most relevant features according to the ML models using MultiAzterTest (Bengoetxea & Gonzalez-Dios, 2021) and calculated their t-test values. The examination of the features yielded valuable insights into these datasets.

For Basque, features with a statistically signiticant difference consistently displayed lower means in BasqueARA compared to the IkasElkar test corpus. The BasqueARA corpus exhibited lower average counts of letters in lemmas, logical connectives, adverbs, indicative verbs, propositions, words, and lower average number of levels within the dependency tree per sentence along with a reduced standard deviation of sentence length.

For Spanish, in the features with significant differences, Agregados-Es presented a higher mean for metrics including: the count of verbs in infinitive form per 1000 words, the occurrence of proper nouns per 1000 words, the proportion of proper nouns relative to all nouns, the average sentences per paragraph, and the similarity between adjacent sentences. On the other hand, the Intef test corpus showed a higher mean for metrics such as the count of adversative connectives, low-frequency words, mean number of sentences per paragraph, and the semantic similarity among all possible sentence pairs within a paragraph.

Numerous attributes showed a statistically significant difference between Agregados-En+ and the NSC test datasets. Notably, within the Agregados-En+ corpus, attributes such as left embeddedness values, mean verb phrase per sentence, and average polysemy values of nouns and verbs displayed higher means. Conversely, the NSC corpus showed higher means in the standard deviation of lemma length, noun density, noun count, syllable count in words, and word length.

Results of Automatic Evaluation

Results show that, in general, our science models do not perform well on unseen data. Let us review their performance per language.

For Basque, the feature-based SVM model displays a decrease in both accuracy and F1 score results when tested on the IkasElkar corpus. For the 4-level scenario (Table 7) the best accuracy of the model is 35.90% (0.34 F1 score) on the balanced corpus configuration. In turn, the DL model does not show a significant improvement, with 38.61% accuracy (0.36 F1 score) on the balanced corpus. In the 2-level scenario (Table 8), the results improve considerably, over 30 points compared to the 4-level scenario. However, the DL models' accuracy drops from the 80.60% obtained in cross-validation to 70.65% for the balanced corpus configuration and to 44.40% for the unbalanced configuration.

The Spanish models trained on Agrega2-Es show a similar behaviour. The accuracy of the SVM models drops more than 10 points when tested on Intef. In turn, the DL

Table 7 Accuracy and weighted F1 score results for the ARA models tested on IkasElkar, Intef and NSC for 4-levels. SVM refers to the ML approach and BERTeus, BERT, BETO and Longfomer to the approaches used for the DL models. The highest value for each language is marked in bold

model	test corpus	balance	accuracy	F1 score
SVM	IkasElkar	No	34.36%	0.34
SVM	IkasElkar	Yes	35.90%	0.34
BERTeus	IkasElkar	No	34.55%	0.33
BERTeus	IkasElkar	Yes	**38.61%**	**0.36**
SVM	Intef	No	32.30%	0.32
SVM	Intef	Yes	37.29%	**0.38**
BETO	Intef	No	**37.52%**	0.37
BETO	Intef	Yes	28.26%	0.24
SVM	NSC	No	34.42%	0.35
SVM	NSC	Yes	**36.61%**	**0.38**
BERT	NSC	No	23.49%	0.15
BERT	NSC	Yes	28.41%	0.14
Longformer	NSC	No	**36.61%**	0.34
Longformer	NSC	Yes	31.14%	0.26

models also perform much worse, with their accuracy decreasing over 30 points. Both approaches perform similarly: the accuracy of the best SVM model is 37.29% while the accuracy of the best DL model is 37.52% (note the F1 score assigns a 0.38 to the former and 0.37 to the latter). The models perform better in the 2-level configuration even if we observe a sharp drop from the results obtained with the cross-validation results. The SVM models perform better than the DL models. For this scenario, the best SVM model obtains an accuracy of 61.28% and the best DL model a 45.84% (0.62 and 0.39 F1 scores, respectively).

Table 8 Accuracy and weighted F1 score results for the ARA models tested on IkasElkar, Intef and NSC for 2-levels. SVM refers to the ML approach and BERTeus, BERT, BETO and Longfomer to the approaches used for the DL models. The highest value for each language is marked in bold

model	test corpus	balance	accuracy	F1 score
SVM	IkasElkar	No	67.18%	0.69
SVM	IkasElkar	Yes	68.33%	0.69
BERTeus	IkasElkar	No	44.40%	0.41
BERTeus	IkasElkar	Yes	**70.65%**	**0.71**
SVM	Intef	No	**61.28%**	**0.62**
SVM	Intef	Yes	60.80%	0.61
BETO	Intef	No	45.84%	0.39
BETO	Intef	Yes	45.60%	0.38
SVM	NSC	No	50.81%	0.51
SVM	NSC	Yes	46.44%	0.50
BERT	NSC	No	43.16%	0.29
BERT	NSC	Yes	43.16%	0.28
Longformer	NSC	No	**73.77%**	**0.73**
Longformer	NSC	Yes	55.73%	0.53

The English models, trained on Agrega2-En+, show the highest drop when compared to previous experiments. The best SVM model obtains an accuracy of 36.61% while the best DL model only achieves an accuracy of 36.61%. This is a high contrast with respect the results of previous experiments, where the best model achieved an accuracy of 86.01% and even the worse model obtained a 61.05% (F1 scores indicate the same trend). The results improve for the 2-level scenario but, again, it is only one of the Longformer-based models that reaches 73.77% of accuracy (0.73 F1 score).

If we focus on the effect of corpus balance, we observe that, following the trend for the cross-validation results for the 4-level setting in Section 5.2, for the ML models, Spanish and English obtain better results with the balanced model. For Basque, the apparent benefit of the unbalanced model from the cross-validation results disappears and both models obtain very similar overall results, with a slightly better accuracy for the balanced model (not for the F1 score). The DL approach trained with the corresponding BERT architectures shows similar trends for Basque and Spanish, but not for English. Basque results are better when using the balanced model but the improvement is not as high as we observed in the cross-validation experiment. Spanish does not benefit from the balanced setting, but the decrease is slightly smaller as compared to the cross-validation experiment. For English, using the balanced corpus results in improved results, which was not the case in the cross-validation experiment or with the Longformer model.

As expected, once again, the models perform substantially better for both the feature-based ML approach and the DL approach for the 2-level setup. SVM models obtain an improvement of around 30 points for Basque and Spanish, and 15 for English. The results of the DL models increase around 15 points for Basque and English, but not for Spanish. In fact, the balance (or lack of it) of the corpus yields divergent results for DL models. For example, for Basque, balancing the corpus increases the results of the BERTeus model in almost 25 points, but the results are very similar for Spanish and English. Then, for the Longformer model, in English, the balanced setting results in a decrease of 18 points. These findings emphasize the necessity for more comprehensive data analysis. The behavior of the models gives rise to thoughts about further investigation and exploration with a larger and more extensive dataset.

Models Trained on Data for Native Learners

From our experiments, data scarcity and domain specificity, among others, appears as an obstacle to build high-performing ARA models. A review of available science

Table 9 Quantitative information for the Siyavula corpus, where *#docs* refers to the number of documents, *#words* to the number of *words, w. avg.* to the average number of words per document and *w. st.dev.* to the standard deviation

level	#docs	#words	w. avg.	w. st.dev.
1	42	38,850	934.00	619.74
2	48	36,548	761.41	574.98
3	74	49,736	672.10	513.90
4	133	87,865	660.63	642.94
Total	297	212,999	718.45	603.41

Table 10 Accuracy and weighted F1 score results of the 4-level ARA models for English trained with texts for native learners (Siyavula corpus) and tested on texts for non-native learners (NSC corpus), where SVM refers to the ML approach and BERTeus, BERT, BETO and Longfomer to the approaches used for the DL models. The highest value is marked in bold

training corpus	test corpus	model	balance	accuracy	F1 score
Siyavula	NSC	SVM	No	26.22%	0.20
Siyavula	NSC	BERT (lr 1e-5)	No	26.22%	0.15
Siyavula	NSC	BERT (lr 3e-5)	No	11.47%	0.04
Siyavula	NSC	BERT (lr 5e-5)	No	14.20%	0.03
Siyavula	NSC	Longformer (lr 1e-5)	No	**28.41**%	0.14
Siyavula	NSC	Longformer (lr 3e-5)	No	27.86%	0.14
Siyavula	NSC	Longformer (lr 5e-5)	No	27.86%	0.14

educational data revealed very limited resources. Our efforts to gather resources for our specific multilingual context also yielded little data. Adding to this sparseness issue, we note that the target users of the collected documents are not always the same. Educational English resources used in the literature were created with native English learners in mind, whereas our English corpora contain text for non-native English speakers. In this scenario, we performed yet another experiment to check whether documents for our ESO grades students could be successfully classified using a model trained on data for native speakers.

To do so, we ran an experiment with the texts coming from the Siyavula project[7], which is an educational initiative with the aim of creating materials in natural sciences and mathematics for high school students. Originating within the South African curriculum, it provides online materials and open access textbooks primarily for native speakers of English. Siyavula proved successful to train a paragraph-level ARA model in a previous study (Nadeem & Ostendorf, 2018). Therefore, if the same corpus could be useful for our non-native context and at document level will be analyzed. To prepare the corpus for it, we downloaded the natural science books for grades 7-10, which are the ones coinciding with the learner ages for the four ESO grades we are focusing on. We then followed the same steps as explained in Section 3 to divide the text into documents.

The compiled corpus consists of 297 documents (212,999 words) distributed across 4 levels (see Table 9). Once again, the number of documents allocated to each level vary, with level 4 gathering the highest number of instances with 133, and level 1 containing the fewest with 42.

We trained the feature-based ML and the DL models following the same procedure as in Section 5. Again, for the feature-based ML model we extracted linguistic features using MultiAzterTest. To build the DL models, we fine-tuned BERT and Longformer-

[7] https://www.siyavula.com/read

Table 11 Accuracy and weighted F1 score results of the 2-level ARA models for English trained with texts for native learners (Siyavula corpus) and tested on texts for non-native learners (NSC corpus), where SVM refers to the ML approach and BERTeus, BERT, BETO and Longfomer to the approaches used for the DL models. The highest value is marked in bold

training corpus	test corpus	model	balance	accuracy	F1 score
Siyavula	NSC	SVM	No	47.54%	0.37
Siyavula	NSC	BERT (lr 1e-5)	No	41.53%	0.24
Siyavula	NSC	BERT (lr 3e-5)	No	41.53%	0.24
Siyavula	NSC	BERT (lr 5e-5)	No	41.53%	0.24
Siyavula	NSC	Longformer (lr 1e-5)	No	**44.80%**	0.31
Siyavula	NSC	Longformer (lr 3e-5)	No	43.16%	0.29
Siyavula	NSC	Longformer (lr 5e-5)	No	43.16%	0.27

based models. For the cross-validation evaluation, once again, the DL models obtain considerably better results in both setups and all the models obtain values ranging from 77% to 88% accuracy. The best accuracy values for the SVM models ranges from 64.98% to 87.87%.

We tested our Siyavula models on the NSC test corpus. Results show a dramatic drop in accuracy and weighted F1 score results for all models (see Tables 10 and 11). For the 4-level setup, the Longformer model (lr 1e-5) scores best with an accuracy of 28.41%. The results for the 2-level setup are better, as happened in previous experiments, but still low, as none of the models reach an accuracy of 50%. The best performing model is the feature-based SVM model, which obtains an accuracy of 47.54%. These results are in line with the results obtained when testing the models trained with our context-specific data for non-native learners of English. In both cases, the scores are low, and in this particular experiment, even slightly lower.

Conclusions and Future Work

In this paper we have explored the performance of feature-based ML and DL models to predict the educational grade of science texts for the Basque curriculum, that is, a multilingual educational context where students require texts in Basque, Spanish and English. To that end, we first compiled three graded corpora for secondary education learners consisting of science documents extracted from material created for our particular context: BasqueARA for Basque, Agrega2-Es for Spanish and Agrega2-En+ for English. Each document was assigned a level 1-4 according to the original ESO grade. Next, we developed feature-based ML models (SVM models) and DL models (BERT-based and Longformer-based models) to test their performance in our context and on unseen data. Additionally, we also tested the performance of an English model trained on texts directed at native English learners (the Siyavula corpus) to classify unseen data for our non-native English learners.

The within-corpus experiments showed promising results. For the 4-level classification, cross-validation experiments for Basque, on the BasqueARA corpus, yielded an accuracy of 74.63%; for Spanish, on the Agrega2-Es corpus, the models reached an accuracy of 73.75%; and for English, on the Agrega2-En+ corpus, 87.33%. The best scoring models were DL models in all cases. The accuracy and F1 score results were even higher for the 2-level classification task, best scoring models reaching well above 80% of accuracy for all languages.

However, results on unseen data showed that the science models created using our context-specific corpora do not generalize well. The accuracy and F1 score when tested with the new context-specific corpora yielded results of around 36-38% for the 4-level setup. Interestingly, the scores for the ML and DL models did not differ much. The scores were higher for the 2-level setup. The best performing language was Basque, with the best performing model (a DL model based on BERTeus) obtained an accuracy of 71.13%. For Spanish, SVM models scored about 15 points higher than the DL models, with the accuracy at 61.28%. The English models obtained the worst results, with the best model (SVM) yielding 50.81%.

The cross-corpus experiments performed using the Siyavula corpus to test the performance of models trained on data directed at native speakers to classify data directed at non-native learners also showed poor results. The results of the models are lower than those obtained with training data directed at non-native speakers. In this final experiment, the best scoring model was a Longformer DL model, which obtained an accuracy of 28.41% for the 4-level setup, and the SVM model, which obtained a 47.54%, for the 2-level setup.

The results highlight the importance of the availability of data when building ARA models. The training sets available to perform the task at hand are small, which affect the performance of the models negatively. To overcome this issue, it is essential to carefully consider strategies to enlarge training corpora. Automatic translation of existing data could be a path that deserves exploring. Automated translation, especially in ARA, is a promising avenue for exploration. However, it is important to consider potential disadvantages. Even though today MT models can provide translations of rather good quality, the generated texts might not always be entirely appropriate, that is, they might include errors. Research would have to be performed to determine the trade-off between quality and quantity of the training corpora used by ARA models.

In addition, our experiments seem to indicate that the task itself is not an easy one in the sense that it is not yet clear what the role linguistic features and the scientific content play in helping differentiate documents per grades for science texts. A more thorough analysis of the characteristics of the texts might shed some light into this issue. It is possible that there might be additional factors, such as content, that can help to extract readability level of scientific documents. Finally, it might be worth testing the multilingual DL models which exploit information from different languages to perform the classification and the big pre-trained models such as FLAN (Wei et al., 2022) or PaLM (Chowdhery et al., 2022) recently published to address our task, as they have shown good results in a variety of NLP-related tasks when having few examples for fine-tuning.

Appendix A

Table 12 Science topics covered in the BasqueARA corpus

Level	Topic	# Documents
ESO 1	Universe	10
	The Earth	4
	Geosphere	11
	Rocks	12
	Atmosphere	10
	Hydrosphere	14
	Living things	15
	Solar system	12
	Misc	14
ESO 2	The Earth	14
	Hydrosphere	4
	Living things	17
	Energy	12
	Heat	6
	Light	8
	Matter	7
	Biosphere	8
	Misc	14
ESO 3	Living beings	26
	Matter	5
	Chemical reactions	22
	Misc	5
ESO 4	Rocks	5
	Energy	1
	Heat	1
	Cell biology	19
	Cell cycle and reproduction	5
	Genetics	8
	Evolution	5
	Ecology	7

Table 13 Science topics covered in the Agregados-Es corpus

Level	Topic	# Documents
ESO 1	Universe and solar system	12
	Vertebrate animals	12
	Invertebrate animals	5
	Minerals and rocks	6
	Beaches	5
	Living beings	10
	Rocks	4
	Atmosphere	2
	Biosphere	7
	Hydrosphere	8
	Matter	5
	Plants and fungi	8
	Mixtures and substances	5
	No theme	13
ESO 2	Natural environment of Andalusia	7
	Aquatic and terrestrial ecosystems	6
	Energy flow in living beings	12
	Energy and environment	7
	Light and sound	14
	Nutrition functions in animals	7
	Matter and energy	7
	Ecosystems	5
	Energy transfer: heat	4
	Vital functions, reproduction	7
	Bodies in motion	1

Table 13 continued

Level	Topic	# Documents
ESO 3	Atom and atomic models	8
	Scientific work	6
	Modern atom	7
	Matter elements	1
	Electrical phenomena and circuits	16
	Electricity, practical applications	12
	Forms of relief	8
	Chemical reactions	21
	Pure substances and mixtures	1
	Blank	10
	Allergies	1
	PISA	3
ESO 4	Waves	16
	Atomic structure and chemical elements	16
	Chemical transformations	8
	Heat and temperature	4
	Carbon compounds	16
	Heat and energy	13
	Circular motion	6
	Pressure	2
	Forces and pressures in fluids	9
	Astronomy and Universal Gravitation	11
	Forces	7
	Work and energy	6
	Terrestrial gravitational field	4
	Linear momentum	3
	Carbon chemistry	5

Table 14 Science topics covered in the Agregados-En+ corpus

Level	Topic	# Documents
ESO 1	Beaches	5
	Matter	25
	City rocks	1
	Health	2
	Trees	1
	Earth and the universe	17
	Atmosphere	2
	Rocks	1
	Blank	18
	Changes in water	1
	Changes of state	1
	Relative movement of the Earth	1
	Rotation of the Earth	1
	Earth orbit	1
	Planets	2
	Zodiac signs	1
	Solar system	1
ESO 2	Blank	14
	Moving bodies	2
	Earth’s internal energy	3
	Ecosystems	2
	Vital functions	2
ESO 3	Atomic theories	4
ESO 4	Blank	2
	Statics	2
	Earth’s gravitational field	4
	Uniform Circular Movement	3
	Motion Physics	1
	Trajectory and displacement	2
	Heat and Temperature	6
	Rectilinear Movement	5
	Undulatory phenomena	5
	Linear momentum	3
	Optics	3
	Work, Power and Energy	2

Table 15 Science topics covered in the IkasElkar corpus

Level	Topic	# Documents
ESO 1	The universe	13
	The Earth	12
	Living beings	20
ESO 2	Recycling	9
	Properties of materials	11
	Energy	7
ESO 3	Cell and ecosystem	10
	Health and life	9
	Ecosystems	11
	Molecules	32
	Electrons	20
ESO 4	Earth’s changes	12
	Evolution and genetics	17
	Ecosystems	13
	Accelerate on the path of inspiration	16
	Creativity	34
	Molecules	13

Table 16 Science topics covered in the Intef corpus

Level	Topic	# Documents
ESO 1	Energy systems on Earth's surface	3
	Slope processes	4
	External geological processes	4
	Soil	3
	Continental water	4
	Streams	1
	Rivers	4
	Groundwater	2
	Glaciers	4
	Wind	2
	Ocean movements	1
	Coastal Modeling	3
	Deltas	1
	External geological agents	31
	Earth's external energy	18
	Human impact on ecosystems	8
	Locomotor system	14
	Systems involved in nutrition	24
	Human reproduction	8
	Endocrine system	5
	Nutrition and health	4
ESO 2	Living beings and environment	16
	Energy transfer and ecosystems	7
	Species maintenance	12
	Functions of living beings	8
	Minerals and rocks	10
	Planetary internal energy	11
	Scientific method	6
	Blank	1

Table 16 continued

Level	Topic	# Documents
ESO 3	Atom and atomic models	8
	Scientific work	6
	Modern atom	7
	Matter	1
	Electric phenomena and circuits	16
	Electricity Practical Applications	12
	Landforms	8
	Chemical reactions	21
	Pure substances and mixtures	1
	Blank	10
	Allergies	1
	PISA	3
ESO 4	The Universe	6
	Matter and energy in Universe	1
	Origin of Universe	1
	Solar system	4
	Earth	2
	Weather	1
	Rocks	1
	History	7
	Natural environment	8
	Tectonic plates	7
	Rocks	11
	Seismic movements	7
	Living being	16
	Reproduction	26
	Genetics	17
	Chemical theory of origin of life	1
	Evolution	7
	Ecosystem dynamics	2
	Matter, energy, and life	1
	Population dynamics	1
	Biogeochemical cycles	1

Table 17 Science topics covered in the NSC corpus

Level	Topic	# Documents
ESO 1	Vertebrates	1
	Solar system planets	1
	Galaxy types	1
	Moon phases	1
	Earth's spheres	4
	Five kingdoms	1
	Monera, protoctists, fungi and plants	1
	Invertebrates	1
	Biodiversity	2
	Atmosphere	7
	Hydrosphere	1
	Minerals and rocks	5
	Matter properties	3
	Matter states	3
	Universe ideas	1
ESO 2	Living organisms	19
	Interaction function	3
	Reproduction function	17
	Ecosystems	11
	Earth's structure	8
	Earth's dynamics	10
	Forces and movements	1
	Energy and its forms	1
	Light and sound	4

Table 17 continued

Level	Topic	# Documents
ESO 3	Blank	1
	Human body	4
	Feeding and nutrition	3
	Nutrition systems	4
	Interaction function	1
	Human reproduction	4
	Health and illness	2
	Energy changes and Earth	3
	Surface processes	2
	Human interaction and environment	3
ESO 4	Cell	5
	Genetic information	4
	Trait transmission	7
	Genetic engineering	3
	Origin of life and evolution	8
	Ecosystems	1
	Ecosystems: Matter and Energy	9
	Lithospheric dynamics	1
	Landform evolution	6
	Earth history	6

Table 18 Science topics covered in the Siyavula corpus

Level	Topic	# Documents
ESO 1	Potential and kinetic energy	4
	Acids, bases and neutral substances	3
	Heat: Energy transfer	5
	Sexual reproduction	2
	Variation	2
	Periodic table of elements	3
	National electricity supply system	2
	Biosphere	5
	Separating mixtures	3
	Sun-Earth relationship	3
	Energy sources	1
	Moon-Earth relationship	3
	Properties of materials	2
	Historical development of astronomy	2
	Heat insulation and energy saving	2
ESO 2	Microorganisms	3
	Looking into space	2
	Visible light	6
	Photosynthesis and respiration	2
	The solar system	3
	Energy transfer in electrical systems	3
	Beyond the solar system	4
	Chemical reactions	2
	Environment interactions and interdependence	7
	Series and parallel circuits	3
	Particle model of matter	8
	Atoms	4
	Static electricity	1

Table 18 continued

Level	Topic	# Documents
ESO 3	Safety with electricity	2
	Human bodys systems	8
	Energy and national electricity grid	3
	Human reproduction	3
	Circulatory and respiratory systems	3
	Cost of electrical energy	2
	Earth as a system	1
	Digestive system	2
	Lithosphere	2
	Compounds	3
	Mining of mineral resources	6
	Cells and life	3
	Forces	3
	Chemical reactions	3
	Atmosphere	5
	Reactions	1
	Birth, life and death of a star	3
	Acids, bases and the pH value	6
	Resistance	3
	Series and parallel circuits	2
ESO 4	Classification of matter	6
	Chemistry of life	8
	Matter states and kinetic molecular theory	2
	Atom	5
	Basic life units	3
	Life sciences	5
	Periodic table	2
	Chemical bonding	5
	Waves	12
	Cell division	3
	Sound	2

Table 18 continued

Level	Topic	# Documents
	Electromagnetic radiation	4
	Plant and animal tissues	5
	Physical and chemical change	2
	Plant support and transport systems	3
	Representing chemical change	1
	Magnetism	2
	Animal support systems	6
	Electrostatics	3
	Electric circuits	5
	Animal transport systems	3
	Reactions in aqueous solution	4
	Skills for science	4
	Biospheres to ecosystems	7
	Quantitative aspects of chemical change	4
	Biodiversity and classification	3
	Vectors and scalars	3
	Motion in one dimension	7
	History of Life on Earth	5
	Mechanical energy	4
	The hydrosphere	5

Appendix B

We utilized scikit-learn's machine learning library (version 1.3.2) Pedregosa et al. (2011) to obtain the metrics. More specifically, we used the function *sklearn.metrics.precision_recall_fscore_support*. The parameter 'average' is set as *weighted*, which, according to their definition, *"calculates metrics for each label, and find their average weighted by support (the number of true instances for each label). This alters macro to account for label imbalance; it can result in an F-score that is not between precision and recall."* We have employed this option to calculate the overall weighted values. Simultaneously, we have set the parameter to *None*, so that the scores for each class are returned.

Table 19 Precision (Pre.), Recall, F1 scores and Weighted (W) scores for the BasqueARA 4-level unbalanced (left) and balanced (right) feature-based ML model (cross-validation)

Pre.	Recall	F1	Level	Pre.	Recall	F1	Level
0.639	0.745	0.688	a = 1	0.508	0.534	0.521	a = 1
0.511	0.522	0.516	b = 2	0.482	0.466	0.474	b = 2
0.243	0.155	0.189	c = 3	0.228	0.224	0.226	c = 3
0.506	0.519	0.513	d = 4	0.414	0.414	0.414	d = 4
0.502	0.526	0.511	weighted	0.408	0.409	0.409	weighted

Table 20 Precision (Pre.), Recall, F1 scores and Weighted (W) scores for the Agregados-Es 4-level unbalanced (left) and balanced (right) feature-based ML model (cross-validation)

Pre.	Recall	F1	Level	Pre.	Recall	F1	Level
0.564	0.559	0.562	a = 1	0.500	0.526	0.513	a = 1
0.500	0.487	0.494	b = 2	0.547	0.603	0.573	b = 2
0.456	0.330	0.383	c = 3	0.469	0.385	0.423	c = 3
0.526	0.646	0.580	d = 4	0.450	0.462	0.456	d = 4
0.514	0.519	0.512	weighted	0.491	0.494	0.491	weighted

Table 21 Precision (Pre.), Recall, F1 scores and Weighted (W) scores for the Agregados-En+ 4-level unbalanced (left) and balanced (right) feature-based ML model (cross-validation)

Pre.	Recall	F1	Level	Pre.	Recall	F1	Level
0.604	0.679	0.640	a = 1	0.549	0.538	0.544	a = 1
0.561	0.525	0.542	b = 2	0.544	0.596	0.569	b = 2
0.653	0.615	0.634	c = 3	0.618	0.654	0.636	c = 3
0.722	0.684	0.703	d = 4	0.756	0.654	0.701	d = 4
0.631	0.629	0.629	weighted	0.617	0.611	0.612	weighted

Table 22 Precision (Pre.), Recall, F1 scores and Weighted (W) scores for the BasqueARA 4-level unbalanced (left) and balanced (right) BERTeus lr3e-5 model (cross-validation)

Pre.	Recall	F1	Level	Pre.	Recall	F1	Level
0.730	0.710	0.720	1	0.720	0.760	0.740	1
0.510	0.540	0.530	2	0.770	0.690	0.730	2
0.560	0.480	0.520	3	0.790	0.840	0.820	3
0.560	0.610	0.590	4	0.700	0.690	0.700	4
0.600	0.600	0.600	weighted	0.750	0.750	0.750	weighted

Table 23 Precision (Pre.), Recall, F1 scores and Weighted (W) scores for the Agregados-Es 4-level unbalanced (left, lr3e-5) and balanced (right) BETO lr5e-5 model (cross-validation)

Pre.	Recall	F1	Level	Pre.	Recall	F1	Level
0.800	0.770	0.780	a = 1	0.770	0.660	0.710	a = 1
0.670	0.770	0.720	b = 2	0.680	0.810	0.740	b = 2
0.690	0.620	0.650	c = 3	0.710	0.560	0.630	c = 3
0.730	0.740	0.730	d = 4	0.650	0.760	0.700	d = 4
0.730	0.720	0.730	weighted	0.700	0.700	0.700	weighted

Table 24 Precision (Pre.), Recall, F1 scores and Weighted (W) scores for the Agregados-En+ 4-level unbalanced (left, lr3e-5e) and balanced (right) BERT lr5e-5e model (cross-validation)

Pre.	Recall	F1	Level	Pre.	Recall	F1	Level
0.920	0.840	0.880	a = 1	0.860	0.810	0.830	a = 1
0.760	0.820	0.790	b = 2	0.770	0.850	0.810	b = 2
0.940	0.850	0.890	c = 3	0.900	0.880	0.890	c = 3
0.770	0.860	0.810	d = 4	0.760	0.750	0.760	d = 4
0.850	0.840	0.840	weighted	0.820	0.820	0.820	weighted

Table 25 Precision (Pre.), Recall, F1 scores and Weighted (W) scores for the Agregados-En+ 4-level unbalanced (left) and balanced (right) Longformer lr5e-5 model (cross-validation)

Pre.	Recall	F1	Level	Pre.	Recall	F1	Level
0.850	0.830	0.840	a = 1	0.910	0.810	0.860	a = 1
0.750	0.770	0.760	b = 2	0.750	0.870	0.800	b = 2
0.920	0.880	0.900	c = 3	0.920	0.880	0.900	c = 3
0.850	0.880	0.860	d = 4	0.830	0.830	0.830	d = 4
0.840	0.840	0.840	weighted	0.850	0.850	0.850	weighted

Appendix C

Table 26 Precision (Pre.), Recall, F1 scores and Weighted (W) scores for the BasqueARA 4-level unbalanced (left) and balanced (right) feature-based ML model tested on IkasElkar

Pre.	Recall	F1	Level	Pre.	Recall	F1	Level
0.274	0.644	0.384	a = 1	0.276	0.711	0.398	a = 1
0.133	0.148	0.140	b = 2	0.133	0.074	0.095	b = 2
0.383	0.280	0.324	c = 3	0.410	0.415	0.412	c = 3
0.524	0.314	0.393	d = 4	0.556	0.238	0.333	d = 4
0.395	0.344	0.343	weighted	0.417	0.359	0.345	weighted

Table 27 Precision (Pre.), Recall, F1 scores and Weighted (W) scores for the Agregados-Es 4-level unbalanced (left) and balanced (right) feature-based ML model tested on Intef

Pre.	Recall	F1	Level	Pre.	Recall	F1	Level
0.246	0.432	0.314	a = 1	0.248	0.446	0.319	a = 1
0.224	0.183	0.202	b = 2	0.261	0.338	0.294	b = 2
0.448	0.291	0.352	c = 3	0.530	0.412	0.464	c = 3
0.350	0.375	0.362	d = 4	0.481	0.305	0.373	d = 4
0.345	0.323	0.323	weighted	0.420	0.373	0.382	weighted

Table 28 Precision (Pre.), Recall, F1 scores and Weighted (W) scores for the Agregados-En+ 4-level unbalanced (left) and balanced (right) feature-based ML model tested on NSC

Pre.	Recall	F1	Level	Pre.	Recall	F1	Level
0.339	0.576	0.427	a = 1	0.367	0.545	0.439	a = 1
0.500	0.324	0.393	b = 2	0.595	0.338	0.431	b = 2
0.174	0.308	0.222	c = 3	0.170	0.346	0.228	c = 3
0.364	0.240	0.289	d = 4	0.385	0.300	0.337	d = 4
0.387	0.344	0.347	weighted	0.436	0.366	0.378	weighted

Table 29 Precision (Pre.), Recall, F1 and Weighted (W) scores for the BasqueARA 4-level unbalanced (left) and balanced (right) BERTeus lr5e-5 model tested on IkasElkar

Pre.	Recall	F1	Level	Pre.	Recall	F1	Level
0.289	0.867	0.433	a = 1	0.385	0.778	0.515	a = 1
0.119	0.185	0.145	b = 2	0.153	0.407	0.222	b = 2
0.458	0.268	0.338	c = 3	0.461	0.439	0.450	c = 3
0.676	0.219	0.331	d = 4	0.889	0.152	0.260	d = 4
0.482	0.344	0.332	weighted	0.589	0.378	0.360	weighted

Table 30 Precision (Pre.), Recall, F1 scores and Weighted (W) scores for the Agregados-Es 4-level unbalanced (left) and balanced (right) BETO lr5e-5 model tested on Intef

Pre.	Recall	F1	Level	Pre.	Recall	F1	Level
0.800	0.770	0.780	a = 1	0.770	0.660	0.710	a = 1
0.670	0.770	0.720	b = 2	0.680	0.810	0.740	b = 2
0.690	0.620	0.650	c = 3	0.710	0.560	0.630	c = 3
0.730	0.740	0.730	d = 4	0.650	0.760	0.700	d = 4
0.730	0.720	0.730	weighted	0.700	0.700	0.700	weighted

Table 31 Precision (Pre.), Recall, F1 scores and Weighted (W) scores for the Agregados-Es 4-level unbalanced (left) and balanced (right) BETO lr3e-5 model tested on Intef

Pre.	Recall	F1	Level	Pre.	Recall	F1	Level
0.286	0.757	0.415	a = 1	0.243	0.905	0.383	a = 1
0.266	0.465	0.338	b = 2	0.147	0.155	0.151	b = 2
0.623	0.257	0.364	c = 3	0.500	0.189	0.274	c = 3
0.775	0.242	0.369	d = 4	0.928	0.101	0.183	d = 4
0.550	0.375	0.370	weighted	0.525	0.283	0.245	weighted

Table 32 Precision (Pre.), Recall, F1 scores and Weighted (W) scores for the Agregados-En+ 4-level unbalanced (left) and balanced (right) Longformer lr5e-5 model tested on NSC

Pre.	Recall	F1	Level	Pre.	Recall	F1	Level
0.920	0.840	0.880	a = 1	0.860	0.810	0.830	a = 1
0.760	0.820	0.790	b = 2	0.770	0.850	0.810	b = 2
0.940	0.850	0.890	c = 3	0.900	0.880	0.890	c = 3
0.770	0.860	0.810	d = 4	0.760	0.750	0.760	d = 4
0.850	0.840	0.840	weighted	0.820	0.820	0.820	weighted

Table 33 Precision (Pre.), Recall, F1 scores and Weighted (W) scores for the Agregados-En+ 4-level unbalanced (left, BERT lr3-e5) and balanced (right) BERT lr5-e5 model tested on NSC

Pre.	Recall	F1	Level	Pre.	Recall	F1	Level
0.196	0.879	0.320	a = 1	0.143	0.030	0.050	a = 1
0	0	0	b = 2	1	0.013	0.026	b = 2
0	0	0	c = 3	0	0	0	c = 3
0.400	0.280	0.329	d = 4	0.286	1	0.444	d = 4
0.145	0.235	0.148	weighted	0.508	0.284	0.141	weighted

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This work was partly supported by: University of the Basque Country UPV/EHU (PIF20/154 UPV/EHU 2020), Basque Goverment (IT1437-22 and IT1570-22) and Spanish Ministry of Science and Innovation (PID2021-127777OB-C21).

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

Agerri, R., Vicente, IS., Campos, J. A., et al. (2020). Give your text representation models some love: the case for Basque. In: Proceedings of the 12th International Conference on Language Resources and Evaluation

Arfé, B., Mason, L., & Fajardo, I. (2018). Simplifying informational text structure for struggling readers. *Reading and Writing, 31*, 2191–2210.

Azpiazu, I. M., & Pera, M. S. (2019). Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics, 7*, 421–436.

Ball, P. (2017). It's not just you: science papers are getting harder to read. *Nature*. https://doi.org/10.1038/nature.2017.21751 published on 30 March 2017

Basch, C. H., Mohlman, J., Hillyer, G. C., et al. (2020). Public health communication in time of crisis: Readability of on-line COVID-19 information. *Disaster Medicine and Public Health Preparedness, 14*(5), 635–637.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. CoRR. arXiv:2004.05150

Bengoetxea, K., & Gonzalez-Dios, I. (2021). Multiaztertest: a multilingual analyzer on multiple levels of language for readability assessment. CoRR. arXiv:2109.04870

Bengoetxea, K., González-Dios, I., & Aguirregoitia, A. (2020). Aztertest: Open source linguistic and stylistic analysis tool. *Procesamiento del Lenguaje Natural, 64*, 61–68.

Bhavani, A., & Kumar, B. S. (2021). A review of state art of text classification algorithms. In: 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), IEEE, pp 1484–1490

Bikandi, U. R., & Valls, A. T. (2008). El tratamiento integrado de las lenguas. *Textos de Didáctica de la Lengua y la Literatura, 47*, 7–9. https://dialnet.unirioja.es/servlet/articulo?codigo=2512674

Castellv,í. C. (2004). La terminología en la traducción especializada. Gonzalo García, Consuelo; García Yebra, Valentín Manual de documentación y terminología para la traducción especializada Madrid: Arco Libros pp 89–125

Cañete, J., Chaperon, G., Fuentes, R., et al. (2020). Spanish pre-trained BERT model and evaluation data. In: PML4DC at ICLR 2020

Chiang-Soong, B., & Yager, R. E. (1993). Readability levels of the science textbooks most used in secondary schools. *School Science and Mathematics, 93*(1), 24–27.

Chowdhery, A., Narang, S., Devlin, J., et al. (2022). PaLM: Scaling language modeling with pathways. arXiv:2204.02311

Conneau, A., Khandelwal, K., Goyal, N., et al. (2020). Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

Association for Computational Linguistics, Online, pp 8440–8451. https://aclanthology.org/2020.acl-main.747

Crossley, S., Heintz, A., Choi, J. S., et al. (2022). A large-scaled corpus for assessing text readability. *Behavior Research Methods* pp 1–17

Daugherty, M. K., Kindall, H. D., Carter, V., et al. (2017). Integrating informational text and STEM: An innovative and necessary curricular approach. *Journal of STEM Teacher Education, 52*(1), 4.

Davison, A., & Kantor, R. N. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly* pp 187–209.

Deutsch, T., Jasbi, M., & Shieber, S. (2020). Linguistic features for readability assessment. In: Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics, Seattle, WA, USA → Online, pp 1–17. https://aclanthology.org/2020.bea-1.1

Devlin, J., Chang, M. W., Lee, K., et al. (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). Association for Computational Linguistics, Minneapolis, Minnesota, pp 4171–4186. https://aclanthology.org/N19-1423

Ehara, Y. (2022). Readability of scientific papers for English learners in various fields of science. In: NeurIPS 2022 AI for Science: Progress and Promises

Feng, L., Jansche, M., Huenerfauth, M., et al. (2010). A comparison of features for automatic readability assessment. In: 23rd International Conference on Computational Linguistics (COLING 2010), Poster Volume, pp 276–284. http://www.aclweb.org/anthology/C10-2032

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

Franco Aixelá, J. (2015). La traducción de textos científicos y técnicos. TonosDigital

François, T., & Miltsakaki, E. (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, pp 49–57

Gonzalez-Dios, I., Aranzabe, M. J., de Ilarraza, A. D., et al. (2014). Simple or complex? assessing the readability of Basque texts. In: Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers, pp 334–344

Guevara Benítez, Y., Cárdenas Espinoza, K., & Reyes Pérez, V. (2015). Levels of reading comprehension in high school students. a comparison by topic. *Actualidades en Psicología, 29*(118), 13–24.

Gutiérrez Rodilla, B. M. (1998). La ciencia empieza en la palabra: análisis e historia del lenguaje científico. Península

Gyasi, W. K. (2013). The role of readability in science education in Ghana: A readability index analysis of Ghana association of science teachers textbooks for senior high school. *IOSR Journal of Research & Method in Education, 2*(1), 9–19.

Hall, M., Frank, E., Holmes, G., et al. (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques". Morgan Kaufmann

Hu, J., Gao, X., & Qiu, X. (2021). Lexical coverage and readability of science textbooks for english-medium instruction secondary schools in Hong Kong. *SAGE Open, 11*(1), 21582440211001868.

Hung, W., Jonassen, D. H., & Rea, Liu. (2008). Problem-based learning. *Handbook of Research on Educational Communications and Technology, 3*(1), 485–506.

Imperial, J. M. (2021). BERT embeddings for automatic readability assessment. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). INCOMA Ltd., Held Online, pp 611–618, https://aclanthology.org/2021.ranlp-1.69

Klare, G. R. (1963). *Measurement of readability*. Iowa State University Press

Lee, B. W., Jang, Y. S., & Lee, J. H. J. (2021). Pushing on text readability assessment: A transformer meets handcrafted linguistic features. arXiv preprint arXiv:2109.12258

Lee, J., & Vajjala, S. (2022). A neural pairwise ranking model for readability assessment. arXiv preprint arXiv:2203.07450

Liu, Z., Lv, X., Liu, K., et al. (2010). Study on SVM compared with the other text classification methods. In: 2010 Second international workshop on education technology and computer science, IEEE, pp 219–222

Ma, Y., Fosler-Lussier, E., & Lofthus, R. (2012). Ranking-based readability assessment for early primary children's literature. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp 548–552

Martinc, M., Pollak, S., & Robnik-Šikonja, M. (2021). Supervised and unsupervised neural approaches to text readability. *Computational Linguistics, 47*(1), 141–179.

Muñoz Calvo, E. M., Muñoz Muñoz, L. M., García González, M. C., et al. (2013). La comprensión lectora de textos científicos en el proceso de enseñanza-aprendizaje. *Humanidades Médicas, 13*(3), 772–804.

Nadeem, F., & Ostendorf, M. (2018). Estimating linguistic complexity for science texts. In: Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications, pp 45–55.

Nwafor, C. E., Abonyi, O. S., Onyema, E. M., et al. (2022). Content coverage and readability of science textbooks in use in Nigerian secondary schools. *Journal of Education and Practice, 13*(7), 43–52.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pérez Zorrilla, M. J. (2005). *Evaluación de la comprensión lectora: dificultades y limitaciones*. Revista de educación

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

Plavén-Sigray, P., Matheson, G. J., Schiffler, B. C., et al. (2017). Research: The readability of scientific texts is decreasing over time. *eLife, 6*, e27725. https://doi.org/10.7554/eLife.27725

Roméu Escobar, A. (2002). La comunicación en la ciencia. una propuesta para la enseñanza interdisciplinaria del discurso científico. *Revista Educación, 17*(5)

Sagen, M. (2021). *Large-context question answering with cross-lingual transfer*. Master's thesis, Uppsala University, Department of Information Technology

Scalabrino, S., Linares-Vásquez, M., Oliveto, R., et al. (2018). A comprehensive model for code readability. *Journal of Software: Evolution and Process, 30*(6), e1958.

Schicchi, D., Pilato, G., & Bosco, G. L. (2020). Deep neural attention-based model for the evaluation of Italian sentences complexity. In: 2020 IEEE 14th International Conference on Semantic Computing (ICSC), IEEE, pp 253–256

Shishkova, T., & Popok, J. (1989). *Stylistics of the Spanish language*. Minsk: Vyschaya Shkola.

Si, L., & Callan, J. (2001). A statistical model for scientific readability. In: Proceedings of the tenth international conference on Information and knowledge management, pp 574–576

Sutton, C. (2003). Los profesores de ciencias como profesores de lenguaje. *Enseñanza de las ciencias: revista de investigación y experiencias didácticas*, pp 21–25

Tekfi, C. (1987). Readability formulas: An overview. *Journal of Documentation, 43*(3), 261–273.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), pp 2214–2218

Uribe, Á. C. (2007). La construcción de sentido en el discurso científico. y algunos apuntes sobre su presencia en la escuela. *Cuadernos de Lingüística Hispánica, 9*, 137–152.

Vajjala, S. (2021). Trends, limitations and open challenges in automatic readability assessment research. arXiv preprint arXiv:2105.00973

Vajjala, S., & Lučić, I. (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, pp 297–304

Vajjala, S., & Meurers, D. (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, pp 163–173

Vajjala, S., & Meurers, D. (2013). On the applicability of readability models to web texts. In: Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations, pp 59–68

Wei, J., Bosma, M., Zhao, V., et al. (2022). Finetuned language models are zero-shot learners. In: International Conference on Learning Representations, https://openreview.net/forum?id=gEZrGCozdqR

Xia, M., Kochmar, E., & Briscoe, T. (2016). Text readability assessment for second language learners. In: Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics, San Diego, CA, pp 12–22, https://aclanthology.org/W16-0502

Xu, W., Callison-Burch, C., & Napoles, C. (2015). Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics, 3*, 283–297. https://doi.org/10.1162/tacl_a_00139. https://arxiv.org/abs/direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00139/1566780/tacl_a_00139.pdf

Zhu, Y., Kiros, R., Zemel, R., et al. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: 2015 IEEE International Conference on Computer Vision (ICCV). IEEE Computer Society, Los Alamitos, CA, USA, pp 19–27, https://doi.ieeecomputersociety.org/10.1109/ICCV.2015.11