Identification of Complex Words in the Academic Domain in Spanish

Jenny Ortiz-Zambrano

University of Guayaquil, Delta Av. s/n, Guayaquil, 090510, Ecuador

Abstract

Research plays a fundamental role in universities and has significant relevance in the academic field. The purpose of this material is to disseminate the discoveries acquired through the study of the Identification of Complex Words in the academic field in Spanish by applying Artificial Intelligence techniques such as language models based on the Transformers architecture and machine learning. Participation in this symposium is of great importance, since the research results seek to generate a significant impact on society.

Keywords

Complex word, transformer based models, machine learning, spanish

1. Introduction

Reading is the key to the development of humanity, since without it an important form of communication is lost. One of the key elements of reading is readability, that is, the ease with which the content can be understood due to the writing style used [1]. The enormous growth of technology in recent decades has made its presence necessary in all spheres of citizenship. Digital information flows through different technological and digital media such as: mobile devices, internet, social networks, emails, electronic commerce, satellite tracking, among others, thus giving birth to the new Knowledge or Information Society [2] giving rise to a group of people who find it difficult to understand certain information and, therefore, cannot assimilate it.

Within the information society, people should be able to access all available information easily and simply, so improving access to written language is a topic of growing interest [3]. The success or failure of the reader in terms of understanding a text will depend on the knowledge or ignorance of the words, since many of the texts are written in a complex way, using a sophisticated and specialized vocabulary and the use of long words such as the case of information from areas such as administrative, legal, government, health, news texts or popular magazines [4], also scientific information [5], among others, which should be accessible to all members of society, especially for that large and heterogeneous target group [3] such as people with dyslexia, people with ASD [6] and it difficult to understand long and unfamiliar words [7] what you create a difficulty to understand the text in its entirety [8], also in this group are the

Doctoral Symposium on Natural Language Processing from the Proyecto ILENIA, 28 September 2023, Jaén, Spain.

jenny.ortizz@ug.edu.ec (J. Ortiz-Zambrano)

0000-0001-6708-4470 (J. Ortiz-Zambrano)

children [9], and the university students are no exception, since they are people who have a high educational level and specialized knowledge in different study subjects but, even so, they could be part of groups of people with reading disabilities [10].

The task of recognizing words in document content that is difficult or complex for a particular group of people is called complex word identification (CWI), this being the basis of many related applications language [11]. Deep learning and its innovative technologies represent edge new technologies for various natural language processing (NLP) tasks [12]. The field of NLP has made tremendous progress in the last two years, especially thanks to the Transformer architecture [13]. This architecture uses a large amount of untagged text corpus [14].

Deep learning models are significantly improved over flat machine learning models with the advent of transfer learning and pre-trained language models. The BERT and XLM-RoBERTa pre-trained deep learning language models are considered to be at the forefront of many NLP tasks [15]. Here, vocabulary complexity prediction (LCP) is no exception [16]. After comparing and analyzing deep learning approaches with other approaches, possible solutions are feasible for English and resource-poor deep learning languages, where deep models are not always available or functional. It should also be noted that the computational requirements for applying deep learning models are significantly higher than the traditional approach [17].

The field of NLP has made tremendous progress in the last two years, especially thanks to the Transformer architecture [13]. This architecture uses a large amount of untagged text corpus [14]. Deep learning models are significantly improved over flat machine learning models with the advent of transfer learning and pre-trained language models. The BERT and XLM-RoBERTa pre-trained deep learning language models are considered to be at the forefront of many NLP tasks [15].

2. Justification

Access to information is a fundamental right for all people, in particular, with regard to people with special abilities, the Convention on the Rights of Persons with Disabilities adopted by the United Nations guarantees access to information for this collective (United Nations, 2006). It is then necessary that all educational institutions and other organizations produce accessible texts for these groups of people; however, it is known that producing accessible texts is very expensive given the degree of specialization required by the editors of these contents [18].

According to the investigations carried out, at present there are no studies realized in the application of techniques for the prediction in the identification of complex words in the Academic Domain in Spanish language that serve as support in the contribution of new investigations for learning of university students. The objective of this research is to predict the complex words that cause barriers in the reading comprehension of the undergraduate students of the State University of Guayaquil through the application of several NLP techniques to evaluate which of the models presents a better performance with the sets of data. in Spanish. This need was detected after a diagnostic process, in which a data collection instrument was designed, preparing a survey. Taking into account the student population enrolled in the period 2015-2016 [19].

3. Related Work

Complex Words are considered as mono-lexical units which are difficult to read (i.e. decode), especially for poor and dyslexic readers [20]. Words identified as complex are on average longer, morphologically more unique, and less frequent in general corpora than noncomplex words [21], [22].

Lexical Complexity Prediction (LCP), is a generalization of Complex Word Identification (CWI). This task is a fundamental component in the Lexical Simplification processes. LCP, which consists of estimating the complexity of words using binary or continuous scores, represents a challenge that has been studied in various domains and natural languages [23].

In the last decades, the identification of complex words was done quite simply: by calculating the number of syllables in the word [24]. Another way was to check if the word was part of a specific list and classify it as simple or complex [25].

[26] presented a system based on the features of the word (using contextual, lexical, and semantic features) and the application of the random forest classifier to determine if the word is complex. In these systems, a total of 45 handwritten features were calculated and each word was modeled as a feature vector, applied surface functions (3 functions), dependency tree functions (8 functions), corpus-based functions (15 functions), WordNet functions (11 functions), and WordNet and corpus frequency functions (4 functions). The best results achieved were an accuracy of 0.186, a recall of 0.673, a G score of 0.750 99, and an F score of 0.292.

Surveys over the past few years have focused on complex word identification - CWI. The goal of these applications is to reduce word complexity based on the composition of the features as outlined in the work done by [27] showing an approach to a set of features in the word embedding of Gloves, InferSent and various language features obtained as predictive sources of vocabulary complexity such as word frequency, word length, or number of syllables. Then, they trained a linear regression model using different subsets of functions, obtaining as a result an MAE = of 0.0853.

[28] performed a machine learning approach based on word level and 15 language features acquired in that environment. They trained a supervised random forest regression algorithm for a set of features. Several runs were performed with different values to observe the performance of the algorithm. The best results obtained were MAE = 0.07347, MSE = 0.00938, and RMSE = 0.096871.

[29] developed a word complexity prediction system for common LCP tasks hosted on SemEval 2021. The Task Organizer distributed the CompLex Corpus [27] to participants in an enhanced version. The task was in the Lexical Semantics track which consisted of predicting the value of word complexity in context.

[30] for the detection of complex words applied a supervised learning approach using the Random Forest algorithm. To execute the system, they needed annotated data that identified the simple or complex word from several words, for which generated a total of 15 features, as in the works of [31] and [32].

[33] Several complex word prediction experiments were carried out with the LegalEc corpus, which has texts whose origin comes from the final degree works of the students of the Law degree at the University of Guayaquil, as well as from several articles of the Constitution. of Ecuador. They extracted 23 linguistic features that they combined with the encodings generated

by models such as XLM-RoBERTa and RoBERTa-BNE (from the MarIA project). The evaluation showed that the combination of these features significantly improves the prediction of lexical complexity.

4. Description of the proposed research, including the main hypotheses

4.1. 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, which translates into greater accuracy in both the prediction and identification of complex words.

5. Methodology and proposed experiments

This research work has a *quantitative* approach because it represents a set of processes and its methodology is sequential and probative. The order is rigorous, although of course, some phase [34] can be redefined. The NLP timeline presented by [35] presents three different types of approaches, which we have taken as a reference to represent the order of development of the research. See Figure 1.

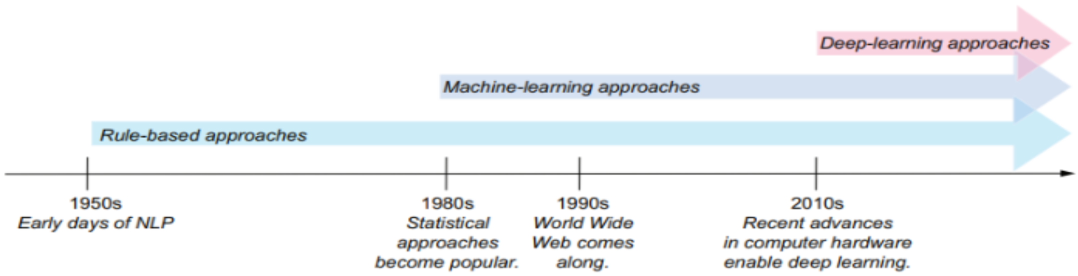

Figure 1: NLP timeline showing three different types of approaches by [35]

The following stages were identified for carrying out the research:

1. Study of the state-of-the-art in Complex Word Identification (CWI).
2. Construction of new corpora and resources.
3. Identification of resources (models and algorithms) and datasets.
4. Identification of relevant linguistic features.
5. Experiments to evaluate different models and combinations of linguistic features for CWI.

The details of each stage are set out below:

5.1. *Study of the state-of-the-art in Complex Word Identification (CWI)*

This type of research has been of vital importance in the exploration and analysis of the relevant documentary sources necessary to support the theories and formulas that support the study of the complexity of words in the Spanish language and its application from various approaches such us the machine learning and deep learning approaches.

5.2. *Construction of new corpora and resources*

As a starting point, the creation of own resources made up of the corpus was carried out: VYTEDU [19], VYTEDU-CW [36], CLexIS[2] [30], LegalEc [33]. These resources constitute the fundamental basis for carrying out experiments in the investigation of complex words in various branches of university studies in Spanish; in view of the fact that resources for Spanish in general and education in particular are scarce. The lack of resources such as parallel corpora and lexical resources represents a difficulty for the advancement of proposals in the area of Lexical Simplification, becoming an evident challenge in the case of the Spanish language [18].

5.3. *Identification of resources (models and algorithms) and datasets*

The exploration of human language generating models such as: BERT, RoBERTa and its variants, GPT-3 with their respective experiments with data sets in English, which served to analyze and evaluate the results of the experiments for the Spanish language, was carried out. Detection systems for the prediction of complex words in English and Spanish were explored, developed and evaluated by applying supervised learning algorithms, and unsupervised learning algorithms with English data sets [37]

In our unsupervised learning experiments, we apply executions with few-shot and zero-shot learning with different indications. We have also varied the values of the scenarios, specifically we are referring to our proposal presented at CLEF-2023 [37]. We observed that when determining the level of complexity (difficult, very difficult, or neutral), the model tended to identify terminology beyond its word-for-word expression in the text. After experiments, we show that this approach is important because the model can be tailored to a specific task.

5.4. *Identification of relevant linguistic features*

According to the experiments and the exploration of scientific material, we have verified that after the executions of the different applied technologies in which the linguistic features have been part of the data set, they play a fundamental role in the prediction of complex words. Therefore, features such as: word length, absolute frequency, Relative frequency, number of syllables, number of words in the sentence, relative frequency of the previous token, relative frequency of the word after the token, length of the previous word, length of the following word, target word position, lexical diversity, number of synonyms, number of hyponyms, number of hyperonyms, Part of Speech, and POS: PROPN, AUX, VERB, ADP, NOUN, NN, SYM, NUM.

5.5. *Experiments to evaluate different models and combinations of linguistic features for CWI*

For the development of the experiments, the Experimental Design was applied, combining several NLP techniques that allowed obtaining the prediction of the complexity of the words and the level of complexity of the texts. Open source applications were used such as: the Python programming language for software development. Lexical complexity metrics and evaluation measures verified by different authors such as [38], [39] and the application of various Machine Learning and Deep learning were applied, executing various language models based on the Transformers architecture for both the language English as for Spanish with the purpose of establishing an analysis of the robustness of the models in terms of the applied language.

6. Conclusions

This research focuses on the development and evaluation of the improvement of the results in the prediction of complex words aimed at the English and language by applying technological solutions with an emphasis on Deep Learning. Our model takes advantage of the combination of advanced NLP techniques applying Transformers-based deep learning models: BERT [40], XLM-RoBERTa [41] and its variants. The datasets are made up of features of a different nature: linguistic, syntactic, statistical, and semantic. The experiments were carried out with the English CompLex 2.0 corpus described in [27] and for the Spanish language, the experiments with the data sets were carried out with the corpus: VYTEDU, VYTEDU-CW, CLexIS[2], LegalEc and AdminLex.

7. Acknowledgments

I would like to express my gratitude to the thesis director, Ph.D. Arturo Montejo-Ráez for his adequate guidance, his scientific support and his ability to foster a critical discussion, which has been essential to make the most of the development of the research. I also want to express my gratitude to Ph.D. Alfonso Ureña, my tutor, for his trust in me and for giving me the opportunity to complete my doctorate at such a prestigious University. Finally, I extend my gratitude to the students of the University of Guayaquil who have collaborated at all times in the development of this research.

References

[1] A. A. Al-Ajlan, H. S. Al-Khalifa, A. S. Al-Salman, Towards the development of an automatic readability measurements for arabic language, in: 2008 Third international conference on digital information management, IEEE, 2008, pp. 506–511.
[2] B. E. S. Requena, Las tic y la educación social en el siglo xxi, Edmetic 5 (2016) 8–24.
[3] F. Dell'Orletta, S. Montemagni, G. Venturi, Read–it: Assessing readability of italian texts with a view to text simplification, in: Proceedings of the second workshop on speech and language processing for assistive technologies, 2011, pp. 73–83.

[4] S. Štajner, G. Glavaš, Leveraging event-based semantics for automated text simplification, Expert systems with applications 82 (2017) 383–395.
[5] L. Ermakova, I. Ovchinnikova, J. Kamps, D. Nurbakova, S. Araújo, R. Hannachi, Overview of the clef 2022 simpletext task 2: Complexity spotting in scientific abstracts, in: Proceedings of the Working Notes of CLEF 2022-Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th-to-8th, 2022, 2022.
[6] J. Carletta, Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus, Language Resources and Evaluation 41 (2007) 181–190.
[7] S. Bott, L. Rello, B. Drndarević, H. Saggion, Can spanish be simpler? lexsis: Lexical simplification for spanish, in: Proceedings of COLING 2012, 2012, pp. 357–374.
[8] D. Torunoglu-Selamet, T. Pamay, G. Eryigit, Simplification of turkish sentences, in: The First International Conference on Turkic Computational Linguistics, 2016, pp. 55–59.
[9] S. Štajner, R. Evans, C. Orasan, R. Mitkov, What can readability measures really tell us about text complexity, in: Proceedings of workshop on natural language processing for improving textual accessibility, Citeseer, 2012, pp. 14–22.
[10] R. Alarcon, L. Moreno, I. Segura-Bedmar, P. Martínez, Lexical simplification approach using easy-to-read resources, Procesamiento del Lenguaje Natural 63 (2019) 95–102.
[11] A. Rico-Sulayes, 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, 2020.
[12] S. Singh, A. Mahmood, The nlp cookbook: Modern recipes for transformer based deep learning architectures, IEEE Access 9 (2021) 68675–68702.
[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.
[14] J. Canete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, Pml4dc at iclr 2020 (2020) 2020.
[15] T. B. Yaseen, Q. Ismail, S. Al-Omari, E. Al-Sobh, M. Abdullah, Just-blue at semeval-2021 task 1: Predicting lexical complexity using bert and roberta pre-trained language models, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), 2021, pp. 661–666.
[16] A. Nandy, S. Adak, T. Halder, S. M. Pokala, cs60075_team2 at semeval-2021 task 1: Lexical complexity prediction using transformer-based language models pre-trained on various text corpora, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), 2021, pp. 678–682.
[17] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big? ��, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 610–623. URL: https://doi.org/10.1145/3442188.3445922. doi:10.1145/3442188.3445922.
[18] S. Bott, H. Saggion, Automatic simplification of spanish text for e-accessibility, in: Computers Helping People with Special Needs: 13th International Conference, ICCHP 2012, Linz, Austria, July 11-13, 2012, Proceedings, Part I 13, Springer, 2012, pp. 527–534.
[19] J. A. Ortiz-Zambrano, A. Montejo-Ráez, Vytedu: Un corpus de vídeos y sus transcripciones

para investigación en el ámbito educativo (2017).
[20] F. Hmida, M. B. Billami, T. François, N. Gala, Assisted lexical simplification for french native children with reading difficulties, in: The Workshop of Automatic Text Adaptation, 11th International Conference on Natural Language Generation, 2018.
[21] G. Paetzold, L. Specia, 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), 2016, pp. 969–974.
[22] S. M. Yimam, C. Biemann, S. Malmasi, G. H. Paetzold, L. Specia, S. Štajner, A. Tack, M. Zampieri, A report on the complex word identification shared task 2018, arXiv preprint arXiv:1804.09132 (2018).
[23] A. V. Abramov, V. V. Ivanov, V. D. Solovyev, Lexical complexity evaluation based on context for russian language, Computación y Sistemas 27 (2023) 127–139.
[24] G. H. Mc Laughlin, Smog grading-a new readability formula, Journal of reading 12 (1969) 639–646.
[25] E. Dale, J. S. Chall, A formula for predicting readability: Instructions, Educational research bulletin (1948) 37–54.
[26] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.
[27] M. Shardlow, M. Cooper, M. Zampieri, Complex: A new corpus for lexical complexity prediction from likert scale data, arXiv preprint arXiv:2003.07008 (2020).
[28] J. A. Ortiz-Zambrano, A. Montejo-Ráez, Complex words identification using word-level features for semeval-2020 task 1, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), 2021, pp. 126–129.
[29] M. Shardlow, R. Evans, G. H. Paetzold, M. Zampieri, Semeval-2021 task 1: Lexical complexity prediction, arXiv preprint arXiv:2106.00473 (2021).
[30] J. A. O. Zambrano, A. Montejo-Raéz, 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), 2021, pp. 1075–1083.
[31] S. Gooding, E. Kochmar, Camb at cwi shared task 2018: Complex word identification with ensemble-based voting, in: Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, 2018, pp. 184–194.
[32] P. Finnimore, E. Fritzsch, D. King, A. Sneyd, A. U. Rehman, F. Alva-Manchego, A. Vlachos, Strong baselines for complex word identification across multiple languages, arXiv preprint arXiv:1904.05953 (2019).
[33] J. A. Ortiz Zambrano, C. Espin-Riofrio, A. Montejo Ráez, Legalec: A new corpus for complex word identification research in law studies in ecuatorian spanish (2023).
[34] R. Hernández, C. Fernández, P. Baptista, Metodología de la investigación sexta edición, México DF: McGraw-Hill (2014).
[35] E. Kochmar, Getting started with natural language processing, (No Title) (2022).
[36] J. Ortiz Zambrano, A. MontejoRáez, K. N. Lino Castillo, O. R. Gonzalez Mendoza, B. C. Cañizales Perdomo, Vytedu-cw: Difficult words as a barrier in the reading comprehension of university students, in: Advances in Emerging Trends and Technologies: Volume 1, Springer, 2019, pp. 167–176.
[37] J. Ortiz-Zambrano, C. Espin-Riofrio, A. Montejo-Ráez, Sinai participation in simpletext task 2 at clef 2023: Gpt-3 in lexical complexity prediction for general audience (2023).

[38] A. Anula, 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 (2008) 162–170.
[39] S. Spaulding, A spanish readability formula, The Modern Language Journal 40 (1956) 433–441.
[40] X. Liu, P. He, W. Chen, J. Gao, Improving multi-task deep neural networks via knowledge distillation for natural language understanding, arXiv preprint arXiv:1904.09482 (2019).
[41] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2019).

A. Publications originating from the doctoral project

Ortiz Zambrano, J. A., & Montejo Ráez, A. (2017). VYTEDU: Un corpus de vídeos y sus transcripciones para investigación en el ámbito educativo.

Available on http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5512

Ortiz Zambrano, J., & Varela Tapia, E. (2018, October). Reading comprehension in university texts: the metrics of lexical complexity in corpus analysis in Spanish. In International Conference on Computer and Communication Engineering (pp. 111-123). Springer, Cham.

Available on https://link.springer.com/chapter/10.1007/978-3-030-12018-4_9

Ortiz Zambrano, J., MontejoRáez, A., Lino Castillo, K. N., Gonzalez Mendoza, O. R., & Cañizales Perdomo, B. C. (2019, March). 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 (pp. 167-176). Springer, Cham.

Available on https://link.springer.com/chapter/10.1007/978-3-030-32022-5_16

Ortiz, Jenny & Montejo-Raéz, Arturo. (2020). Barriers in Reading Comprehension of University Students: Analysis of the Complicated Words Annotated in the VYTEDU-CW Corpus. International Journal on Advanced Science, Engineering and Information Technology. 10. 1798. 10.18517/ijaseit.10.5.10809.

Available on http://ijaseit.insightsociety.org/index.php?option=com_content&view=article&id=9&Itemid=1&article_id=10809

Ortiz-Zambranoa, J. A., & Montejo-Ráezb, A. (2020). Overview of alexs 2020on lexical analysis at sepln. In Proceedings of : First workshop the Iberian Languages Evaluation Forum (IberLEF 2020).

Available on https://ceur-ws.org/Vol-2664/

Ortiz-Zambrano, J. A., & Montejo-Ráez, A. (2021, August). Complex words identification using word-level features for SemEval-2020 Task 1. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021) (pp. 126-129).

Available on https://aclanthology.org/2021.semeval-1.11/

Zambrano, J. A. O., & Montejo-Ráez, A. (2021, September). Clexis2: A new corpus for complex word identification research in computing studies.

Available on https://aclanthology.org/2021.ranlp-1.121/

Ortiz-Zambrano, J., Espin-Riofrio, C., & Montejo-Ráez, A. (2022). Transformers for Lexical Complexity Prediction in Spanish Language. Procesamiento del Lenguaje Natural, 69, 177-188.

Available on http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6438

Ortiz-Zambrano, J. A., Espin-Riofrio, C., & Montejo-Ráez, A. (2022). Combining Transformer Embeddings with Linguistic Features for Complex Word Identification. Electronics, 12(1), 120.

https://www.mdpi.com/2079-9292/12/1/120

Ortiz-Zambrano, J., Espin-Riofrio, C., & Montejo-Ráez, A. (2023). SINAI Participation in SimpleText Task 2 at CLEF 2023: GPT-3 in Lexical Complexity Prediction for General Audience.

https://ceur-ws.org/Vol-3497/paper-255.pdf

Ortiz Zambrano, J. A., Espin-Riofrio, C., & Montejo Ráez, A. (2023). LegalEc: A New Corpus for Complex Word Identification Research in Law Studies in Ecuatorian Spanish.

Available on http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6557