GovAI*Ec*: A Lexical Complexity Corpus for Spanish in Ecuadorian public documents

Anonymous ACL submission

Abstract

In this article, we present GovAI*Ec*, a new annotated corpus of complex lexicon created with institutional texts in Ecuadorian Spanish, and we detail the process of compiling and annotating this corpus. With the aim of providing a valuable resource to the scientific community to advance research in the field of Lexical Simplification in the Spanish language, we carried out several complex word prediction experiments using this corpus. The complex word labeling process was carried out with a group of annotators with different levels of literacy, in order to ensure a comprehensive evaluation. We use Lexical Complexity metrics as units of analysis, and apply advanced multilingual language models such as XLM-RoBERTa-Base, RoBERTa-large-BNE, XLM-RoBERTa-Large and BERT to evaluate the corpus. This corpus is invaluable for identifying words that represent barriers in the reading comprehension of users who interact with bureaucratic procedures of various entities in Ecuador.

1 Introduction

In recent times, the use of artificial intelligence (AI) has seen a notable increase to address the governance challenges facing cities. Given its advanced capabilities, AI is expected to become a critical resource for local governments in their pursuit of smart and sustainable development (Son et al., 2023). Although the potential of Artificial Intelligence has been widely explored in the private sector, its usefulness in the public sphere is increasingly being recognized by governments themselves, who are adopting AI to strengthen their performance in various areas (Vélez et al., 2022), which includes an important challenge: improving communication between the government and citizens, an aspect that has represented a problem for a long time and is showing significant improvements in user satisfaction (Insapillo Fatama, 2023).

Many individuals encounter significant obstacles in understanding texts related to public administration (Yuan et al., 2023). These challenges may stem from struggles in deciphering lengthy sentences, technical jargon, uncommon terminology, or complex linguistic structures. Such hurdles directly impact individuals with intellectual disabilities or those with limited literacy skills. Even individuals with advanced education, such as university students specializing in various fields of study, may find themselves among those affected by reading difficulties (Alarcón et al., 2020). Public institutions are not exempt from this reality. Frequently, the content of texts directed towards citizens contains vocabulary that is challenging to comprehend, thereby complicating interpretation and the commencement of activities and administrative procedures by users (Roundy et al., 2023).

Reading comprehension is understanding a text in its entirety (Simanjuntak et al., 2024). For many people, the way a text is written can become an obstacle to understanding its content (Saggion et al., 2015). It is essential to note that complex words can present significant challenges, as their meaning is often intrinsically linked to context and cannot be easily deduced (Zaharia et al., 2021). The presence of infrequent or unknown words in the content of the texts significantly hinders the reader's understanding (North et al., 2023).

Predicting which terms may be difficult to understand for a specific group of people is known as complex word identification (CWI) (Shardlow et al., 2020). The identification of complex words involves the detection of terms within documents that could present difficulties or be confusing to understand for individuals belonging to certain groups (Rico-Sulayes, 2020).

The main purpose of public companies is to provide high-quality services to citizens. In Ecuador, specifically in the city of Guayaquil, various state institutions such as such as: 1) *Illustrious Munici-*

pality of Guayaquil (GMO)[1]. 2) *The Internal Revenue Service* (IRS)[2]. 3) *The National Telecommunications Corporation* (NTC)[3]. 4) *The National Electoral Council* (NEC)[4]. 5) *The Municipal Transit Authority* (MTA)[5]. These institutions have the responsibility of informing users about the available services and their improvements, as well as facilitating the necessary administrative procedures through various processes that must be completed by users. These institutions have a large number of users and are the ones in which we have carried out this research.

The objective of this research is to provide an essential resource to advance the study of Lexical Simplification, specifically in the identification of complex words in Spanish texts issued by various Ecuadorian public institutions. Our corpus has been annotated and evaluated by applying complexity metrics for Spanish. Additionally, we have conducted experiments using Transformers-based language models, evaluating their performance with common error metrics. The contributions of this research can be summarized as follows:

- A new corpus consists of 1,500 texts in Spanish which we have called GovAI*Ec*. The texts that make up this corpus come from various sources of Ecuadorian public service institutions. A total of 7,813 complex words identified and a total of 12,095 annotations.

- The corpus has been evaluated by lexical complexity metrics for Spanish.

- We calculated 23 linguistic features, which we combined with the encodings generated by the models based on the Transformers architecture: XLM-RoBERTa-Base, RoBERTa-large-BNE, XLM-RoBERTa-Large and BERT, with the purpose of evaluating the results obtained in our research and determining whether they support or contradict the statement formulated in our hypothesis.

Our hypothesis "The implementation of Large-scale Language Models that combine features of diverse nature, such as linguistic features and encodings, leads to better model performance, resulting in higher accuracy in both prediction and identification of complex words".

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, introduce to GovAI*Ec* corpus and the annotation process. Section 4, presents the experimentations ans results and analysis on them. Section 5, summarizes main contributions and provides some insights on future work.

2 Related Works

Previous studies on Spanish corpora creation for complex word identification can be categorized into two sections. The first section encompasses works offering background, context, and theoretical foundations, underscoring the relevance and originality of our research. The second section focuses on studies applying Lexical Complexity Measures in Spanish.

2.1 Corpora for lexical complexity in Spanish

Pitkowski and Gamarra (2009) defines a *corpus* as an extensive collection of texts, whether written or oral, containing millions of words in electronic format. An annotated *corpus* is a fundamental resource for any Natural Language Processing (NLP) task (Quevedo-Marcos, 2020).

The development of effective Natural Language Processing (NLP) tools relies heavily on the existence of large annotated corpora of texts. Although annotated corpora in English are common, extensive corpora in Spanish are less frequent. Furthermore, corpora available in Spanish often lack the necessary annotations to facilitate the development of beneficial tools (Davidson et al., 2020). Creating an annotated corpus is a time-consuming process.

[1]Illustrious Municipality of Guayaquil (GMO). The Municipal Palace of Guayaquil, also known as the Porteño town Council or simply as the Municipality, is the headquarters of the Very Illustrious Municipality of the city, that is, the Municipal Council and Mayor's Office of Guayaquil. - Available on `https://www.guayaquil.gob.ec/`

[2]Internal Revenue Service (IRS). The Internal Revenue Service is an autonomous body of the State of Ecuador, whose main function is the administration of taxes, based on a taxpayer database. Available on `https://www.sri.gob.ec/web/intersri/home`

[3]National Telecommunications Corporation (NTC). The National Telecommunications Corporation, is an Ecuadorian state telecommunications company, operating local, regional and international fixed telephone services, standard and high-speed internet access. Available on `https://www.cnt.com.ec/`

[4]National Electoral Council (NEC). The National Electoral Council of the Republic of Ecuador is the highest voting body in the country. - Available on `https://www.cne.gob.ec/`

[5]Municipal Transit Authority (MTA). The Municipal Public Company of Transit and Mobility of Guayaquil, better known simply as the Transit and Mobility Agency. - Available on `https://www.atm.gob.ec/`

Furthermore, even with human annotation, discrepancies can arise between annotators or within the same annotator, which could compromise the quality of the corpus. Therefore, a lack of supervision in the annotation process can result in a low-quality corpus (García-Díaz et al., 2020).

Saggion et al. (2015) presented the results of the Simplext project, focused on the automatic simplification of texts in Spanish. This modular system focused on syntactic and lexical simplification, based on the analysis of a manually simplified corpus for people with special needs. They carried out an evaluation using Spanish readability metrics, such as the lexical and sentence complexity index proposed by Anula (2008), as well as the readability of Spanish according to Spaulding (1956).

In his research, Segura-Bedmar and Martínez (2017) used the corpus *EasyDPL* (Easy Drug Package Leaflets), which consists of 306 leaflets written in Spanish. These brochures are manually annotated with 1,400 adverse effects of medications and their simplest synonyms, since patients often have problems understanding the sections that describe the dosage (dosage amount and prescription), contraindications, and adverse reactions to the medication providing an automated approach that helped pharmaceutical companies write drug package inserts in easy-to-understand language.

Ortiz-Zambrano and Montejo-Ráez (2017) developed the VYTEDU corpus (Videos and Transcriptions for research in the Educational field) in Spanish, obtained from the transcriptions of videos recorded during university classes. For its construction, 55 videos were filmed during classes of different careers at the University of Guayaquil. The system incorporates indicators selected by Saggion et al. (2015) and applies the seven metrics of lexical complexity for Spanish, which allowed them to analyze the complexity of the text at different levels, such as the lexical and sentence complexity index.

The annotated corpus known as *VYTEDU-CW* was introduced by Ortiz Zambrano et al. (2019). This corpus arises from the process of identifying and labeling complex words in the Spanish texts of the VYTEDU corpus, carried out by students from various disciplines at the University of Guayaquil. This resource was offered to the participants of the ALexS 2020 workshop (Lexical Analysis at SEPLN 2020[6]) as part of the second edition of IberLEF 2020[7]

Zambrano and Montejo-Raéz (2021) introduces *CLexIS*[2], a new annotated corpus in Spanish aimed at researching complex words in computational studies. Seven textual complexity metrics were used to evaluate the complexity of the texts. Furthermore, as a point of reference, two experiments were carried out to predict word complexity: one using a supervised learning approach and another using an unsupervised approach based on word frequency in a general corpus.

Ferrés and Saggion (2022) introduced ALEXSIS, the initial dataset designed to assess lexical simplification in Spanish. This data set incorporated potentially valuable details for lexical simplicity ranking and provided a higher average number of unique synonyms. "ALEXSIS facilitated a comparison of several neural methods", including their adaptation of LSBert to Spanish, along with other neural approaches that rely on pre-trained.

Alarcon et al. (2023) introduced the EASIER corpus, a valuable source that facilitates the construction of lexical simplification methods to process texts in Spanish, regardless of their specific domain. This corpus is composed of 260 documents, meticulously annotated with 8,155 words identified as complex and 5,130 words that have at least one contextually suggested synonym. To guarantee the reliability of the corpus, an agreement test between annotators was carried out, yielding a Fleiss Kappa coefficient of 0.641, indicating moderate consistency.

Ortiz Zambrano et al. (2023) introduced *LegalEc*, a novel corpus annotated with complex lexicon derived from legal content in Ecuadorian Spanish. They also outlined the compilation and annotation process in detail. To establish baseline cases for the scientific community, several experiments predicting complex words were conducted on this corpus. They extracted 23 linguistic features, which were combined with encodings generated by models like XLM-RoBERTa and RoBERTa-BNE from the MarIA project. Evaluation results demonstrated a significant enhancement in lexical complexity pre-

[6]SEPLN 2020 - Available on `http://sepln2020.sepln.org/index.php/iberlef/`

[7]IberLEF 2020: Iberian Language Evaluation Forum. Available on `https://ceur-ws.org/Vol-2664/`. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020) - Available on `https://ceur-ws.org/Vol-2664/`. Track 1: Lexical Analysis at SEPLN (ALexS). ALexS 2020: Lexicon Analysis Task @ SEPLN (Ortiz-Zambrano and Montejo-Ráez, 2020)

diction with the amalgamation of these linguistic features.

Sierra et al. (2024) presented a valuable resource in the form of an aligned parallel corpus consisting of Spanish Bible translations. This corpus comprises 11 translations of the Bible into Spanish, spanning different centuries and geographic regions, including Spain and Latin America, and representing various religious denominations, such as Protestants and Catholics. This corpus provides a valuable tool for various linguistic analyses, such as the detection of paraphrases, semantic clustering and the exploration of possible biases present in the texts specified for monolingual studies.

2.2 Measures of Lexical Complexity for Spanish

A strong indicator of writing quality lies in the use of a measure of lexical complexity, which encompasses the size, variety, and quality of vocabulary (Crossley et al., 2012). Another method to determine the lexical complexity of words in Spanish is based on the metrics proposed by Anula (2008) and Spaulding (1956). These metrics have been used in research on the simplification of texts in Spanish, such as the work carried out by Saggion et al. (2015), Ortiz-Zambrano and Montejo-Ráez (2017), Zambrano and Montejo-Raéz (2021), Ortiz Zambrano et al. (2023) among other notable examples. The formulas were proposed by Anula (2008) except the SSR formula corresponds to Spaulding (1956). For better understanding, the Table 1 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.

We have added:

ARI: Automated Readability Index.
PM: Punctuation Mark.

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

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

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

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 2.

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

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

$$ASL = N_w/N_s \quad (6)$$

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

3 The GovAI*Ec* corpus

GovAI*Ec* provides a collection of 1,500 texts obtained mainly from two sources: notifications and instructions for administrative procedures that users receive through emails or find on the websites of public institutions. GovAI*Ec* has a total of 7,813 complex words identified and a total of 12,095 annotations. The objective of GovAI*Ec* is to contribute to research on the identification of complex words in state documents of Ecuador, specifically from public institutions with the largest number of users. This corpus is invaluable for two fundamental reasons. Firstly, it makes it possible to identify terms that hinder the understanding of readers who participate in administrative processes of various organizations. Secondly, it provides a valuable resource for the scientific community, allowing progress in research within the field of Lexical Simplification in the Spanish language.

For the construction of the data set, we followed the format of the data set provided by the SemEval-2021[8] competition, for the proposal of the Task 1[9]: Lexical Complexity Prediction; and the efforts made in creating labeled corpora for research on complex word identification (Shardlow et al., 2020), (Zambrano and Montejo-Raéz, 2021),

[8] SemEval-2021 - The 15th International Workshop on Semantic Evaluation. Available on `https://semeval.github.io/SemEval2021/`

[9] SemEval 2021- Task 1: Lexical Complexity Prediction. Available on `https://semeval.github.io/SemEval2021/tasks`

(Ortiz Zambrano et al., 2023).

Each sample of the GovAI*Ec* data set contains the following fields:

- **Id:** The identification number of each record.
- **Source:** The description of the source where the text comes from, that is, of the public institution.
- **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. Zeng et al. (2024) refers to linguistic features as indicators used to describe the linguistic properties of texts. The linguistic features that we have calculated correspond to the works presented by (Shiroyama, 2022), (Ronzano et al., 2016), (Shardlow et al., 2020), (Paetzold, 2021), Mosquera (2021), (Desai et al., 2021), (Shiroyama, 2022)

  1. The absolute frequency .
  2. The number of characters of the token.
  3. The relative frequency of the word before the token.
  4. The relative frequency of the word after the token.
  5. The relative frequency of the token.
  6. The number of syllables.
  7. The position of the target word in the sentence.
  8. Number of words in sentence.
  9. The number of characters in the word before the token.
  10. The number of characters in the word after the token.
  11. The Part Of Speech category.
  12. Lexical diversity.
  13. The number of synonyms.
  14. The number of hyponyms.
  15. The number of hyperonyms.
  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.

3.1 Annotation Process

3.1.1 Description of the annotation system

A graphical user interface (GUI) was created using the Tkinter library, designed to offer an intuitive and easy-to-use experience to users of the annotation system. The user had to select the words that were difficult to understand and assign them a level of complexity, which could be neutral, difficult or very difficult. This interface provided various functionalities related to the research processes, including user registration, complex word identification, linguistic feature extraction, and data set generation.

3.1.2 Labelers selection criteria

For the selection process of users in charge of tagging complex words in public texts, a total of 30 users who had carried out processes in the institutions mentioned in the 1 section were chosen.

A selection criterion was established based on the academic level of the users made up of young people, adults, and older adults: 10 users were selected, equally distributed between men and women, that is, 5 men were chosen and 5 women. In this way, the representation of users with a basic or lower academic level was guaranteed, these being people who only finished school or dropped out, made up of young people, adults, and older adults. Likewise, another 10 users with a medium academic level were selected, we refer to those users who finished secondary school as high school graduates; and 10 additional users with a university academic level or higher.

4 Results

Several experiments were carried out for the evaluation of the GovAI*Ec* corpus to demonstrate its relevance and usefulness. Details the order of executions:

Firstly: We apply the Lexical Complexity metrics for Spanish to the GovAI*Ec* corpus.

Second: The application of the Fleiss-Kappa Coefficient as a measure of agreement to evaluate the consistency of annotations.

Third: Evaluation applying LLMs, specifically: XLM-RoBERTa-Base, RoBERTa-large-BNE, XLM-RoBERTa-Large and BERT.

4.1 Lexical Complexity Variables for Spanish

It was necessary to calculate the lexical complexity variables for Spanish in order to subsequently obtain the corpus statistics.

Some statistics on the corpus texts are presented in Table 2, while the definition of the variables is shown in table 1. It is notable that the number of rare words (N_{rw}) is considerably greater than that of less frequent words (N_{lfw}).

Table 3 presents several examples of the words identified and annotated as complex in the corpus during the GovAI*Ec* tagging process.

4.2 Inter-annotator Agreement

A total of 7,813 complex words identified and a total of 12,095 annotations in the 1,500 texts that make up the GovAI*Ec* corpus were labeled by the annotators when reviewing the public texts of the GovAI*Ec* corpus.

Below are some examples of the complex words noted by three taggers: *jurisdiction*, *hierarchization*, *climatological*, *apprehended*, *aquaplaning*.

Some examples of the words noted by 2 taggers were: *sporadic*, *scheduling*, *certification*, *homologation*, *stirrups*, *therapeutic*.

Other examples of the words selected by 1 annotator: *emanated*, *regulations*, *regulations*, *legalization*, *will establish*, *preservation*.

We applied the Fleiss-Kappa coefficient as a measure of agreement to evaluate the consistency of annotations made by multiple taggers. We obtained a value of 0.165, which, according to the reference table, indicates a slight level of agreement. It is important to highlight that the annotators were users of the mentioned public institutions and came from various academic levels. This means that for some, ignorance of certain terminology may have complicated the understanding of the notifications or the understanding according to the instructions of the steps necessary to carry out a certain bureaucratic procedure in certain state institutions in Ecuador.

The table 4 shows the number of annotations made by the annotators in each category. It is observed that the annotators of the low academic level have made more annotations compared to those of the medium level. On the other hand, scorers at the middle level have made fewer annotations in relation to those at the previous level. Furthermore, the annotators of the high level have recorded a smaller number of complex words compared to the annotators of the previous groups.

4.3 Application of complexity measures for Spanish

In the second phase of results, we evaluate the GovAI*Ec* corpus using seven complexity measures for Spanish detailed in detail in the section 2.2. See Table 5.

The application of the LC metric has been fundamental to evaluate the quality of the texts by analyzing their level of lexical complexity within the GovAI*Ec* corpus. We have obtained an average value of 28.87, which indicates considerable complexity in the evaluated texts, which corresponds to various aspects, such as lexical diversity, the length of the words and the breadth of the vocabulary present in each text. Another relevant result to highlight is the average obtained through the ASL (Average Sentence length) metric, which reveals the complexity of the texts based on the average length of the sentences. The value of 40.43 is the average value referring to the number of words per sentence, indicating greater complexity and difficulty in understanding the texts, especially for those readers with limited linguistic skills.

4.4 Evaluation applying LLMs

In this study, to meet our ultimate goal of evaluating the GovAI*Ec* corpus, models known for their robustness and effectiveness in investigating lexical complexity in Spanish texts were used, such as XLM-RoBERTa-Base, RoBERTa-large-BNE, XLM-RoBERTa-Large and BERT. hese models have been widely used to create state-of-the-art solutions for numerous tasks (Paetzold, 2021). These models were trained and evaluated using the GovAI*Ec* corpus in Spanish.

Executions were carried out with each of the models, applying 30, 50, 70 and 100 epochs. These experiments are part of a series aimed at exploring different approaches. Our strategy focuses on integrating the 23 linguistic features of the corpus with the encodings generated by previously trained

The Statistics of GovAI*Ec*

	N_{chrs}	N_w	N_{dcw}	N_{cw}	N_{lfw}	N_{rw}	N_s	N_{cs}
Mode	197.00	38.00	29.00	19.00	5.00	20.00	1.00	0.00
Median	250.00	44.00	34.00	24.00	6.00	23.00	1.00	1.00
Mean	278.00	49.37	36.94	26.75	6.97	25.18	1.31	0.55
Std.Dev	118.17	21.66	12.48	11.14	3.57	10.79	0.59	0.59
Min	93.00	15.00	12.00	9.00	0.00	7.00	1.00	0.00
Max	1024.00	192.00	105.00	96.00	26.00	84.00	5.00	3.00

Table 2: Descriptive Statistics of different counters over documents in GovAI*Ec*.

Words tagged by the annotators in the texts of the corpus GovAI*Ec*

ID	Sentence	Complexity
CNE-3432	La Secretaría General del Consejo Nacional Electoral - CNE [..] [..] remitir a la Dirección Nacional de Organizaciones Políticas, que será la encargada de emitir el informe correspondiente, [..]	0.33
CNT-4334	La CNT EP, no cobrará ningún valor por las reparaciones de los daños producidos entre la central y la caja de **dispersión** inclusive si el daño se localiza entre la caja de **dispersión** y el aparato [..]	1.00
ATM-0097	De no haberse efectuado la aprehensión del o los vehículos [..] [..] el agente fiscal podrá solicitar al Juez de Tránsito disponga las [..] cautelares **pertinentes** para la práctica de las mencionadas [..].	1.00
SRI-7274	De acuerdo a lo señalado en el Código **Tributario** Artículo 153 (Plazos para el pago), el porcentaje para el pago de la primera cuota siempre será del 20% de la obligación tributaria, por lo que este [..]	0.33
6613	Retiro Temporal, con el que debe **acudir** a la Ventanilla # 38 [..]	

Table 3: Examples of words tagged by the annotators in the texts of the GovAI*Ec* corpus.

Agreement between labelers

	Low Academic level		Middel Academic level		University Academic level	
PI	**women taggers**	**men taggers**	**women taggers**	**men taggers**	**women taggers**	**men taggers**
MTA	467	598	457	346	420	317
NEC	551	963	511	452	290	387
NTC	547	541	363	484	246	451
GMO	562	653	442	318	164	295
IRS	348	209	188	189	163	173
# tagged words	2475	2964	1961	1789	1283	1623

Table 4: Analysis of the degree of agreement between the annotators in relation to the academic level.

models. The objective is to evaluate whether this combination provides satisfactory answers to our research hypothesis.

We have carried out experiments without tuning the encoders, using only pre-trained models. Runs were performed to determine whether combining linguistic features (LF) represents an improvement over full end-to-end approaches. Integrating linguistic features involves concatenating them, after applying min-max scaling, with the embeddings resulting from the last encoding layer, and before reaching the classification header. See Figure 1.

The table 6 shows the results of the executions with the different models. We have evaluated the results based on the mean absolute error (MAE). We observe that the BERT model, particu-

Lexical Complexity Metrics for Spanish in GovAI*Ec*									
	LDI	**ILFW**	**LC**	**SSR**	**ASL**	**CS**	**SCI**	**ARI**	**PM**
Mode	29.00	33.33	27.00	244.80	38.00	0.00	19.00	19.16	3.00
Median	29.00	27.07	28.51	256.65	38.00	0.29	19.00	24.22	4.00
Mean	30.67	27.07	28.87	258.17	40.43	0.44	0.43	25.53	4.46
Std.Dev	11.04	11.00	7.25	33.47	16.83	0.46	8.46	8.44	2.83
Min	6.00	0.00	7.5	148.28	8.67	0.00	4.5	6.63	1.00
Max	96.00	76.92	58.41	464.25	177.00	1.00	89.00	94.31	23.00

Table 5: Results of the application of lexical complexity metrics for Spanish in corpus GovAI*Ec*.

Spanish Language Model* pre-trained with GovAI*Ec				
with 50 epochs				
Model	**MAE**	**MSE**	**RMSE**	**R2**
XLM-RoBERTa-Large	0.20618	0.05533	0.23521	-0.00692
XML-RoBERTa-Large ⊕ LF	0.19824	0.06221	0.24943	-0.11003
RoBERTa-Large-BNE	0.21077	0.05623	0.23712	-0.00321
RoBERTa-Large-BNE ⊕ LF	0.20399	0.05112	0.22610	0.02651
XML-RoBERTa-BASE	0.16807	0.06074	0.24645	-0.09903
XML-RoBERTa-BASE ⊕ LF	0.19492	0.04825	0.21966	0.09482
BERT	0.14641	0.05374	0.23181	0.01266
BERT ⊕ LF	0.14777	0.05151	0.22697	0.01875

Table 6: Results of the pre-trained models applying 50 epochs.

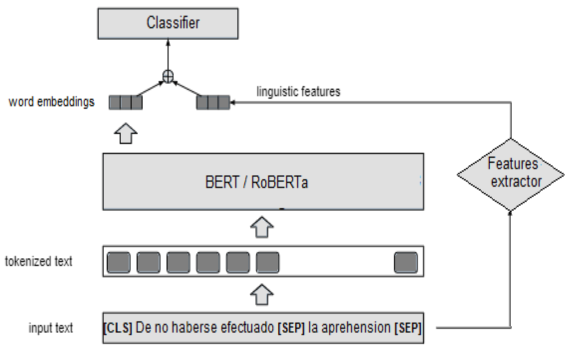

Figure 1: Process flow methodology integrating linguistic features.

larly its Spanish-adapted variant known as BERT BASE Spanish wwm uncased, offers superior performance compared to other models. This optimal performance was achieved with 50 execution epochs. It should be noted that the whole-word masking technique applied in this model contributes to its effectiveness. An interesting finding is that by incorporating the 23 additional linguistic features into the data set, even better results are obtained with the same BERT model.

5 Conclusions and Recommendations

To ensure the quality and reliability of the corpus, we carried out a comprehensive evaluation. We used measures of complexity and readability, as well as the Fleiss-Kappa coefficient to assess agreement between annotators. In addition, we perform performance tests on models based on the Transformers architecture, trained with the corpus, to validate their effectiveness in identifying complex words in Ecuadorian public documents.

The lexical complexity metrics for Spanish demonstrated that the terminology used in texts addressed to users, both in notifications and in bureaucratic processes, becomes a difficulty for recipients to understand state documents and their implications.

The experiments revealed a significant improvement in the performance of the models when integrating linguistic features obtained from the texts. Furthermore, the evaluation results indicated that the combination of these features contributes to improving the prediction of lexical complexity.

Based on the findings of this research, we recommend promoting more agile, open and innovative governments through the use of emerging technologies such as artificial intelligence and accessible websites. These technologies can improve the efficiency of government processes by facilitating the understanding of the content of public documents, which in turn guarantees the quality of services offered to citizens.

6 Limitations

The restrictions and challenges that result from the applicability of our work are:

Corpus timing:

The corpus is constructed from texts from a specific time period, which may not reflect changes in language and terminology over time. The evolution of institutional language and the emergence of new terminologies may not be represented.

Diversity of Sources:

The sources of the texts focus on a specific context, which is the government sphere. The corpus is limited to documents from certain Ecuadorian public institutions, the results could vary in other contexts.

Annotation Quality:

The annotation process for complex words may be subject to human error or tagger bias. Of course, annotators could have different criteria for identifying complex words, which could affect the consistency of the corpus.

Complexity Criteria:

The criteria used according to other research carried out to define and measure the complexity of words may not capture all dimensions of lexical complexity due to contextual, cultural or context- and language-specific factors, in our case although the language is Spanish and the public study institutions are Ecuadorian, these factors could influence the perception of complexity and not be fully considered.

Data Access and Use:

Access to certain documents sent to users through notifications by public institutions could be restricted for reasons of privacy or confidentiality, which would limit the inclusion of certain types of texts in the corpus.

Applicability of Results:

The results derived from this study may not be easily generalizable and applicable to other languages or dialects of Spanish. Regional linguistic variability could limit the generalizability of the conclusions.

References

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

Rodrigo Alarcon, Lourdes Moreno, and Paloma Martínez. 2023. Easier corpus: A lexical simplification resource for people with cognitive impairments. *Plos one*, 18(4):e0283622.

Alberto Anula. 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.

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

Sam Davidson, Aaron Yamada, Paloma Fernandez Mira, Agustina Carando, Claudia H Sanchez Gutierrez, and Kenji 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.

Abhinandan Tejalkumar Desai, Kai North, Marcos Zampieri, and Christopher 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. Association for Computational Linguistics.

Daniel Ferrés and Horacio Saggion. 2022. ALEXSIS: A dataset for lexical simplification in Spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3582–3594, Marseille, France. European Language Resources Association.

José Antonio García-Díaz, Ángela Almela, Gema Alcaraz-Mármol, and Rafael Valencia-García. 2020. Umucorpusclassifier: Compilation and evaluation of linguistic corpus for natural language processing tasks. *Procesamiento del Lenguaje Natural*, 65:139–142.

Milagros del Pilar Insapillo Fatama. 2023. Implementación de chatbot con inteligencia artificial para el mejoramiento del sistema helpdesk en el gobierno regional loreto, iquitos 2023.

Alejandro Mosquera. 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.

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

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

Jenny Ortiz-Zambrano and Arturo 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.

Jenny Ortiz Zambrano, Arturo MontejoRáez, Katty Nancy Lino Castillo, Otto Rodrigo Gonzalez Mendoza, and Belkis Chiquinquirá Cañizales Perdomo. 2019. Vytedu-cw: Difficult words as a barrier in the reading comprehension of university students. In *The International Conference on Advances in Emerging Trends and Technologies*, pages 167–176. Springer.

Jenny Alexandra Ortiz Zambrano, César Espin-Riofrio, and Arturo Montejo Ráez. 2023. Legalec: A new corpus for complex word identification research in law studies in ecuatorian spanish.

Gustavo Paetzold. 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.

Elena Fabiana Pitkowski and Javier Vásquez 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.

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

Antonio Rico-Sulayes. 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.

Francesco Ronzano, Luis Espinosa Anke, Horacio 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.

Philip T Roundy, John M Trussel, and Stephan A Davenport. 2023. The text complexity of local government annual reports. *Local Government Studies*, 49(5):1135–1156.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana 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.

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

Matthew Shardlow, Michael Cooper, and Marcos 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. European Language Resources Association.

Tomotaka Shiroyama. 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.

Gerardo Sierra, Gemma Bel-Enguix, Ameyali Díaz-Velasco, Natalia Guerrero-Cerón, and Núria Bel. 2024. An aligned corpus of spanish bibles. *Language Resources and Evaluation*, pages 1–31.

Syahdani Azhara Simanjuntak, Dian Fajrina, Nurul Inayah, and Saiful Marhaban. 2024. The correlation between students' vocabulary knowledge and reading comprehension outcome. *Research in English and Education Journal*, 9(1):10–17.

Tim Heinrich Son, Zack Weedon, Tan Yigitcanlar, Thomas Sanchez, Juan M. Corchado, and Rashid Mehmood. 2023. Algorithmic urban planning for smart and sustainable development: Systematic review of the literature. *Sustainable Cities and Society*, 94:104562.

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

María Isabel Vélez, Cristina Gómez Santamaría, and Mariutsi Alexandra Osorio Sanabria. 2022. Conceptos fundamentales y uso responsable de la inteligencia artificial en el sector público. informe 2.

Yun-Peng Yuan, Yogesh K. Dwivedi, Garry Wei-Han Tan, Tat-Huei Cham, Keng-Boon Ooi, Eugene Cheng-Xi Aw, and Wendy Currie. 2023. Government digital transformation: Understanding the role of government social media. *Government Information Quarterly*, 40(1):101775.

George-Eduard Zaharia, Dumitru-Clementin Cercel, and Mihai Dascalu. 2021. Upb at semeval-2021 task 1: Combining deep learning and hand-crafted features for lexical complexity prediction. *arXiv preprint arXiv:2104.06983*.

Jenny A Ortiz Zambrano and Arturo Montejo-Raéz. 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.

Jinshan Zeng, Xianchao Tong, Xianglong Yu, Wenyan Xiao, and Qing Huang. 2024. Interpretara: Enhancing hybrid automatic readability assessment with linguistic feature interpreter and contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19497–19505.

A Anexo 1: The files that correspond to the GovAI*Ec* corpus

The link is shared where the material corresponding to the GovAI*Ec* corpus is stored. You can also contact the authors.

https://ugye-my.sharepoint.com/:f:/g/personal/jenny_ortizz_ug_edu_ec/EjBB5s1CzjNMty6GRXUhAIsBzIM3DzHD31OPzyVBo6p9xA?e=7XEvaC