Procesamiento del Lenguaje Natural, Revista nº 71, septiembre de 2023, pp. 247-259 recibido 01-04-2023 revisado 24-04-2023 aceptado 03-05-2023

LegalEc: A New Corpus for Complex Word Identification Research in Law Studies in Ecuatorian Spanish

LegalEc: Un nuevo corpus para la investigación de la identificación de palabras complejas en los estudios de Derecho en español ecuatoriano

Jenny A. Ortiz-Zambrano,[1] **César Espin-Riofrio,**[1] **Arturo Montejo-Ráez**[2]
[1]Universidad de Guayaquil, Guayaquil, Ecuador
[2]Universidad de Jaén, 23071 Jaén, Spain
{jenny.ortizz, cesar.espinr}@ug.edu.ec, amontejo@ujaen.es

Abstract: In this paper, we present *LegalEc*, a new annotated corpus of complex lexis constructed from legal texts in Ecuadorian Spanish. We detail its compilation and annotation process. In order to provide a resource for the scientific community to continue research in the area of Lexical Simplification in the Spanish language, several complex word prediction experiments have been carried out on this corpus. We extracted 23 linguistic features which we combined with the encodings generated by models such as XLM-RoBERTa and RoBERTa-BNE (from the MarIA project). The evaluation shows that the combination of these features improves the prediction of lexical complexity.
Keywords: Lexical complexity, feature integration, corpus generation, legal language, Spanish.

Resumen: En este trabajo, presentamos a *LegalEc*, un nuevo corpus etiquetado con léxico complejo construido con textos de contenido legal en español ecuatoriano. Detallamos el proceso de compilación y anotación del mismo. Para proporcionar casos base a la comunidad científica, se han realizado varios experimentos de predicción de palabras complejas sobre este corpus. Extrajimos 23 caracterísiticas linguísticas que combinamos con las codificaciones generadas por modelos como XLM-RoBERTa y RoBERTa-BNE (del proyecto MarIA). La evaluación muestra que la combinación de estas características mejora notablemente la predicción de la complejidad léxica.
Palabras clave: Complejidad léxica, integración de características, generación de corpus, lenguaje jurídico, español.

1 Introduction

For many people, the way a text is written can cause a barrier to understanding its content (Saggion et al., 2015). The presence of infrequent or unknown words in the content of the texts drastically complicates their understanding for the reader (North, Zampieri, and Shardlow, 2023). The success or failure of reading comprehension will depend on the reader's prior knowledge about the meaning of the words (Anula, 2008).

Radical changes have been occurring over the past two decades in the way we access information. Information technologies provide people with abundant information in various fields such as education, news, social, health, or government, among others. However, this information constitutes a barrier in the comprehensibility of its content for many people, finding certain words difficult to read, interpret or learn. In readability and text simplification (TS) literature, these words are known as complex words (North, Zampieri, and Shardlow, 2022), being directly affected non-native speakers, people with low literacy rates, people with cognitive problems (Saggion et al., 2015), with a reading disability, such as dyslexia or aphasia (North, Zampieri, and Shardlow, 2022) and even some young university students, despite their high level of education and possessing specialized knowledge in various fields of study, could have a reading disability (Alarcón, Moreno, and

ISSN 1135-5948 DOI 10.26342/2023-71-19

Martínez, 2020).

Words identified as complex are on average longer, morphologically more unique, and less frequent in general corpora than non-complex words (Paetzold and Specia, 2016b), (Yimam et al., 2018). Shardlow et al. (2021) manifest that, the identification of complex words in texts (CWI) is the task of detecting in their contents, the words that are difficult or complex for people of a certain group and that it is an important first step in text simplification systems (Rico-Sulayes, 2020). CWI and the substitution of words identified as complex can significantly improve the readability and comprehension of a given text (Zotova et al., 2020).

There is a clear need to increase the scope of lexical simplification in terms of language coverage, given its social significance to make information accessible to broader audiences (Saggion et al., 2022). Being the Spanish language is one of the most spoken languages in the world (fourth in number of speakers), it usually has fewer terminological resources compared to other languages as English (Segura-Bedmar and Martínez, 2017) which has had the focus of a high number of investigations in the area of lexical simplification, as evidenced by the shared tasks of CWI in SemEval 2016 (Paetzold and Specia, 2016a), NAACL-HTL 2018 (Yimam et al., 2018), the ALexS task at IberLEF 2020 (Ortiz-Zambrano and Montejo-Ráez, 2020), the 15th edition of SemEval and the first Lexical Complexity Prediction task (Shardlow et al., 2021) and TSAR-2022, the workshop on Text Simplification, Accessibility, and Readability (Saggion et al., 2023), among other initiatives in this sense.

In addition to this, there is little research on natural language processing (NLP) tools to support students and teachers of Spanish, as well as the development of effective NLP applications aimed at teaching, given that the resources for Spanish that are still available do not contain annotations that facilitate the contribution of possible solutions (Davidson et al., 2020).

Currently, in Ecuador, the use of online procedures has led people with limited legal understanding to face unfamiliar terms. There are notable differences in the level of legal knowledge of the general population. Our research is another step in finding mechanisms to help university students in the understanding of legal jargon within the different domains where text simplification becomes useful. The legal domain is relevant to a wider range of people, as citizens have to tackle with legal processes and administrative in everyday life formalities without the help of an expert (Döring, 2021). The contributions of this research can be summarized as follows:

- A new corpus of 6,594 Spanish texts in the legal domain has been generated, manually annotated, named *LegalEc*, whose objective is to contribute to research on the identification and prediction of the complex words in Ecuadorian Spanish.
- The analysis of the content of the texts was carried out by applying several complexity metrics for Spanish where the evaluation of the results determined that the texts contain words that become difficult to understand due to the level of complexity they present.
- We present some experiments on the corpus. The experiments showed that the performance of the transformer-based models can be improved by integrating linguistic information automatically derived from texts.

The rest of the article is organized as follows:

Section 2 describes the work related to lexical simplification focused on systems based on lexical complexity metrics for Spanish and on linguistic models. Section 3, exposes the construction process of the corpus *LegalEc*. Section 4, presents the experimental results and an analysis on them. Section 5 exposes the discussion of the results obtained. Section 6 summarizes main contributions and provides some insights on future work.

2 Related Work

Pitkowski and Gamarra (2009) define a *corpus* as a large-volume compilation made up of different types of texts, written or oral, made up of several million words in electronic format. An annotated *corpus* becomes an essential resource for any PLN task (Quevedo-Marcos, 2020). While annotated English learner corpora are still widely available, large Spanish corpora are less common.

Developing an annotated corpus is a time consuming task. In addition, even when human annotation is performed, there may be discrepancies between annotators or within the same annotator, which could affect the quality of the corpus. Consequently, the lack of supervision over the annotation process can lead to a low-quality corpus (García-Díaz et al., 2020).

2.1 Lexical complexity prediction

Segura-Bedmar and Martínez (2017) used the corpus *EasyDPL* (Easy Drug Package Leaflets) in their research, a collection of 306 leaflets written in Spanish and manually annotated with 1,400 adverse drug effects and their simpler synonyms.

Ortiz-Zambrano and Montejo-Ráez (2017) created a new corpus of videos with their transcriptions named VYTEDU (Videos and Transcriptions for research in the Education domain) developed at the State University of Guayaquil for the study of text simplification systems in the educational field. For this, 55 videos were produced during the classes of the teachers within the academic classrooms in the different careers of the University of Guayaquil. The videos contain in their recording different themes that correspond to several of the subjects of the academic programs. The system measures some of the indicators selected by (Saggion et al., 2015) and constitutes a set of metrics that allow analyzing the complexity of the text at various levels: the lexical complexity index and the sentence complexity index, proposed by (Anula, 2008), and the legibility of Spaulding's Spanish (Spaulding, 1956).

The annotated corpus called *VYTEDU-CW* was proposed by Ortiz-Zambrano et al. (2019). This corpus is the result of the process of identification and labeling of the complex words contained in the texts of the VYTEDU corpus carried out by students of the different careers of the University of Guayaquil This resource was provided to the participants of the ALexS 2020 workshop (Lexical Analysis Task in SEPLN 2020) as part of the second edition of IberLEF 2020[1] (Forum for the Evaluation of Iberian Languages) (Ortiz-Zambrano and Montejo-Ráez, 2020).

The objective of this task was to contribute to the advancement of methods and techniques for the effective identification of complex words, since the substitution of complex words in texts improves the understanding of a given text by the reader (thus displaying a better level of for readability). Most of the works combined strategies to generate different functions with machine learning algorithms, including classical and deep neural networks (Ortiz-Zambrano and Montejo-Ráez, 2020).

Ortiz-Zambrano and Montejo-Ráez (2021) introduced *CLexIS*[2], a new Spanish annotated corpus of complex words in computational studies. A total of seven textual complexity metrics were used to assess the complexity of the texts. Furthermore, as a baseline, two experiments were performed to predict word complexity: one using a supervised learning approach and the other using an unsupervised approach whose solution was based on word frequency in a general corpus.

2.2 Measures of Lexical Complexity for Spanish

A good indicator of the quality of writing is to use a measure of lexical complexity that refers to the size, variety, and quality of a student's vocabulary (Crossley, Salsbury, and McNamara, 2012). Another way to determine the lexical complexity of words for Spanish is based on the metrics proposed by Anula (2008) and Spaulding (1956). These measures have been applied in recent years in research on the simplification of texts for Spanish, such as the work carried out by Saggion et al. (2015), Ortiz-Zambrano and Montejo-Ráez (2017), Ortiz-Zambrano and Varela Tapia (2019), Camposa et al. (2020), Ortiz-Zambrano and Montejo-Ráez (2021) to cite a few examples. The formulas were proposed by Anula (2008) except the SSR formula corresponds to Spaulding (1956). For better understanding, the Table 2 shows the definition of the variables.

LC: The Lexical Complexity Index.
LDI: Lexical Distribution Index.
ILFW: Index of Low Frequency Words.
SSR: Spaulding's Spanish Readability Index.
SCI: The Sentence Complex Index.
ASL: The Average Sentences Length.
CS: The Percentage of Complex Sentence.

$$LC = (LDI + ILFW)/2 \qquad (1)$$

$$LDI = N_{dcw}/N_s \qquad (2)$$

[1] https://ceur-ws.org/Vol-2664/

Variable	Total number of...
N_w	words
N_{cw}	content words
N_{dcw}	distinct content words
N_{rw}	rare words
N_{lfw}	frequent words
N_s	sentences
N_{cs}	complex sentences
	... per document

Table 1: Definition of the columns in Table 3.

$$ILFW = N_{lfw}/Ncw * 100 \quad (3)$$

$$SSR = 1.609N_w/N_s + 331.8N_{rw}/N_w + 22.0 \quad (4)$$

$$SCI = (ASL + CS)/2 \quad (5)$$

$$ASL = N_w/N_{sc} \quad (6)$$

$$CS = N_{cs}/N_s \quad (7)$$

In the Table 1, the definition of the variables is detailed.

For the interpretation of the results of the SSR formula, Spaulding (1956) proposed a table of values as presented in Table 2 (Camposa et al., 2020).

Interpretation of the SSR measure

Score SSR	Readability
less than 40	material very simplified
40-60	Very easy
61-80	Easy
81-100	Moderate difficulty
101-120	Difficult
121 o más	Very difficult

Table 2: Interpretation of the SSR metric.

3 *The LegalEc corpus*

The corpus entitled *LegalEc* is a corpus in Law studies in Spanish Ecuadorian aims to contribute to the research in the area of Lexical Complexity Prediction, specifically in the identification of complex words in the legal domain. *LegalEc* offers a collection of 900 texts from two main sources: the final degree projects of the students of the Law course of the University of Guayaquil, and various articles of the Constitution of the Republic of Ecuador.

The documents referring to the degree projects of the students were selected from the DSpace repository of the University of Guayaquil[2]. We took as a reference the works carried out in different topics that address the legal field. Regarding the texts of the articles of the Constitution of Ecuador, those that are directed to the duties and rights of citizens were chosen preferably. The content of the texts is written in the Spanish language spoken by Ecuadorians. It should be noted that the University of Guayaquil is a higher education center, and is the largest and oldest public institution in the country with an average of 65,649 students.

For the construction of the data set we followed the format used in the SemEval-2021 task 1 competition[3] whose objective was to predict lexical complexity (Shardlow, Cooper, and Zampieri, 2020). Each sample in the *LegalEc* dataset contains the following fields:

- **Id:** The identification number of each record.
- **Source:** The description of the source where the text comes from.
- **Sentence:** The set of words for which complexity was needed to be measured.
- **Token:** The word identified as complex for the annotator to understand. The only word needed to measure complexity.
- **Complexity:** It is the level of complexity of the word whose value is within the range [0, 1].
- **Features**: To strengthen the data set, a set of 23 linguistic features was included and computed for each sentence. We indicate these features and some of the research papers that have also considered them:

  1. The absolute frequency (Paetzold, 2021).
  2. The relative frequency of the target word.
  3. The number of characters of the token (Paetzold, 2021).
  4. The number of syllables (Shardlow, 2013), (Ronzano et al., 2016), (Shardlow, Cooper, and Zampieri,

[2] http://repositorio.ug.edu.ec/

[3] https://sites.google.com/view/lcpsharedtask2021

2020), (Paetzold and Specia, 2016b).

5. The position of the target word in the sentence (Shardlow, 2013), (Ronzano et al., 2016).
6. Number of words in sentence (Shardlow, 2013), (Ronzano et al., 2016).
7. The Part Of Speech category (Ronzano et al., 2016).
8. The relative frequency of the word before the token (Paetzold, 2021).
9. The relative frequency of the word after the token (Paetzold, 2021).
10. The number of characters in the word before the token (Ronzano et al., 2016).
11. The number of characters in the word after the token (Ronzano et al., 2016).
12. Lexical diversity (Shiroyama, 2022).
13. The number of synonyms (Mosquera, 2021).
14. The number of hyponyms (Mosquera, 2021).
15. The number of hyperonyms (Mosquera, 2021).
16. The number of nouns, singular or massive.
17. The number of auxiliaries verbs.
18. The number of adverbs.
19. The number of symbols.
20. The number of numeric expressions.
21. The number of verbs.
22. The number of nouns.
23. The number of pronouns.

The last eight features (from 16 to 23) are traditional categories of the POS (Part Of Speech) applied in the investigations of (Ronzano et al., 2016), (Paetzold and Specia, 2016b), (Desai et al., 2021).

Some statistics on the corpus texts are presented in Table 3. The Table 4 shows the definition of the variables. As can be seen, the number of rare words (N_{rw}) is much higher than that of less frequent words (N_{lfw}), although an average of 10% of words in sentences is considered with low frequency.

3.1 Annotation Process

For the annotation process of complex words, the participants (student volunteers) were grouped according to the semester of study they were in:

- Group A (*Basic Level*): students from first to third semester of studies.
- Group B (*Middle Level*): students from fourth to sixth semester of studies.
- Group C (*Advanced Level*): students from seventh to ninth semester of studies.

A total of 27 students were selected as annotators, all of them over 18 years of age, and deciding to participate voluntarily. The students were distributed in different groups by level of study. The nine students in each group were divided into separate subgroups of three annotators. The texts were also classified into basic, intermediate and advanced levels depending on their content. Finally, each student was assigned a total of 300 texts to carry out this process. Figure 1 presents the methodology with which the texts were assigned to the annotators. Annotators only have to mark those words considered difficult. A custom tool was developed to this end.

A total of 6,594 words were tagged by annotators. There are words that were selected as complex by all of the nine students from different study levels, as it is the case of the words: *suprayacente, imprescriptible, ineludible, antropogénicos. Toxicomanía, decorosa, conculcar, primigenias, impugnación, circunscripción* were words chosen by eight of the nine annotators, to mention a few examples. Table 5 presents several examples of the words identified and annotated as complex in the corpus during the *LegalEc* tagging process. Next, Table 6 shows the number of words annotated by each tagger.

3.2 Inter-annotator Agreement

To analyze the concordance of the data of the groups of each level, the *Kappa Fleiss* index was measured. The results of the matches between students showed a value of p = 0.13, which indicates that it is in the range of 0.00 - 0.20, therefore, there is a *low* level of agreement (Cabrera-Meléndez et al., 2022) according to the Table 7. Finally, Table 7 shows the total number of matches obtained between taggers.

The Statistics of *LegalEc*

	N_{chrs}	N_w	N_{dcw}	N_{cw}	N_{lfw}	N_{rw}	N_s	N_{cs}
Valid	900	900	900	900	900	900	900	900
Missing	0	0	0	0	0	0	0	0
Mean	505.40	92.13	61.17	46.96	10.15	44.02	3.69	1.02
Std.Dev	151.20	28.48	15.09	14.44	4.57	14.24	2.35	0.87
Min	232.00	37.00	29.00	20.00	1.00	15.00	1.00	0.00
Max	1,239.00	223.00	128.00	118.00	33.00	111.00	24.00	5.00
Sum	4.55e+5	8.29e+4	5.50e+4	4.23e+4	9,138.00	3.96e+4	3,320.00	919.00

Table 3: Descriptive Statistics of different counters over documents in *LegalEc*.

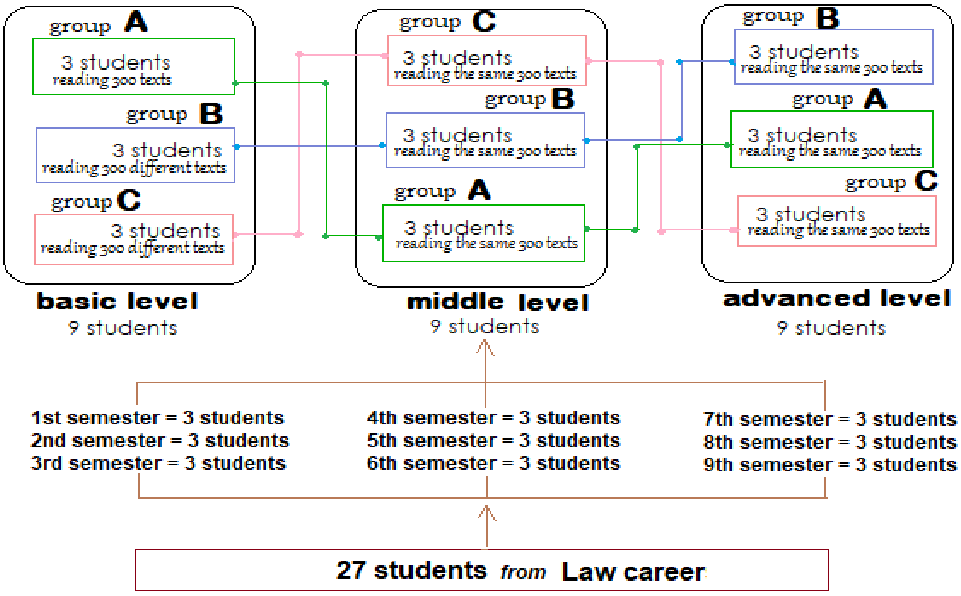
group A
3 students
reading 300 texts
group B
3 students
reading 300 different texts
group C
3 students
reading 300 different texts
basic level
9 students
group C
3 students
reading the same 300 texts
group B
3 students
reading the same 300 texts
group A
3 students
reading the same 300 texts
middle level
9 students
group B
3 students
reading the same 300 texts
group A
3 students
reading the same 300 texts
group C
3 students
reading the same 300 texts
advanced level
9 students
1st semester = 3 students
2nd semester = 3 students
3rd semester = 3 students
4th semester = 3 students
5th semester = 3 students
6th semester = 3 students
7th semester = 3 students
8th semester = 3 students
9th semester = 3 students
27 students from Law career

Figure 1: Applied methodology for the assignment of *LegalEc* texts.

Variable	Total number of...
N_{chrs}	characters
N_w	words
N_{cw}	content words
N_{dcw}	distinct content words
N_{rw}	rare words
N_{lfw}	less frequent words
N_s	sentences
N_{cs}	complex sentences
	... per text

Table 4: Definition of the columns.

3.2.1 Analysis of textual complexity

The lexical complexity metrics for Spanish described in section 2.2 were applied for *LegalEc* dataset, see Table 8. For the analysis of the data and interpretation of the results according to the applied formulas, we consider the works of Saggion et al. (2015) and Camposa et al. (2020). The **SSR** formula made it possible to measure the complexity of the texts using the average number of words per line and the percentage of complex words according to a list created by the author; the results are in the range 130.20 and 413.20, which shows that the readability of the texts is *very difficult*, according to Spaulding's table of interpretations. See Table 2.

For the calculation of the **LC**, the formula uses a list of words taken from the CREA[4] lexicon whose frequency is less than 1,000. It has a similar behavior to the SSR, that is, the greater the complexity, the greater the value of return of this formula (Camposa et al., 2020). The results of the LC are in the range 5.86 and 53.09, which shows that the content of the texts have a high level of lexical complexity, since they are based on an infrequent lexicon. The **SCI** metric, allowed to measure the average complexity of the sen-

[4] http://corpus.rae.es/lfrecuencias.html

Words tagged by the annotators in the texts of the corpus *LegalEc*		
No. text	**Sentence**	**Complexity**
00042	Se prohíbe la emisión de publicidad que induzca a la [..] el **sexismo**, la intolerancia religiosa o política, [..]	0.33
00077	El Estado responderá **civilmente** por los daños y [..]	0.33
00078	Conservar la propiedad imprescriptible de sus tierras comunitarias, que serán **inalienables**, [..]	0,44
00152	Aportes, **subvenciones** y subsidios que fueren acordados en su favor por instituciones públicas y privadas, [..]	0,89
00230	[..] y aún después de aprobados les quedará **expedito** su recurso a la justicia, contra toda lesión o perjuicio [..]	0.33
00234	Artículo 75 Constitución de la República del Ecuador [..] y a la tutela efectiva, imparcial y **expedita** de sus [..]	0.56
00241	Si por un acto de partición se adjudican a varias personas inmuebles o parte [..] que antes se poseían **proindiviso**, [..]	0.89
00279	El error de hecho vicia el consentimiento cuando [..] como si una de las partes entendiese **empréstito**, [..]	0.89

Table 5: Examples of words tagged by the annotators in the texts of the *LegalEc* corpus.

Annotations made in *LegalEc* by each tagger									
	BASIC level			**MIDDLE level**			**ADVANCED level**		
Annotator	S1	S2	S3	S4	S5	S6	S7	S8	S9
Group **A**	1,113	380	495	558	782	804	1,001	543	635
Group **B**	413	707	509	517	451	377	643	641	439
Group **C**	771	624	497	365	681	833	317	781	387

Table 6: Total number of words annotated by each annotator in corpus *LegalEc.*

tence, with a result in the range 1.79 and 75.00, demonstrating that the texts have a high complexity index at sentence level.

4 Experimental setup

In their recent overview on lexical complexity prediction research, North, Zampieri, and Shardlow (2023) they found that transformer based models, when combined in ensembles, are the state of the art for machine learning approaches. Also, multi-word expressions detection for LCP (Lexical Complexity Prediction) is one of the most challenging task.

We have carried out a series of experiments to serve as base line for other investigations. Our approach is based on the combination of the 23 linguistic features included in the corpus in combination with the encodings generated by several transformer based models for Spanish such as XLM-RoBERTa-Base, XLM-RoBERTa-Base, and XLM-RoBERTa-large pre-trained models, that have been widely used to create state-of-the-art solutions for numerous tasks (Paetzold, 2021).

We have experimented without fine-tuning the encoders (using only pre-trained models). We carried out runs to test whether the combination of linguistic features (LF) supposes an improvement compared to full end-to-end approaches. The way linguistic features are integrated is by concatenating them, after a min-max scaling, with the embeddings resulting from the last encoding layer, and before reaching the classification head (see Figure 2).

The steps are the following:

1. The input sequence was extended with the target term for which the complexity estimate is determined. This term is placed before a [SEP] token. The [SEP] token helps to clearly separate and distinguish the different parts of a text input in the model, making it easier to process it properly.
2. Once the input sequence passes through the encoder (XLM-RoBERTa-Base, XLM-RoBERTa-Base, or XLM-RoBERTa-large), the sentence embedding is concatenated with a min-max

Annotators agreement

	BASIC level			MIDDLE level			ADVANCED level		
No. annotators	S1	S2	S3	S4	S5	S6	S7	S8	S9
Group **A**	1,085	324	85	716	378	224	1,143	347	114
Group **B**	645	309	122	707	220	66	752	289	131
Group **C**	791	300	167	855	320	128	934	235	27
Total	2,521	933	374	2,278	918	418	2,829	871	272
AVR	66%	24%	10%	63%	25%	12%	71%	22%	7%

Table 7: Number words tagged as complex agreed by different number of annotators.

Lexical Complexity Metrics for Spanish in *LegalEc*

	LC	**LDI**	**ILFW**	**SSR**	**SCI**	**ASL**	**CS**	**MTLD**
Valid	900	900	900	900	900	900	900	900
Missing	0	0	0	0	0	0	0	0
Mean	22.46	22.73	22.18	235.30	16.95	33.54	0.36	13.41
Std.Dev	8.17	14.78	8.75	36.58	10.93	21.69	0.34	0.57
Min	5.86	2.38	2.33	130.20	1.79	3.58	0.00	11.92
Max	53.09	91.00	54.84	413.20	75.00	149.00	1.00	15.61

Table 8: Results of the application of lexical complexity metrics for spanish in corpus *LegalEc*.

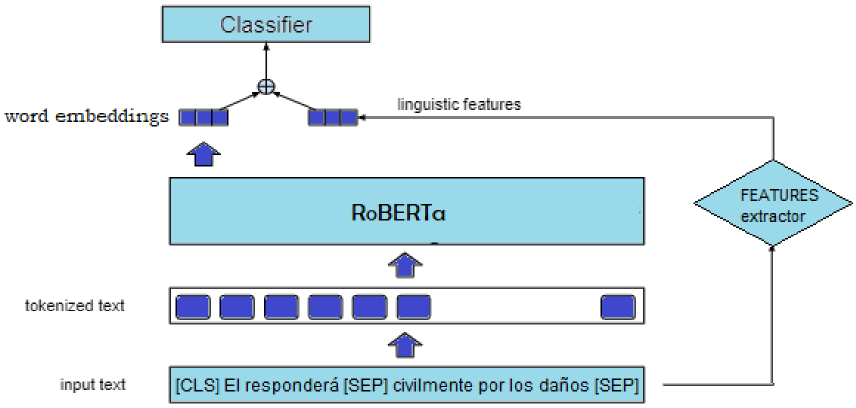

Figure 2: Process flow methodology integrating linguistic features.

scaled vector of the linguistic features.

3. The resulting vector enter the classification layer. The classification head is composed of a pair of dense layers preceded by a dropout layer and an activation *tanh* layer after the first dense layer. Therefore, the entire network can be tuned even if linguistic features are injected. The models were trained with a batch size of 32 for 10, 30 and 50 epochs.

5 *Results*

The metric used to evaluate the different configurations are Mean Absolute Error (MAE), Mean Square Error (MSE), root mean square error (RMSE) and Pearson's correlation coefficient. The following sections show the results obtained for the different executions explored. To carry out the experiments we apply the models RoBERTa-large-bne, XLM-RoBERTa-base and XLM-RoBERTa-large as in the work carried out by (Taya et al., 2021). The first trainings were done over 10 epochs and a batch-size of 32. In order to attempt a better result in the predictions, the linguistic features (LF) were included in the next execution. The results of these experiments can be seen in Table 9. The best result was achieved by the RoBERTa-large-BNE model with a MAE value of 0.1560.

Then, the classification heads were trained over an increased number of epochs (30). Now, it can be noticed that the complexity prediction has a significant improvement when including linguistic features, reaching the model XLM-RoBERTa-Base + LF a MAE of 0.1341; RoBERTa-large-BNE + LF a MAE of 0.1363, and XLM-RoBERTa-large + LF a MAE of 0.1360. See Table 10 for detailed results.

In view of the fact that the improvement of the results was evident when increasing the epochs, we carried out one more training over 50 epochs. The prediction error reached by RoBERTa-large-BNE + LF is a MAE of 0.1349, and XLM- RoBERTa-large + LF with a MAE of 0.1338, achieving an improvement over previous runs (see Table 11). A finding is that the inclusion of the linguistic features and the increase in the number of epochs in the executions of the models achieves a relevant gain in the performance of the algorithms for lexical complexity prediction as it is graphically shown in Figure 3.

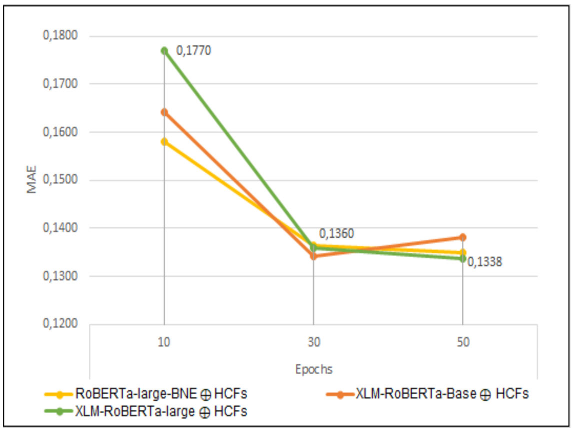

Figure 3: Integration of HCF and increase in times.

6 Discussion

The Table 12 summarizes the results on the performance of the combinations of features with the different language models based on Transformers targeting the Spanish language with the *LegalEc* dataset.

The performance gain appears to be related to the number of epochs plus the inclusion of additional features. The use of these features in conjunction with pre-trained model encodings turn out to be performance friendly. The best result was obtained by applying the XLM-RoBERTa-large + LF model with a MAE of 0.1338, after training for 50 epochs. The other models also reported higher performance during the different executions. In this research we have shown how the inclusion of the linguistic features and the increase of epochs substantially improve the prediction of lexical complexity, even for smaller models like XLM-RoBERTa-base. See Figure 3.

7 Conclusions and Further Work

This work presents a new corpus for lexical complexity prediction research on Spanish in the legal domain. The corpus is freely available and, although there is a low level of agreement among annotators, *learning from disagreement* researcher can use this dataset for analyzing this phenomena. Anyhow, the 900 texts compiled exhibits a large number of complex terms agreed by reviewers, so it is also suitable for testing solutions on LCP.

Anyhow, in the near future, we plan to make use of NLP tools that help the process of compiling and annotating linguistic corpora such as the one proposed by García-Díaz et al. (2020) to avoid the errors that are made during the corpus annotation stage as a result of the differences between annotators since they can affect the quality of the corpus.

A comprehensive set of experiments was performed to test the desirability of combining transformer encodings with lexical features traditionally used in complex word identification. The experiments directed to the Spanish language were applied to study how complex are these texts and if it is feasible to automatically obtain the level of lexical complexity. State-of-the-art language models were tested and also combined with linguistic features as hybrid solution, to serve as baseline results for future experiments.

It is clear that more research needs to be done on the combination of features in deep learning models. In general, the XLM-RoBERTa-large-bne, XLM-RoBERTa-base, XLM-RoBERTa models reported better results when combined with linguistic features and the number of epochs was increased. However, deep learning models work like a black box, so understanding how linguistic features complement deep features requires work on the explainability of the deep model itself, as transformers can encode in-

Spanish Language Model* pre-trained with *LegalEc

with 10 epochs

Model	MAE	MSE	RMSE	R2	Poisson	Pearson
XLM-RoBERTa-Base	0.2286	0.0710	0.2665	-0.0010	0.3970	0.0323
XLM-RoBERTa-Base ⊕ LF	0.1642	0.0345	0.2669	0.5458	0.2269	0.9948
RoBERTa-large-BNE	0.1560	0.0461	0.2149	-0.1394	0.3129	0.2109
RoBERTa-large-BNE ⊕ LF	0.1580	0.0630	0.2510	-0.0528	0.4126	0.5353
XLM-RoBERTa-large	0.2100	0.0789	0.2828	-0.1038	0.4652	0.0040
XLM-RoBERTa-large ⊕ LF	0.1770	0.0682	0.2612	-0.0821	0.4388	-0.0400

Table 9: Results of the pre-trained models applying 10 epochs.

Spanish Language Model* pre-trained with *LegalEc

with 30 epochs

Model	MAE	MSE	RMSE	R2	Poisson	Pearson
XLM-RoBERTa-Base	0.2255	0.0729	0.2701	-0.0149	0.4076	0.1230
XLM-RoBERTa-Base ⊕ LF	0.1341	0.0242	0.1558	0.6581	0.1652	0.9948
RoBERTa-large-BNE	0.2284	0.0789	0.2809	-0.0428	0.4288	0.0910
RoBERTa-large-BNE ⊕ LF	0.1363	0.0268	0.1638	0.6548	0.1628	0.9936
XLM-RoBERTa-large	0.2360	0.0744	0.2728	-0.0058	0.4068	0.0573
XLM-RoBERTa-large ⊕ LF	0.1360	0.0256	0.1600	0.6614	0.1632	0.9959

Table 10: Results of the pre-trained models applying 30 epochs.

Spanish Language Model* pre-trained with *LegalEc

with 50 epochs

Model	MAE	MSE	RMSE	R2	Poisson	Pearson
XLM-RoBERTa-Base	0,2239	0.0724	0.2691	-0.0148	0.4058	-0.0628
XLM-RoBERTa-Base ⊕ LF	0.1382	0.0259	0.1612	0.6564	0.1682	0.9958
RoBERTa-large-BNE	0,2085	0.0765	0.2766	-0.0781	0.4462	-0.0332
RoBERTa-large-BNE ⊕ LF	0.1349	0.0270	0.1646	0.6418	0.1575	0.9948
XLM-RoBERTa-large	0.2375	0.0761	0.2760	-0.0070	0.4107	0.0417
XLM-RoBERTa-large ⊕ LF	0.1338	0.0252	0.1587	0.6611	0.1594	0.9950

Table 11: Results of the pre-trained models applying 50 epochs.

Best Results of the execution of the pre-trained models

Models	10 epochs	30 epochs	50 epochs
XLM-RoBERTa-Base	0.2286	0.2255	0.2239
XLM-RoBERTa-Base ⊕ LF	0.1642	0.1341	0.1382
RoBERTa-large-BNE	0.1560	0.2284	0.2085
RoBERTa-large-BNE ⊕ LF	0.1580	0.1363	0.1349
XLM-RoBERTa-large	0.2100	0.2360	0.2375
XLM-RoBERTa-large ⊕ LF	0.1770	0.1360	0.1338

Table 12: Best Results of the execution of the pre-trained models through the MAE metric.

formation related to syntax, dependencies, grammar, gender, negation, semantics, etc. inside the layers.

We hypothesize that the wealth of knowledge present in transformer-based models can help in extracting complementary clues of contextual complexity (Paetzold, 2021). We plan to explore which language features are adding additional information to the network, so ablation tests on the use of these language features are envisioned by gradually introducing them into the model. Additional feature selection and feature transformation strategies could be evaluated. We believe that tuning network parameters with these external features in mind could eventually lead to better performance.

We consider that this corpus intends a valuable contribution to the scientific community to continue advancing in the studies of NLP techniques for lexical simplification in the identification of complex words that affect the comprehensibility of the content of the texts in the legal domain. To obtain this resource you can contact the authors.

References

Alarcón, R., L. Moreno, and P. Martínez. 2020. Hulat-alexs cwi task-cwi for language and learning disabilities applied to university educational texts. In *IberLEF@ SEPLN*, pages 24–30.

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

Cabrera-Meléndez, J. L., D. Iparraguirre-León, M. Way, F. Valenzuela-Oré, and D. B. Montesinos-Tubée. 2022. The applicability of similarity indices in an ethnobotanical study of medicinal plants from three localities of the yunga district, moquegua region, peru. *Ethnobotany Research and Applications*, 24(16).

Camposa, R. A., P. Estrella, J. A. Castillo, and W. A. Grijalba. 2020. Estudio de la complejidad del español para la simplificación textual. *Revista Tecnología en Marcha*, pages ág–45.

Crossley, S. A., T. Salsbury, and D. S. McNamara. 2012. Predicting the proficiency level of language learners using lexical indices. *Language Testing*, 29(2):243–263.

Davidson, S., A. Yamada, P. F. Mira, A. Carando, C. H. S. Gutierrez, and K. Sagae. 2020. Developing nlp tools with a new corpus of learner spanish. In *Proceedings of the 12th language resources and evaluation conference*, pages 7238–7243.

Desai, A. T., K. North, M. Zampieri, and C. Homan. 2021. LCP-RIT at SemEval-2021 task 1: Exploring linguistic features for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 548–553, Online, August. Association for Computational Linguistics.

Döring, M. 2021. How-to bureaucracy: A concept of citizens' administrative literacy. *Administration & Society*, 53(8):1155–1177.

García-Díaz, J. A., Á. Almela, G. Alcaraz-Mármol, and R. Valencia-García. 2020. Umucorpusclassifier: Compilation and evaluation of linguistic corpus for natural language processing tasks. *Procesamiento del Lenguaje Natural*, 65:139–142.

Mosquera, A. 2021. Alejandro mosquera at semeval-2021 task 1: Exploring sentence and word features for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 554–559.

North, K., M. Zampieri, and M. Shardlow. 2022. An evaluation of binary comparative lexical complexity models. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 197–203, Seattle, Washington, July. Association for Computational Linguistics.

North, K., M. Zampieri, and M. Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Computing Surveys*, 55(9):1–42.

Ortiz-Zambrano, J. and A. Montejo-Ráez. 2017. Vytedu: Un corpus de vídeos y sus transcripciones para investigación en el ámbito educativo.

Ortiz-Zambrano, J. and A. Montejo-Ráez. 2020. Overview of alexs 2020: First workshop on lexical analysis at sepln. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*, volume 2664, pages 1–6.

Ortiz-Zambrano, J. and A. Montejo-Ráez. 2021. Clexis2: A new corpus for complex word identification research in computing studies. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1075–1083.

Ortiz-Zambrano, J., A. MontejoRáez, K. Lino Castillo, O. Gonzalez Mendoza, and B. Cañizales Perdomo. 2019. Vytedu-cw: Difficult words as a barrier in the reading comprehension of university students. In *Advances in Emerging Trends and Technologies: Volume 1*. Springer, pages 167–176.

Ortiz-Zambrano, J. and E. Varela Tapia. 2019. Reading comprehension in university texts: the metrics of lexical complexity in corpus analysis in spanish. In

Computer and Communication Engineering: First International Conference, ICCE 2018, Guayaquil, Ecuador, October 25–27, 2018, Proceedings 1, pages 111–123. Springer.

Paetzold, G. 2021. Utfpr at semeval-2021 task 1: Complexity prediction by combining bert vectors and classic features. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 617–622.

Paetzold, G. and L. Specia. 2016a. Semeval 2016 task 11: Complex word identification. pages 560–569, 01.

Paetzold, G. and L. Specia. 2016b. Sv000gg at semeval-2016 task 11: Heavy gauge complex word identification with system voting. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 969–974.

Pitkowski, E. F. and J. V. Gamarra. 2009. El uso de los corpus lingüísticos como herramienta pedagógica para la enseñanza y aprendizaje de ele. *Tinkuy: boletín de investigación y debate*, (11):31–51.

Quevedo-Marcos, B. 2020. Análisis de las herrramientas de procesamiento de lenguaje natural para estructurar textos médicos.

Rico-Sulayes, A. 2020. General lexicon-based complex word identification extended with stem n-grams and morphological engines. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR-WS, Malaga, Spain*, volume 23.

Ronzano, F., L. E. Anke, H. Saggion, et al. 2016. Taln at semeval-2016 task 11: Modelling complex words by contextual, lexical and semantic features. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1011–1016.

Saggion, H., S. Štajner, S. Bott, S. Mille, L. Rello, and B. Drndarevic. 2015. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):1–36.

Saggion, H., S. Štajner, D. Ferrés, K. C. Sheang, M. Shardlow, K. North, and M. Zampieri. 2022. Findings of the tsar-2022 shared task on multilingual lexical simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 271–283, Abu Dhabi, United Arab Emirates (Virtual), December. Association for Computational Linguistics.

Saggion, H., S. Štajner, D. Ferrés, K. C. Sheang, M. Shardlow, K. North, and M. Zampieri. 2023. Findings of the tsar-2022 shared task on multilingual lexical simplification. *arXiv preprint arXiv:2302.02888*.

Segura-Bedmar, I. and P. Martínez. 2017. Simplifying drug package leaflets written in spanish by using word embedding. *Journal of biomedical semantics*, 8(1):1–9.

Shardlow, M. 2013. A comparison of techniques to automatically identify complex words. In *51st annual meeting of the association for computational linguistics proceedings of the student research workshop*, pages 103–109.

Shardlow, M., M. Cooper, and M. Zampieri. 2020. CompLex — a new corpus for lexical complexity prediction from Likert Scale data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 57–62, Marseille, France, May. European Language Resources Association.

Shardlow, M., R. Evans, G. H. Paetzold, and M. Zampieri. 2021. SemEval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online, August. Association for Computational Linguistics.

Shiroyama, T. 2022. Comparing lexical complexity using two different ve modes: a pilot study. *Intelligent CALL, granular systems and learner data: short papers from EUROCALL 2022*, page 358.

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

Taya, Y., L. Kanashiro Pereira, F. Cheng, and I. Kobayashi. 2021. OCHADAI-KYOTO at SemEval-2021 task 1: Enhancing model generalization and robustness for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 17–23, Online, August. Association for Computational Linguistics.

Yimam, S. M., C. Biemann, S. Malmasi, G. Paetzold, L. Specia, S. Štajner, A. Tack, and M. Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana, June. Association for Computational Linguistics.

Zotova, E., M. Cuadros, N. Perez, and A. G. Pablos. 2020. Vicomtech at alexs 2020: Unsupervised complex word identification based on domain frequency. In *IberLEF@ SEPLN*, pages 7–14.