



Universidad de Jaén

Escuela de Doctorado

Identification of Complex Words in the Academic Domain in Spanish

Autor: Jenny Ortiz Zambrano

Director de la tesis: Ph.D. Arturo Montejo Ráez
Departamento: Informática

Fecha: 01/11/2024

ISBN:
Licencia CC

RUJA



Universidad de Jaén

Escuela de Doctorado

DOCTORAL THESIS

Identification of Complex Words in the Academic Domain in Spanish

PRESENTED BY:

Jenny A. Ortiz Zambrano

SUPERVISED BY:

Ph.D. Arturo Montejo Ráez

JAÉN, Nov 1, 2024

Dedictory

To my family, the fundamental pillar of my being, the heart that gives life to my dreams, and the strength that gives meaning to each step, whose love and support are the most valuable treasure, and my essential source of inspiration.

Jenny Alexandra Ortiz Zambrano

Acknowledgements

At this moment of academic completion and achievement, I wish to render my deepest gratitude to God, my inexhaustible source of strength and guidance. The completion of this doctoral thesis is the result of a journey full of challenges and learning, and I recognize that everything has been possible thanks to divine grace.

I want to express my gratitude to Ph.D. Luis Alfonso Ureña, my tutor, for his trust in giving me the opportunity to carry out doctoral studies at such a prestigious University.

It gives me great pleasure to express my deep gratitude to my thesis supervisor, Ph.D. Arturo Montejo-Ráez, for his expert and adequate guidance, his scientific support, constant support, for his ability to foster critical discussion, his vision, patience and dedication unwavering that has been essential to make the most of the development of the investigation throughout this investigative process and bring it to its conclusion.

I want to thank my family who, since the beginning of this arduous path, have been my constant source of inspiration and motivation. Your words of encouragement and unconditional love have sustained me in moments of exhaustion. I understand that, on many occasions, my dedication to this project kept me away from important moments in our family life. Therefore, I thank you for your patience and understanding. I know it wasn't always easy, but your constant support allowed me to focus on this important goal and finally achieve it.

I express my most sincere gratitude to the authorities of the University of Guayaquil, especially to the authorities of the Faculty of Mathematical Sciences for their invaluable cooperation, without your kind help my work would not have been possible. Their collaboration and support was important to the success of my doctoral thesis.

Last but not least, I would want to thank everyone who took part in my research for their time and cooperation. Without their assistance, my work would not have been possible. This accomplishment belongs to all of you who have contributed in some way to my doctoral thesis, not just to myself. I hope we can continue to work together on academic projects in the future.

With love and gratitude, *Jenny*

Contents

Introduction	1
Context	1
Motivation	3
Difficulty of the task	5
Objective	7
Results	7
Main contributions	8
Structure of the doctoral thesis	19
 1 Chapter I. The Lexical Complexity Problem	 24
1.1 Introduction	24
1.2 Research Hypothesis	29
1.3 Justification For The Research	29
1.4 Scientific Motivation	31
1.5 Research Objectives	32
1.5.1 General Objective	32
1.5.2 Specific Objectives	32
 2 Chapter II. Theoretical framework	 34
Foundations	34
2.1 Natural Language Processing	34
2.2 Textual Simplification	36
2.2.1 Syntactic Simplification	37
2.2.2 Complex Words Identification	39
2.2.3 Lexical Simplification	40
2.3 Lexical Complexity Prediction	41
2.3.1 Lexical Complexity Metrics	42

2.4	Artificial Intelligence and Machine Learning	43
2.4.1	Machine Learning Taxonomy	45
	Supervised Machine Learning	46
	Unsupervised Machine Learning	46
	Reinforcement Learning	47
2.4.2	Other Concepts in Machine Learning	47
2.4.3	Main Algorithms Used In Machine Learning	51
2.4.4	Artificial Neural Networks	54
	Deep Learning	56
	Large Language Model	58
	LLM - Based On Transformers	59
	Large Language Models Based On Transformers: BERT and RoBERTa	
	Models	60
	LLM Type GPT (Generators)	64
2.5	Evaluation	67
2.5.1	Evaluation Of Complex Word Identification	68
2.5.2	Evaluation Of Lexical Complexity Prediction	69
2.6	Previous Work	71
2.6.1	ALexS 2020 - Lexical Analysis in SEPLN 2020 Lexical	75
2.6.2	LCP-2021 at SemEval 2021	78
2.6.3	SimpleText at CLEF 2023	79
2.6.4	DeTermIt! at LREC-COLING 2024	83
2.6.5	SimpleText at CLEF 2024	86
3	Chapter III. Methodological Framework	92
3.1	Research Approach and Design	92
3.1.1	Research Approach	92
3.1.2	Research Design	93
3.2	Research Procedure	93
3.2.1	SOTA - State Of The Art in Complex Word Identification - CWI and Lexical Complexity Prediction - LCP	94
3.2.2	Creation Of The Corpora - The Creation Of Resources	95
3.2.3	Identification of Relevant Linguistic Features	95
3.2.4	Baseline With CREA	96

3.2.5	Machine Learning Approaches	96
3.2.6	Deep Learning Approaches - The Implementation Of The Models Based On The Transformer Architecture	96
3.3	Methodology	96
3.4	Variables	99
3.4.1	Evaluation Metrics	101
3.4.2	Features Sets	101
3.5	Available Resources	105
3.6	Information Collection And Analysis Techniques	111
4	Chapter IV. Resources, Experiments and Results	113
4.1	Introduction	113
4.2	Creation Of Resources	121
4.2.1	VYTEDU - A Corpus of Videos and Transcriptions for Research in the Education Domain	122
4.2.1.1	The Methodology	123
4.2.1.2	Analysis	128
4.2.2	VYTEDU-CW - A Corpus of Videos and Transcriptions for Research in the Education Domain - Complex Word	135
4.2.2.1	Origin Of VYTEDU-CW	135
4.2.2.2	Related Work	135
4.2.2.3	Construction Process Of The VYTEDU-CW Corpus	136
4.2.2.4	Corpus Statistics VYTEDU-CW	139
4.2.2.5	Analysis	141
4.2.3	CLexIS ² - Lexical Complexity in Computer Systems and Software Engi- neering Careers	144
4.2.3.1	The Methodology	145
4.2.3.2	About Annotators	147
4.2.3.3	Labeling Process	147
4.2.3.4	Lexical Complexity Of The Corpus	149
4.2.4	LegalEc	153
4.2.4.1	Motivation	153
4.2.4.2	The <i>LegalEc</i> Corpus	153
4.2.4.3	Annotation Process	155

4.2.4.4	Inter-annotator Agreement	156
4.2.4.5	Analysis Of Textual Complexity	156
4.3	Proposed Solution To The Problem Of Detecting Complex Words	159
4.3.1	Complex Word Identification Based ON CREA	160
4.3.2	Applying Linguistic Features With Classical Machine Learning Algorithms In English Dataset	161
4.3.3	Applying Linguistic Features With Classical Machine Learning Algorithms In Spanish Dataset	165
4.3.4	Applying Linguistic Features With Transformers Embeddings (pre-trained) Executed On Various Classic Machine Learning Algorithms In Spanish Dataset	167
4.3.5	Applying Linguistic Features With Transformers Embeddings (pre-trained) Executed On Various Classic Machine Learning algorithms In English dataset	177
4.3.6	Applying Linguistic Features With Transformers Embeddings (tuned) In Spanish And English Corpora (End-to-End Approach)	182
4.3.7	Applying Prompting (unsupervised): Zero-Shot, Few-Shot, and Chain-Of- Thought	196
5	Chapter V. Conclusions	231
5.1	Main Conclusions	231
5.2	Discussion	235
5.3	Scientific Evaluation And Future Work	239
5.4	Limitations	248
6	Resumen escrito en Español. Identificación de las Palabras complejas en el Dominio Académico en Español	249
6.1	Introducción	249
6.2	Hipótesis de Investigación	252
6.3	Justificación de la Investigación	252
6.4	Motivación Científica	254
6.5	Objetivos de la Investigación	255
6.5.1	Objetivo General	255
6.5.2	Objetivos Específicos	255
6.6	Metodología	256

6.7	Resultados	257
6.7.1	SOTA - Estudio del Estado del Arte en Identificación de Palabras Complejas (CWI)	258
6.7.2	Identificación de Características Lingüísticas Relevantes	258
6.7.3	Creación de Corpus - La Creación de Recursos	258
6.7.4	Línea de Base con CREA	259
6.7.5	Enfoques de Aprendizaje Automático	259
6.7.6	Enfoques de Aprendizaje Profundo: la implementación de modelos basados en la arquitectura Transformer	260
6.8	Conclusiones	261
6.9	Limitations	269
	Bibliography	270
A	Contributions and participations in workshops and international conferences	289
A.1	Contributions	289
A.1.1	JCR-Indexed Journal	290
A.1.2	International Conferences - Main Publications	290
A.1.3	International Workshop	292
A.1.4	Other Publications	294

List of Figures

1	Publications made throughout the development of the research.	12
2	EwLex Interface.	13
3	EwLex Interface.	14
4	GovAIEasy Interface.	15
1.1	Example of text with infrequent words addressed to general users.	25
1.2	Example of a text posted by a public institution with rare words addressed to grandparents of children of deceased parents.	26
2.1	Artificial Intelligence domain areas by Rouhiainen, 2018.	39
2.2	Artificial Intelligence domain areas by Rouhiainen, 2018.	44
2.3	Scheme of the relationship between Artificial Intelligence, Machine Learning, and Deep Learning by Torra, 2011.	45
2.4	Machine learning types by Als Salman, 2018.	45
2.5	Supervised learning Workflow by Mahesh, 2020.	46
2.6	Classic regression and classification algorithms by Huertas Mora, 2020.	49
2.7	Perceptron Neuron Model with Hard-Limid Transfer Function by Valbuena, 2021.	56
2.8	Topology of a Multilayer Perceptron by Valbuena, 2021. Model taken and edited from Hopgood, 2021.	56
2.9	The Transformer - model architecture by Vaswani et al., 2017.	60
2.10	Input embeddings by Rothman, 2021.	62
2.11	Prompting mechanisms in large language models by Miao et al., 2024.	65
2.12	Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted (Wei et al., 2023).	68
3.1	NLP timeline showing three different types of approaches by Kochmar, 2022.	93

3.2	Combining encodings of pre-trained language models with different linguistic features.	104
4.1	The tasks covered by our research based on the proposed structure in the timeline of Kochmar, 2022.	114
4.2	The methodological process carried out during the development of the research on the identification of complex words in the academic domain in Spanish. . . .	114
4.3	Scientific production carried out during the development of the research.	120
4.4	The own resources created by authors	121
4.5	Main hardware and performance characteristics of the Samsung S4 (mobile) device.	124
4.6	The VYTEDU corpus data model.	124
4.7	Examples of video files from the VYTEDU corpus.	125
4.8	Execution of the application for the calculation of the metrics of the lexical complexity in the VYTEDU corpus.	130
4.9	EIL Software.	137
4.10	EIL Software: Complex word labeled system.	138
4.11	EIL Software - Labeling of difficult words identified in the text.	138
4.12	Students labeling difficult words in texts.	139
4.13	Students labeling difficult words in texts.	139
4.14	The data library of the corpus VYTEDU-CW.	140
4.15	LC in the texts of Software Engineering and Computer Systems careers.	152
4.16	LC in the texts of Software Engineering and Computer Systems careers.	152
4.17	Applied methodology for the assignment of <i>LegalEc</i> texts.	156
4.18	Representation of the explored solutions to the problem of detecting complex words.	159
4.19	Process for the detection of complex words in Spanish in the corpus <i>CLexIS²</i> using CREA.	160
4.20	Relationship between the predictions of a model and the actual classes of a dataset.	161
4.21	Prediction of lexical complexity by applying the Random Forest Regressor with linguistic features algorithm by Zambrano and Montejo-Ráez, 2021.	164
4.22	Prediction of word complexity applying Random Forest algorithm in Spanish dataset.	166
4.23	Representation of the workflow of the process for LCP.	171
4.24	Architecture of the system.	178

4.25 Process flow methodology integrating linguistic features with embeddings from the Transformer model.	182
4.26 Integration of HCF and increase in times.	185
4.27 Process flow methodology integrating linguistic features.	194
5.1 Integration of HCF and increase in times.	234
6.1 Cronología de la PNL que muestra tres tipos diferentes de enfoques de Kochmar, 2022.	256
6.2 Integración de HCF e incremento de épocas.	260
A.1	290

List of Tables

1	Several examples of words labeled as complex.	4
2	Several examples of words labeled as complex.	5
3	Several examples of words labeled as complex.	5
4	Several examples of words labeled as complex.	6
5	The activities illustrate the progress of research and the dissemination of knowledge throughout the project.	19
2.1	Definition of the basic variables used in complexity metrics	43
2.2	Interpretation of the SSR metrics by Camposa et al., 2020.	43
2.3	The statistics for CompLex. Each cell shows three values, which are split according to the statistics for ‘All’ / ‘Single Words’ / ‘Multi Words’.	73
2.4	Turkish CWI Dataset.	75
2.5	ALexS - Official results from (Ortiz-Zambrano and Montejo-Ráezb, 2020).	77
2.6	Final results of the LCP task on the single words dataset according to MAE.	79
2.7	SimpleText Task 2.1: Results for the official runs on train and test sets. The evaluation is done based on the terms with their limits correctly detected.	82
2.8	SimpleText Task 2.2: Results for the official runs on the test set.	82
2.9	SimpleText Task 2.2: Results for the official runs on the abbreviation expansion task.	83
2.10	Volumetric information for the single and multiple words in each subcollection of the CompLex dataset.	84
2.11	Categories on the Likert scale proposed by (Shardlow, Cooper, and Zampieri, 2020b).	85
2.12	Examples form the CompLex dataset where the complex word is highlighted in bold.	85
2.13	Summary of team approaches and models used in the task.	88
2.14	Results for CLEF 2024 SimpleText Task 2 for all terms.	89
2.15	Results for CLEF 2024 SimpleText Task 2 for difficult terms.	90

2.16	Results for CLEF 2024 SimpleText Task 2 for the quality of definitions	91
4.1	Implementation of statistical formulas in the results from the Lexical Complexity in the archives of the VYTEDU corpus.	131
4.2	Calculation of the SSR-Spaulding's Spanish Readability.	132
4.3	Calculation of the LC-Lexical Complexity and SSR-Spaulding's Spanish Readabil- ity formulas.	132
4.4	Calculation of the Sentence Complexity Index - SCI.	133
4.5	Calculation of the ARI formula - Automated Readability Index.	134
4.6	Calculation of the PM formula - Punctuation Marks.	134
4.7	The statistics of the corpus VYTEDU-CW.	140
4.8	Contents of the <i>CLexIS²</i> corpus.	146
4.9	Descriptive Statistics of different counters over documents in <i>CLexIS²</i>	146
4.10	Definition of the columns in Table 4.9.	146
4.11	Examples from the <i>CLexIS²</i> dataset.	148
4.12	Examples from the <i>CLexIS²</i> dataset (translated).	148
4.13	Total number of complex words annotated by the taggers per semester in the careers of Software and Engineering in Computer Systems.	149
4.14	Lexical complexity measurements by subject.	150
4.15	Descriptive Statistics of different counters over documents in <i>LegalEc</i>	154
4.16	Definition of the columns.	155
4.17	Examples of words tagged by the annotators in the texts of the <i>LegalEc</i> corpus. .	157
4.18	Examples of words tagged by the annotators in the texts of the <i>LegalEc</i> corpus (translated).	157
4.19	Number words tagged as complex agreed by the different number of annotators. .	158
4.20	Interpretation of the Fleiss Kappa statistic by Landis and Koch, 1977.	158
4.21	Interpretation of the SSR metric by Camposa et al., 2020.	158
4.22	Results of the application of lexical complexity metrics for Spanish in corpus <i>LegalEc</i>	159
4.23	Results obtained in the execution of the algorithm in the training set of the simple words.	165
4.24	Final results of the Lexical Complexity Prediction task on the single words dataset by Shardlow et al., 2021a.	165
4.25	Feature sets.	171

4.26 Results of the model BERT pre-trained with features of different nature.	172
4.27 Results of the model BERT tuned with features of different nature.	173
4.28 Results of the model XLM-RoBERTA pre-trained with features of different nature.	174
4.29 Results of the model XLM-RoBERTA fine-tuned with features of different nature.	174
4.30 Best results with pre-trained models.	175
4.31 Results of the RoBERTa-large-BNE pre-trained model with features of different nature.	175
4.32 Results of the RoBERTa-large-BNE model fine-tuned with features of different nature.	176
4.33 Best results with fine-tuned models.	176
4.34 Summary of best results on the CLexIS ² dataset.	177
4.35 Description of the explored feature sets.	180
4.36 Results for different feature combinations, models, and classifiers.	181
4.37 Results of the pre-trained models applying 10 epochs.	183
4.38 Results of the pre-trained models applying 30 epochs.	184
4.39 Results of the pre-trained models applying 50 epochs.	184
4.40 Best Results of the execution of the pre-trained models through the MAE metric.	185
4.41 Summary of the best results on the CLexIS ² corpus.	185
4.42 Summary of the best results on the CompLex corpus.	186
4.43 Results of the fine-tuned end-to-end approach.	186
4.44 Descriptive Statistics of different counters over documents in GovAIEc.	191
4.45 Definition of the columns in Table 4.44.	192
4.46 Examples of words tagged by the annotators in the texts of the GovAIEc corpus.	192
4.47 Analysis of the degree of agreement between the annotators in relation to the academic level.	193
4.48 Results of the application of lexical complexity metrics for Spanish in corpus GovAIEc.	193
4.49 Results of the pre-trained models applying 50 epochs.	194
4.50 GPT-3 model details.	199
4.51 Predictions generated by applying GPT-3 with <i>few-shot learning</i>	202
4.52 The Experiments applying GPT-3 with Zero-shot learning.	204
4.54 SINAI results for task 2.2.	205

4.53 SINAI results for task 2.1. The ranking reached over the 34 runs submitted by different teams is shown. The results obtained by the best submitted run are in the last row.	205
4.55 SimpleText Task 2.1: Results for the official runs Ermakova et al., 2023.	207
4.56 SimpleText Task 2.1: Results for the official runs on train and test sets. The evaluation is done based on the terms with thier limits correctly detected Ermakova et al., 2023.	208
4.57 Volumetric information for the single and multiple words in each subcollection of the CompLex dataset.	208
4.58 Categories on the Likert scale proposed by (Shardlow, Cooper, and Zampieri, 2020b).	209
4.60 GPT-3 Model Configuration.	210
4.61 Prompt example	211
4.62 Methods applied to calculate the level of lexical complexity.	214
4.63 Probabilities associated with the predictions generated by the GPT-3 model that correspond to the category of the complex word.	214
4.64 Standard applied for the construction of the prompt variants.	215
4.65 The results applying the davinci-002 and davinci-003 models and few shots learning approach.	215
4.66 The results of the probabilities associated with the level of complexity predictions generated by the GPT-3 model applying the davinci-002 and few shots learning approach.	216
4.67 GPT-4-Turbo model details.	220
4.68 Predictions generated by applying GPT-4-Turbo with zero-shot learning in Sub-task 2.1-Prompt PRM_ZS_TASK2_V1.	226
4.69 Predictions generated by applying GPT-4-Turbo with Zero-Shot learning in Sub-task 2.2 - Prompt PRM_ZS_TASK2_V1.	227
4.70 Results for CLEF 2024 SimpleText Task 2 for the quality of definitions.	228
5.1 Best results applying features with Classical Machine Learning algorithms presented in various scientific events.	235
5.2 Running the pre-trained models on the Spanish datasets.	239

Abstract

The identification of complex words within academic texts is a critical task for enhancing reading comprehension, particularly for non-native speakers and individuals with reading difficulties. This doctoral thesis focuses on the “Identification of Complex Words in the Academic Domain in Spanish” aiming to develop and evaluate advanced methodologies to accurately identify and assess the complexity of words in academic texts. The research is structured around several key objectives: the integration of linguistic features into both classical machine learning algorithms and deep neural network models, the exploration of unsupervised techniques such as auto-regressive models for Lexical Complexity Prediction (LCP), and the development of new reference corpora as essential resources for linguistic complexity research. Initially, the study identifies and calculates a wide range of linguistic features, including morphological, syntactic, and semantic characteristics, which are crucial for capturing the complexity of words. These features are then integrated into classical machine learning algorithms such as Support Vector Machines (SVM), Decision Trees, and Random Forests, as well as advanced deep learning architectures like Transformer-based models. To evaluate the performance of these models, extensive experiments are conducted using datasets in both Spanish and English. The results demonstrate that the integration of linguistic features significantly enhances the accuracy and effectiveness of both classical and deep learning models in identifying complex words and predicting their lexical complexity. The hybrid approach of combining linguistic features with deep learning models proves particularly effective, offering a robust solution for analyzing linguistic complexity in multilingual academic contexts. Moreover, the research explores the capabilities of large generative pre-trained language models, such as GPT-3, for unsupervised lexical complexity prediction. By leveraging auto-regressive models, the study examines their ability to generate high-quality predictions and contextual explanations for complex words, showcasing their potential in addressing the task without extensive labeled datasets. The development of new reference corpora in Spanish is another significant contribution of this thesis. These corpora provide valuable resources for future research in linguistic complexity,

enabling the development of more accurate and reliable models. In conclusion, this thesis presents a comprehensive approach to the identification of complex words in academic texts, integrating linguistic features with both classical and advanced machine learning techniques, and exploring the potential of unsupervised models. The findings highlight the effectiveness of these methods in improving reading comprehension and accessibility, paving the way for future advancements in the field of computational linguistics and language education.

Keywords: Lexical complexity prediction, linguistic features, machine learning algorithms, transformer-based models, Spanish language, academic domain.

Introduction

Context

Reading, an essential skill for academic success (Crossley et al., [2023](#)), emerges as the fundamental key to the progress of humanity, constituting an essential element of communication. In this context, readability stands as one of the fundamental pillars of reading, representing the ease with which the content can be understood thanks to the writing style used. Understanding and valuing readability not only enriches the reading experience, but also preserves an important avenue for the exchange of ideas and knowledge (Al-Ajlan, Al-Khalifa, and Al-Salman, [2008](#)).

The dizzying technological advance of recent decades has given rise to an urgent need to integrate it into all areas of daily life. Digital information spreads through a wide variety of technological and digital channels, ranging from mobile devices, the Internet, and social networks to web applications, email, e-commerce, satellite tracking systems, and an even broader range of innovations. This diverse digital fabric has woven an interconnected network that redefines the way we interact with the world, consolidating the presence of technology as a central element in our daily lives. This phenomenon has led to the emergence of a new era known as the Knowledge or Information Society (Requena, [2016](#)). In this context, groups of individuals face challenges in understanding certain types of information, which limits their ability to fully assimilate it. Not all information is equally accessible to everyone, due to various reasons. For example, some people with cognitive illnesses or language difficulties find the structures used in these texts complicated. This challenge also occurs in those who are immersed in the process of learning foreign languages (Gonzalez-Dios, [2017](#)). The information accessibility gap highlights the importance of addressing and overcoming barriers to full understanding, thereby ensuring that knowledge is accessible to all individuals, regardless of their circumstances or abilities.

In the current information age, it becomes imperative that people can access all available

knowledge in an accessible and direct manner. In response to this need, improving the accessibility of written language has emerged as a concern of increasing relevance in our society. By facilitating access to written information, we not only promote inclusion and learning but also empower people to pursue knowledge and actively participate in the ever-evolving digital society (Dell’Orletta, Montemagni, and Venturi, 2011). This approach not only enriches individual experience but also contributes to collective development by ensuring that information is an accessible and transformative resource for all.

A reader’s success or failure in understanding a text is intrinsically linked to his or her familiarity or unfamiliarity with the words used in it. Texts frequently adopt a complex structure and make use of sophisticated and specialized vocabulary. This characteristic is notably evident in fields such as administration, legal, government, and health, as well as in journalistic texts or scientific journals. The lexical complexity and terminological specialization present in these contexts demand more advanced linguistic competence, highlighting the importance of addressing the knowledge gap and facilitating understanding to ensure equitable access to knowledge in all areas. These disciplines often incorporate technical terms, jargon, or extensive expressions, which can constitute significant obstacles for those who are not familiar with this type of language (Štajner and Glavaš, 2017).

This challenge is featured prominently in scientific information (Ermakova et al., 2022a), a vital resource that should be available to all members of society, especially for diverse and large groups, such as people with dyslexia, those facing disorder of autism spectrum (ASD) (Carletta, 2007), and who may have difficulties understanding long and unknown words (Bott et al., 2012), which generates complications in the comprehensive understanding of the text (Torunoglu-Selamet, Pamay, and Eryigit, 2016). This problem also affects children (Štajner et al., 2012) and university students, despite their high educational level and specialization in various areas of study. Even among those with a solid academic background, some could be part of groups with reading difficulties (Alarcon et al., 2019). Overcoming these barriers is essential to ensure that information, regardless of its complexity, is accessible and understandable to everyone.

Motivation

On many occasions, difficulties in understanding the content of a text are faced through tactics such as taking notes, detecting important words reviewing parts that generate confusion, or seeking explanations when doubts or misunderstandings arise. Under various circumstances, understanding the content of a text can be challenging, especially for college students with disabilities or reduced levels of reading ability.

Access to information stands as a fundamental right that concerns all people and is of particular importance for those individuals with special abilities. In this context, the Convention on the Rights of Persons with Disabilities, adopted by the United Nations in 2006, establishes safeguards to ensure equitable access to information for this group, thus underlining the need to promote inclusion and equal opportunities in all spheres of life (Dobransky and Hargittai, 2006). Therefore, all educational institutions and other organizations must commit to creating textual content that is easily accessible and understandable to all. However, it is important to recognize that the production of accessible texts can be a costly process due to the specialization required on the part of the editors of such content. It is essential to address this challenge to ensure that information is accessible regardless of economic barriers, and to promote inclusion and equality in the dissemination of knowledge.

In the current landscape of natural language processing research applied to the university academic context in Spanish, there is a notable lack of studies focused on the identification of complex words, especially in the country where this research was carried out (Ecuador) where, despite the growing relevance of technology and the increasing availability of NLP (acronym for Natural Language Processing) tools, it seems that this specific area has remained largely unexplored to date. The use of NLP tools to identify and simplify complex terms could significantly improve students' comprehension and performance. However, the lack of research in this area suggests the need for the application of artificial intelligence algorithms and the development of tools in the university academic environment in Spanish. When dealing with large language models, the task of complex word identification is a meta-cognitive related tasks, as it requires that the models evaluates terminology at a higher level, beyond semantics (Mahowald et al., 2024)

After conducting research at the University of Guayaquil in January 2017, it was detected

that a large part of the academic units do not implement accessible technological tools that contribute to the learning of the student group. Additionally, the need to develop an automatic text simplification system for the University of Guayaquil was demonstrated after conducting a survey taking into consideration the student population enrolled in the period 2015-2016, consulting 2000 students from all branches taught in the different faculties that make up the above. The results indicate that the student community highly values the need to have technological tools that provide monitoring and understanding of the academic content taught by teachers in the classrooms and that facilitate the learning of students who have special abilities.

The following table shows illustrative examples of annotations prepared by Computer Systems Engineering students and Software students in texts corresponding to their respective academic disciplines. This process shows the identification and annotation of complex words that may represent obstacles to reading comprehension. Words identified as complex are highlighted in bold, thus providing a clear overview of terms that might require greater attention and understanding from students. This approach makes visible the specific difficulties that students of these disciplines may face when encountering specialized terminology, thus facilitating the implementation of pedagogical strategies to improve the understanding of the contents.

Note: The original text is written in Spanish, so we have included the respective translation into the working language of this thesis.

Words identified and annotated in the texts by students of Computer Systems and Software Engineering careers		
No. text	Sentence	Complexity
text33	Dividir el trabajo en paquetes poco acoplados	0,2
text74	¿Cuáles son las clases de sujetos del derecho? Existen: sujeto activo y sujeto pasivo	0,2
text72	Recuerden que para hacer el proceso léxico necesita una secuencia de caracteres para el token	0,8
text54	En el onceavo principio de materialidad , este [..] [..] las transacciones de poco valor significativo	0,6
text80	[..] Establecer una política formal que especifique la periodicidad y las características [..]	0,4

TABLE 1: Several examples of words labeled as complex.

In addition to the previously presented examples in Tabla 3 and Table 4, another set of cases is included to illustrate the lexical complexity in academic texts in Spanish. This new example complements the analysis and highlights the diversity of terms that may hinder comprehension,

**Words Identified and Annotated in the Texts by Students of
Computer Systems and Software Engineering Careers**

Text No.	Sentence	Complexity
<i>text33</i>	<i>Dividing the work into loosely coupled packages</i>	0.2
<i>text74</i>	<i>What are the types of legal subjects?</i>	
	<i>There are: active subject and passive subject</i>	0.2
<i>text72</i>	<i>Remember that to perform the lexical process</i>	
	<i>you need a sequence of characters for the token</i>	0.8
<i>text54</i>	<i>In the eleventh principle of materiality, this [..]</i>	
	<i>[..] the transactions of little significant value</i>	0.6
<i>text80</i>	<i>[..] Establishing a formal policy that specifies the</i>	
	<i>periodicity and the characteristics [..]</i>	0.4

TABLE 2: Several examples of words labeled as complex.

particularly in specialized environments such as education and professional fields. The Table 3 presents other examples of complex words tagged in the texts of the LegalEc corpus by Law students, these words are those that are in bold (the creation of the LegalEc corpus was part of the development of this study).

Words identified and annotated in the texts by Law students

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 3: Several examples of words labeled as complex.

Difficulty of the task

However, the difficulty of the texts not only constitutes a challenge for humans but also impacts on advanced applications of Natural Language Processing (NLP). These applications do

Words identified and annotated in the texts by Law students		
Text No.	Sentence	Complexity
00042	The publication of advertisements that promote [.] sexism , religious or political intolerance, [.] is prohibited	0.33
00077	The State will be civilly liable for the damages and [.]	0.33
00078	Preserve the inalienable ownership of their community lands, which shall be unalienable , [.]	0.44
00152	Contributions, subsidies , and grants that have been granted in their favor by public and private institutions, [.]	0.89
00230	[.] and even after being approved, they will have open access to justice, against any injury or harm [.]	0.33
00234	Article 75 Constitution of the Republic of Ecuador [.] and to effective, impartial, and expeditious protection of their [.]	0.56
00241	If by an act of partition several people are assigned real estate or part [.] that was previously held pro indiviso , [.]	0.89
00279	A factual error invalidates consent when [.] as if one of the parties understood it as a loan , [.]	0.89

TABLE 4: Several examples of words labeled as complex.

not manage to efficiently process large and complex sentences, therefore, they seek to improve their performance by automatic simplification (Gonzalez-Dios, 2017).

To achieve adequate understanding and effective handling of complex words both in academic and other contexts, it is essential to address the challenges that can arise for many reasons. “Some of the common difficulties that have arisen in our analysis regarding complex words can be summarized into the following aspects:”

1. *The context*: Refers to the cultural and linguistic context: Complex words are likely influenced by cultural and linguistic factors, which would be a possibility that would complicate the investigation.
2. *Ambiguity*: Complex words often have multiple meanings or interpretations, which can make it difficult to understand their precise use in various contexts. This requires careful consideration and analysis.
3. *Area of study*: A complex word in one discipline may not be complex in another.
4. *Polysemous words*: Some complex words are polysemous, that is, they have different possible meanings depending on the context in which they are used.

Objective

The central purpose of this research project is to identify the linguistic complexity present in Spanish academic university texts. To achieve this goal, we will explore several deep learning models, as well as apply a combination of widely recognized techniques in the field of Natural Language Processing, including Fine-tuning and Masked Language Modeling. Additionally, we will implement several deep learning models specifically designed for NLP tasks, such as innovative languages based on the Transformers architecture, including BERT, XLM-RoBERTa-base, XLM-RoBERTa-large, RoBERTa-large-BNE, GPT and its successor versions. This multidimensional approach will allow us to comprehensively explore and assess the complexity inherent in linguistic content in the academic context, providing a detailed and advanced perspective on the nature of such complexity in the Spanish language.

These models, used as both text and word encoding techniques, play an important role in the creation of the datasets. In this process, we were able to integrate features of different nature, which allowed us to prepare heterogeneous datasets for application in various classical machine learning algorithms. This heterogeneous feature integration approach provides a solid foundation for effective executions in a variety of contexts.

Results

As a relevant part of this research, we maintain that the dissemination of the results is intrinsically linked to the crystallization of the research objective as an essential component of this process, which has allowed us to share our advances in the field of Lexical Simplification of complex words in Spanish that we intend to serve as a benefit to the scientific and academic community so that the most recent research is known and to use the advances achieved to generate possible implementations as a solution to the problem of the complexity of words in the academic community, which will be beneficial for the student body.

The findings, data, and conclusions that we have obtained as a result of the experiments carried out in our study have provided us with concrete and quantitative information. This information has been meticulously compiled and analyzed to answer the research questions posed in the context of our project. It is important to note that our results in this field of study

have been shared in world-renowned congresses dedicated to the dissemination of high-quality research in Natural Language Processing.

The following section presents the main contributions and Chapter IV presents a detailed summary of the achievements achieved in the creation of the corpora, the execution of models based on the Transformers architecture, the application of NLP techniques, applied in the development of the experiments. In addition, all scientific production generated during the research is comprehensively covered. This chapter offers a complete and consolidated view of the progress and results obtained in each phase of the project, providing an essential guide to understanding the scope and contribution of the research in its entirety.

Main contributions

During the development of our research, we have made significant contributions that have considerably enriched the topic of the identification of complex words in the academic field in Spanish. Our contributions range from the creation of our resources to the implementation of various LLM adjustment techniques in data sets along with the integration of linguistic features, applied not only in Spanish but also in English, which has been very useful to evaluate the performance of the applied models, and finally the application of new cutting-edge technological trends such as unsupervised models such as the GPT-3 model.

Moreover, through the long path of this process of experimentation, our research has shed light on fundamental issues in the prediction of complex words and has provided valuable starting points for future research in the field of Lexical Simplification. These contributions have not only broadened our knowledge in the area but have also had a potentially beneficial impact on the academic community and society at large.

Below, we present a list of notable international events in the field of computational linguistics and natural language processing, in which we have had the honor of participating and contributing to the submission and exhibition of our solution proposals. Events we have participated in include:

SEPLN¹ - *Sociedad Española para el Procesamiento del Lenguaje Natural*

It is a scientific association that aims to promote research in the field of Natural Language Processing or Human Language Technologies. SEPLN has received a certificate of excellence from the FETYC and has been included in the ESCI index, being ranked by Google Scholar with an h-index of 11. SEPLN holds an annual congress that serves as a meeting point for the different research groups working in the area of natural language processing.

Year 2017 - At the XXXIII International Congress Of The Spanish Society For The Processing of Natural Language - SEPLN 2017² was presented to VYTEDU (Ortiz-Zambrano and Montejoráez, 2017), which is made up of texts from several careers at the University of Guayaquil. The reference of the texts is that they correspond to transcriptions of the classes recorded inside the classrooms while they were being taught by the teachers.

RanLP³ - Recent Advances in Natural Language Processing.

The International Conference Recent Advances in Natural Language Processing proceedings are indexed by Scopus - Recent Advances in Natural Language Processing.

Year 2021 - This congress was held in online mode in the year 2021. In this congress was presented CLexIS² (Ortiz-Zambrano and Montejoráez, 2021), a corpus that gathers the texts of the careers of Computer Systems Engineering and Software Engineering. Its particularity is that the texts correspond to the transcripts of the classes recorded by the teachers whose modality was virtual due to the pandemic of coronavirus.

Year 2023 - The LegalEc corpus was presented at the XXXIX International Congress of the Spanish Society for Natural Language Processing - SEPLN 2023⁴. This corpus is based on the Final Degree Projects of the students of the Law School of the University of Guayaquil, as well as on several articles of the Constitution of the Republic of Ecuador. What distinguishes this

¹SEPLN - Sociedad Española para el Procesamiento del Lenguaje Natural. <http://www.sepln.org/sepln>

²SEPLN 2017. Held in the city of Murcia - Spain. Available in <https://sepln2017.um.es/>

³RanLP 2021 - Internacional Conference Recent Advances in Natural Language Processing. Available in <https://ranlp.org/ranlp2021/start.php>

⁴SEPLN 2023. Celebrated in Jaén - Spain (Ortiz-Zambrano, Espin-Riofrio, and Montejoráez, 2023a)) Available in <http://sepln2023.sepln.org/>

corpus is that its texts are written in Ecuadorian Spanish.

IberLEF⁵ - *Iberian Languages Evaluation Forum*

IberLEF aims to encourage the research community to organize competitive text processing, comprehension, and generation tasks to define new research challenges and establish new state-of-the-art results for the Natural Language Processing community, involving at least one of the following Iberian languages: Spanish, Portuguese, Catalan, Basque or Galician.

Year 2020 - IberLEF 2020 coincided with the SEPLN Conference, which in 2020 was held online (due to COVID-19). For this international workshop, our initiative was the proposal of the Task on Lexical Analysis which we called ALexS 2020.⁶ - Lexical Analysis at SEPLN 2020, which was held in Malaga (Spain) in the framework of the second edition of IberLEF (Iberian Languages Evaluation Forum) which combined the efforts of the IberEval and TASS workshops. We provided participants with the VYTEDU-CW corpus (Ortiz-Zambrano et al., 2019), a corpus in Spanish, containing texts tagged by 430 annotators (students) corresponding to different careers. The task was to predict the complexity value of words in context.

SemEval⁷.

SemEval is a series of international natural language processing (NLP) research workshops whose mission is to advance the current state of the art in semantic analysis and to help create high-quality annotated datasets in a range of increasingly challenging problems in natural language semantics. Each year's workshop features a collection of shared tasks in which computational semantic analysis systems designed by different teams are presented and compared.

Year 2021 - SemEval - 2021 - The 15th International Workshop on Semantic Evaluation. SemEval-2021 is co-located (virtually) with The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021) on August 5–6, 2021 in Bangkok, Thailand (online).

⁵IberLEF - Iberian Languages Evaluation Forum. <https://sites.google.com/view/iberlef2020/home>

⁶IberLEF 2020: Iberian Languages Evaluation Forum. Available in <https://sites.google.com/view/iberlef2020/home>

⁷SemEval 2021 - The 15th International Workshop on Semantic Evaluation. Available in <https://semeval.github.io/SemEval2021/>

We participate in the International workshop Task 1: Lexical Complexity Prediction⁸ (Zambrano and Montejo-Ráez, 2021). Our proposal consisted of executing the data set to which we added 15 linguistic features so that it was then executed in the traditional Machine Learning algorithm called Random Forest.

CLEF⁹ - Conference and Labs of the Evaluation Forum.

Year 2023 - CLEF 2023 is the 14th CLEF conference. Held in Thessaloniki - Greece. CLEF conferences contribute to the systematic evaluation of information access systems, mainly through experimentation in shared tasks.

We participated in the International workshop SimpleText@CLEF-2023¹⁰ - CLEF 2023 SimpleText Track, in task 2: Complexity Spotting: Identifying and explaining difficult concepts for general audience (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023b). Our participation played a key role in this conference, as we took first place in Task 2.1 by achieving the highest number of correctly detected terms and the highest scores among all participants.

Year 2024 - CLEF 2024 is the 15th CLEF conference continuing the popular CLEF campaigns which have run since 2000 contributing to the systematic evaluation of information access systems, primarily through experimentation on shared tasks.

We participated in the International workshop SimpleText@CLEF-2024¹¹ - CLEF 2024 SimpleText Track, in task 2: Identifying and explaining difficult concepts (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024).

Year 2024 - LREC-COLING 2024 - Language Resources and Evaluation. LREC is the International Language Resources and Evaluation Conference. This conference is a key event in the field of computational linguistics, organised by two major international players, ELRA (European Language Resources Association), an organisation dedicated to promoting the creation and use of language resources in Europe, and ICCL (International Committee for Computational

⁸SemEval-2021: Program - <https://semeval.github.io/SemEval2021/schedule>

⁹CLEF 2023 - Conference and Labs of the Evaluation Forum.

Available in <https://clef2023.clef-initiative.eu/>

¹⁰SimpleText@CLEF-2023. CLEF 2023 SimpleText Track.

Available in <http://simpletext-project.com/2023/clef/tasks>

¹¹SimpleText@CLEF-2024. SimpleText is a track organised as a part of CLEF 2024 conference, initiated by CLEF initiative.

Available in <https://simpletext-project.com/2024/en/>

Linguistics), an organisation focused on research and development in computational linguistics.

We participated in the International workshop DeTermIt!¹² - Co-Located with the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation LREC-COLING 2024 (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024b).

Below, The Figure 1 presents the details of our findings and contributions:

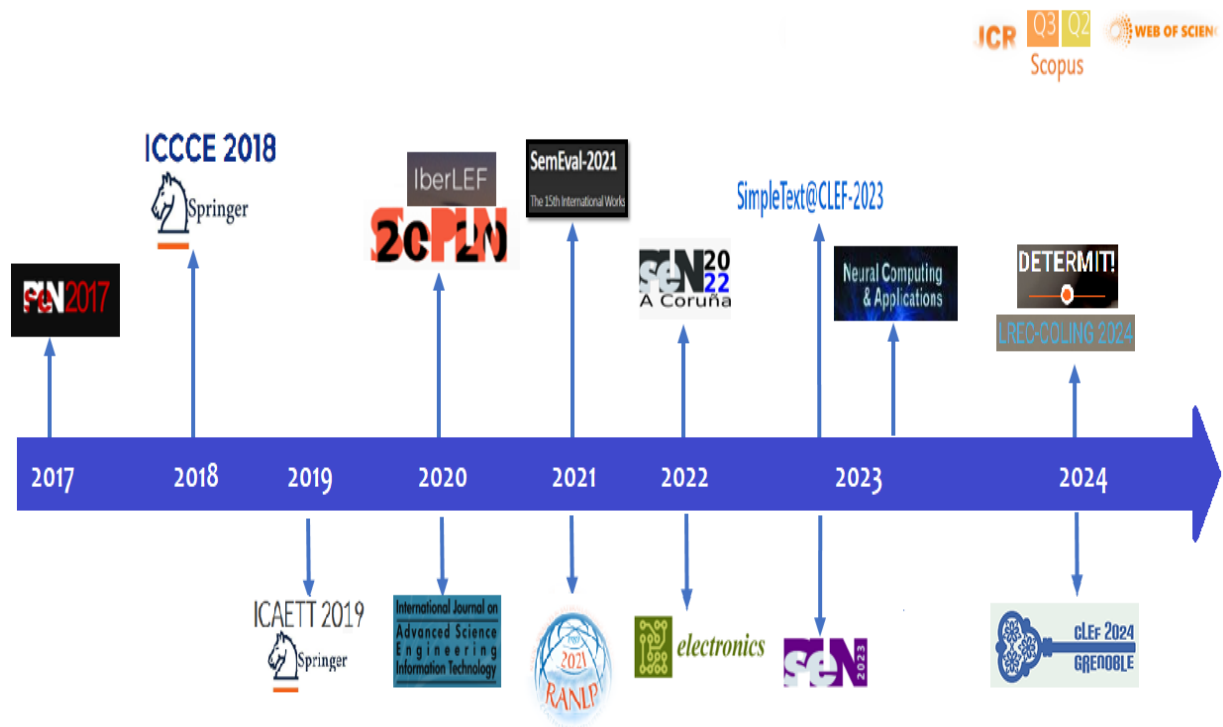


FIGURE 1: Publications made throughout the development of the research.

Applications Prototype

Our first demonstration was VYTEDU (Videos and Transcriptions in Education), an innovative corpus of videos and their transcripts developed at the State University of Guayaquil for the study of text simplification in education. 55 videos were produced with their respective transcripts, which were made available to the scientific community as a research tool. This demonstration sought to promote the dissemination and early use of this material, maximizing its potential in educational and technological research. VYTEDU is available at the following

¹²DeTermIt!. DeTermIt! Workshop Evaluating Text Difficulty in a Multilingual Context. Available in <https://determit2024.dei.unipd.it/>

link <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/issue/view/234>.

Because the University of Guayaquil does not have accessible platforms to support students who face challenges in understanding academic texts, a platform has been developed especially for those with special needs and students who have difficulty understanding reading. According to our studies, students are exposed daily to reading comprehension barriers. Our platform is designed to help them overcome these obstacles and enrich their learning process effectively and equitably.

This demonstration implementation marks a milestone at the University of Guayaquil, being the first of its kind. It has been developed specifically to offer support to students of Computer Systems Engineering and Software Engineering. It is called *EwLex* - Easy with Lexical Simplification. It is a platform that is available to all students, regardless of their academic program. Its operation is simple: the student only needs to identify the words he/she considers complex and then he/she can choose to know their meaning or discover synonyms for them. EwLex is available at the following link <https://ewlex.pythonanywhere.com/>. Figures 2 and 3 present the EwLex interfaces.

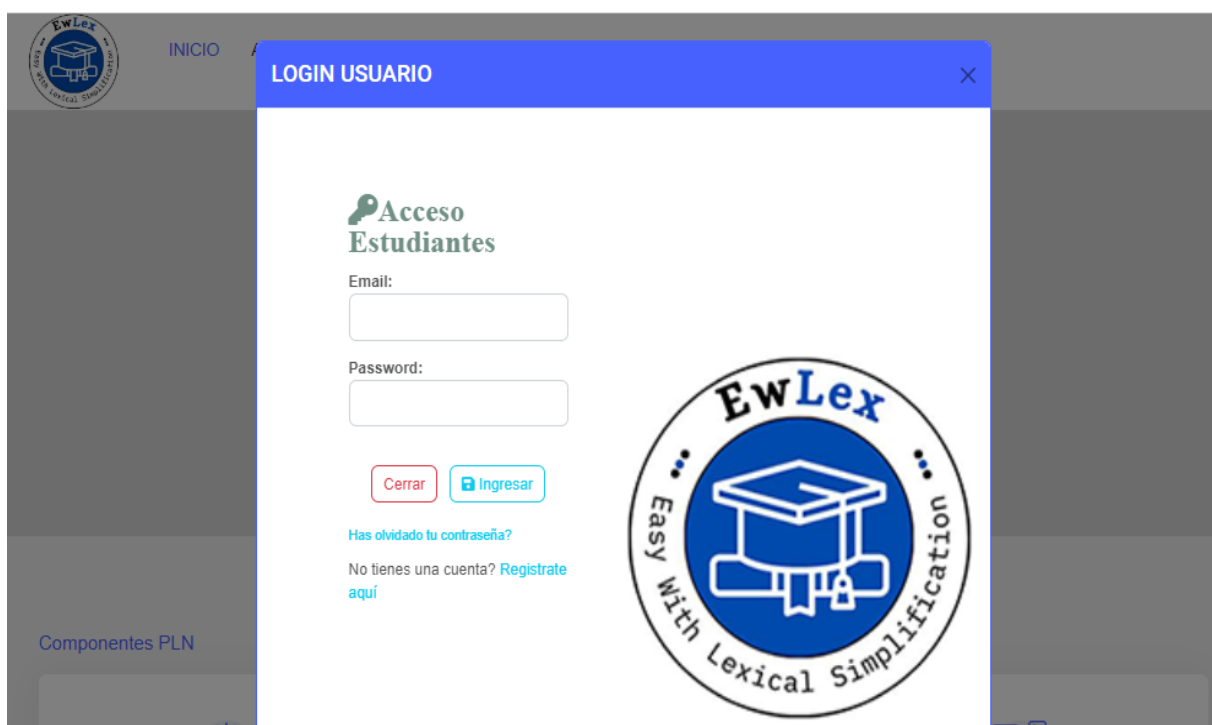


FIGURE 2: EwLex Interface.

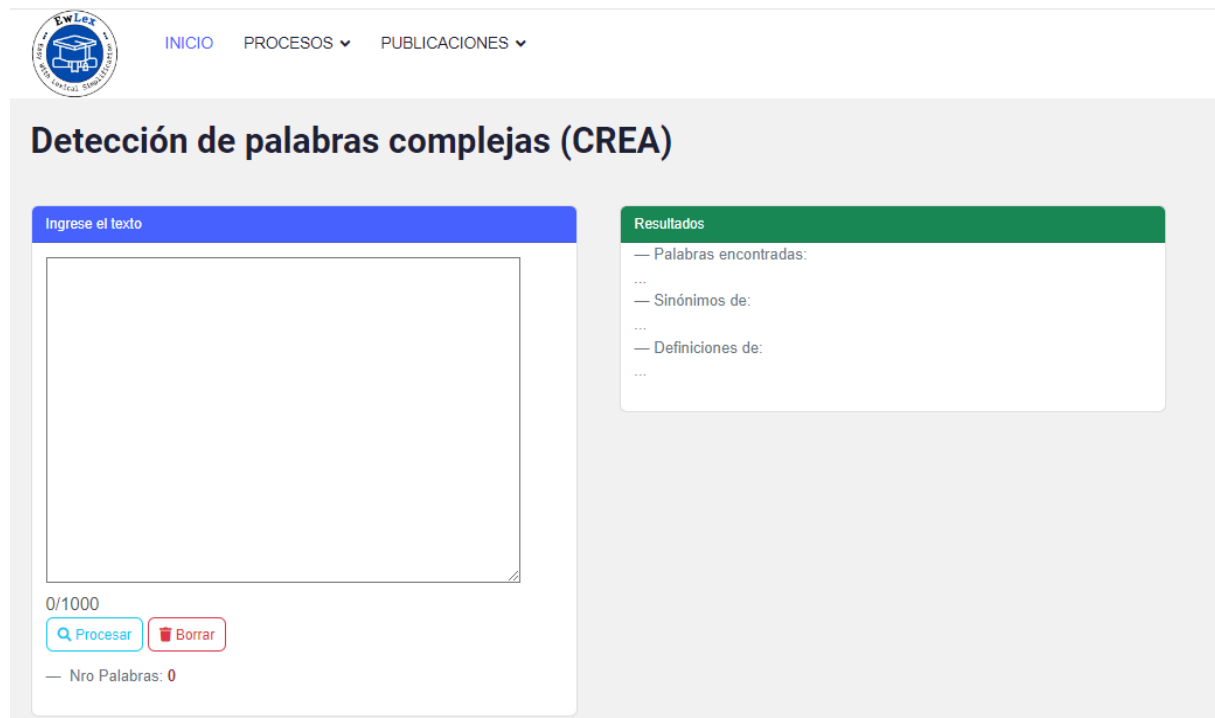


FIGURE 3: EwLex Interface.

We developed a prototype aimed at users of the busiest public institutions in Ecuador. GovAIEasy (Government Artificial Intelligence Easy) is a digital tool that facilitates text analysis through the use of artificial intelligence. This application helps users understand complex texts by providing definitions, examples and use cases of complex terms not only identified by the GPT-4o model but also complex words identified by users. GovAIEc generates a summary, that is, an accessible text, which can be useful first for understanding the content of the texts of public documents, and then for executing bureaucratic procedures. GovAIEc is a very useful tool in detecting complex terms, and suggests more accessible alternatives. GovAIEc is available at the following link <https://www.govaieasy.com/>. Figure 4 presents the GovAIEc interface.

Table 5 presents a detailed summary of the academic and scientific activities carried out during the development of the project, including publications, participation in conferences, and collaborations in evaluation forums. This summary reflects the progress and impact generated not only in the academic community, but also in the scientific community thanks to the advances in research linked to the main objective of the project. In Annex A (see Appendix A) provides a comprehensive validation of the academic and scientific activities carried out, ensuring the transparency and veracity of the achievements obtained throughout the project.

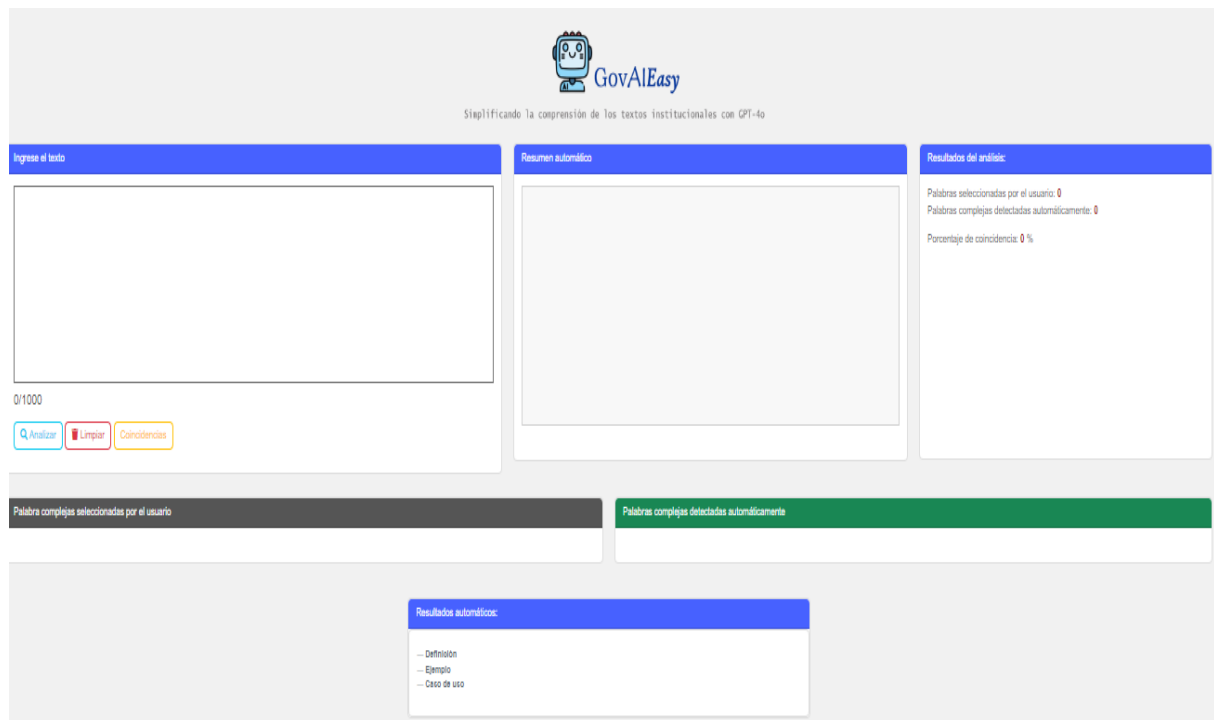


FIGURE 4: GovAIEasy Interface.

Summary of Academic and Scientific Activities Carried Out During the Development of the Project

JOURNAL PUBLICATIONS = 7

SEPLN 2017

(Ortiz-Zambrano and Montejó-Ráez, 2017)

<http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/issue/view/234>

IJASEIT 2020

(Ortiz-Zambrano and Montejó-Ráez, 2020)

<https://ijaseit.insightsociety.org/index.php/ijaseit/issue/view/65>

Electronics

(Ortiz-Zambrano, Espin-Riofrio, and Montejó-Ráez, 2022a)

<https://www.mdpi.com/search?authors=ortiz+zambrano&journal=electronics>

SEPLN 2021

(Ortiz-Zambrano and Montejó-Ráez, 2021)

<http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6438>

SEPLN 2022

(Ortiz-Zambrano, Espin-Riofrio, and Montejó-Ráez, 2022b)

<http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6438>

SEPLN 2023

Summary of Academic and Scientific Activities Carried Out During the Development of the Project (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023a) http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6557 Springer Journal 2023 (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024a) Paper under review Neural Computing and Applications
PUBLICATIONS AT CONFERENCES = 4 ICCCE 2018 (Ortiz-Zambrano and Varela Tapia, 2018) https://link.springer.com/chapter/10.1007/978-3-030-12018-4_9 ICAETT 2019 (Ortiz-Zambrano et al., 2019) https://link.springer.com/chapter/10.1007/978-3-030-32022-5_16 SemEval 2021 (Zambrano and Montejo-Ráez, 2021) https://aclanthology.org/2021.semeval-1.11/ RanLP 2021 (Ortiz-Zambrano and Montejo-Ráez, 2021) https://ranlp.org/ranlp2021/proceedings-20Sep.pdf GovAIEc: A Lexical Complexity Corpus for Spanish in Ecuadorian public documents Paper under review
PARTICIPATION IN EVALUATION FORUMS = 4 IberLEF at SEPLN 2020 (Ortiz-Zambrano and Montejo-Ráezb, 2020) https://ceur-ws.org/Vol-2664/ SimpleText@CLEF-2023 (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023b) https://ceur-ws.org/Vol-3497/ Determin- LREC-COLING-2024 (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024b) https://aclanthology.org/volumes/2024.determin-1/ SimpleText@CLEF-2024

Summary of Academic and Scientific Activities Carried Out During the Development of the Project
(Ortiz-Zambrano, Espin-Riofrio, and Montejó-Ráez, 2024) https://ceur-ws.org/Vol-3740/
DOCTORAL RESEARCH STAYS = 2 Year 2017 Year 2023
DEMONSTRATIONS / DEVELOPED PROTOTYPES = 3 VYTEDU http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/issue/view/234 EwLex https://ewlex.pythonanywhere.com/ GovAIEc https://www.govaieasy.com/
PROPOSED WORKSHOPS IN LEXICAL ANALYSIS = 1
ALexS 2020 at SEPLN 2020 (Ortiz-Zambrano and Montejó-Ráezb, 2020) https://ceur-ws.org/Vol-2664/
RELATED PROJECTS = 1
Artificial Intelligence in Public Management and its Social Effects on the Understanding of Bureaucratic Processes to Improve Efficiency. Start date: September 2023 End date: September 2025 Resolution No. R-CSU-UG-SE34-313-14-09-2023
SUPERVISED BACHELOR'S THESIS PROJECTS IN RELATED FIELDS = 16
Development of a Web system applying lexical simplification to contribute to the reading comprehension of EGB students in the city of Guayaquil. https://repositorio.ug.edu.ec/items/7ef04d80-3b8e-4350-a226-76b7825bb3ef

Summary of Academic and Scientific Activities Carried Out During the Development of the Project

Evaluating the prediction of the level of complexity of words in Law studies by applying the language model based on Transformers XLM-RoBERTa.

<https://repositorio.ug.edu.ec/items/c4b2f635-09db-41fb-952a-4208fbbcad0e>

Implementation of a website for the detection of complex words in Spanish to contribute to the linguistic understanding of university students.

<https://repositorio.ug.edu.ec/items/b0caa03c-be98-46cb-88c6-609c03db371c>

Masked language modeling and fine tuning with Bert and XLM-RoBERTa models in evaluating the prediction of complex words in the English language.

<https://repositorio.ug.edu.ec/items/7d22b22a-6b32-496c-8c2c-d84deb3d1597>

A new labeled corpus for researching complex words in public administration procedures. Analysis and evaluation using language models.

<https://repositorio.ug.edu.ec/items/263f2b40-a19f-447b-b51d-92879eda1cc5>

Complex Word Detection System in Spanish Applying Lexical Simplification Strategies to Contribute to the Linguistic Comprehension of University Students.

<https://repositorio.ug.edu.ec/items/d277a5b2-832d-4eaa-90b0-1f391194a554>

Lexical complexity prediction system to help reduce reading comprehension barriers in English.

<https://repositorio.ug.edu.ec/items/597655b6-de4a-42ad-b455-f00a241854f6>

Linguistic features and their contribution to the prediction of lexical complexity in the execution of language models based on Transformers.

<https://repositorio.ug.edu.ec/items/57220d88-80d7-4dc4-a67c-583e212534d9>

Construction of a corpus of scientific literature with GPT-3 predictions applying the Chain of Thought (CoT) Prompting technique. Analysis and evaluation.

<https://repositorio.ug.edu.ec/items/e9fe8b6e-99b9-46fd-8bf9-df1da9cbf9c6>

Creation of a corpus of university texts in Spanish for the identification of complex words in the area of lexical simplification.

<https://repositorio.ug.edu.ec/items/ab5677af-6848-4a97-98bf-8f533a978420>

Performance evaluation of transformer models for the prediction of lexical complexity for the Spanish language.

<https://repositorio.ug.edu.ec/items/6b965d1f-cba7-4cd2-82c4-f3607e67a117>

MarIA: Robust model in the investigation of the detection of complex words in the Spanish language".

<https://repositorio.ug.edu.ec/items/9ebe5c7c-1340-4a09-a256-37216c840539>

A new labeled corpus for research on complex words in the Law program at the University of Guayaquil.

<https://repositorio.ug.edu.ec/items/68b1d52e-8dd0-44b9-a81b-adb70dd80b96>

Summary of Academic and Scientific Activities Carried Out During the Development of the Project
Lexical complexity prediction system implementing machine learning and neural networks to reduce reading comprehension barriers in university students. https://repositorio.ug.edu.ec/items/5983cc7a-3d31-4bc0-bd6a-079c086fee21
Predicting text complexity in public documents using the pre-trained generative model transformer 3 (gpt-3): comparative analysis with other models based on the transformer architecture (in the process of publication).
Implementation of a website for the identification of complex words and their linguistic understanding in public documents using GPT-3 (in the process of publication).

TABLE 5: The activities illustrate the progress of research and the dissemination of knowledge throughout the project.

Structure of the doctoral thesis

To provide a more robust understanding of the content of this doctoral research report, the following structure is established:

Chapter I: The first chapter, corresponding to the Introduction, constitutes the initial section of this research. In this chapter, we have directed our attention to the following key aspects:

Contextualization of the Research Problem: An exhaustive contextualization of the research problem of complex words was carried out with emphasis on the academic field in the Spanish language. This analysis provided a solid framework for understanding the nature and importance of the challenge we addressed.

Formulation of the Research Question or Hypothesis: As a fundamental piece of this scientific research process, we have carefully formulated the research question or hypothesis. In this formulation, we have clearly and precisely addressed the major linguistic models and the linguistic features that intervene in the identification of complex words in university texts in Spanish. This approach has enabled us to precisely define the scope and objectives of our study, establishing a solid foundation for the research.

Distinguish the Relevance and Rationale of the Study: A comprehensive evaluation has been carried out to highlight the significance and explain in detail why the study is essential, necessary,

and valuable. In this analysis, we have placed special emphasis on identifying gaps in current knowledge that research on complex word identification aims to address. Furthermore, the unique contribution that is anticipated to be made is highlighted, underlining how the expected results will offer practical applications and concrete benefits in the academic field.

Establishment of General and Specific Objectives: The general objective is a general goal that we seek to achieve with the investigation of the identification of complex words; and the specific objectives that have provided us with guides on the tasks to be carried out to achieve the proposed goal; which have been outlined clearly and precisely, defining broad and detailed goals that we seek to achieve during the development of the study.

Chapter II: This chapter establishes the fundamental foundations in the field of Natural Language Processing, focusing specifically on Lexical Simplification and Deep Neural Networks, with a particular focus on the LLM (Large Language Model), especially Models based on the Transformers architecture. Additionally, the current state of research is addressed, highlighting the most significant advances in addressing lexical complexity.

The theoretical, conceptual, and legal landscape that supports these areas is thoroughly explored, providing a comprehensive understanding of the context in which the research takes place. Special attention is paid to the Transformers architecture within the framework of Natural Language Processing, highlighting its importance in the evolution of linguistic models.

Within this context, a detailed *analysis of the state of the art* is presented, highlighting outstanding works that have effectively addressed the challenge of lexical complexity. This approach provides a critical view of the most relevant contributions and establishes a reference point to situate the present research in the broader context of the discipline. This structure not only underlines the rigor of the methodological approach adopted but also highlights how each component of the research process has contributed to the achievement of reliable and valid results.

Chapter III: *The methodological framework* is presented as a detailed guide that outlines the execution of the study, ensuring the quality and reliability of the data collected. This approach not only facilitates the understanding of the present work but also provides other researchers with the necessary tools to evaluate the validity of the results. Relevant aspects of this chapter include:

Type and approach of research: We detail the type of research carried out and the methodological approach adopted for the development of research in the identification of complex words, which provides a clear context for understanding the study design.

Sources of Information: We identify and explain the sources of information used, the resources from which we have obtained the data, and the relevant evidence or materials to carry out our study, which has provided us with the basis for obtaining relevant and reliable data.

Data Collection Techniques: We describe the collection techniques used, highlighting the specific methods applied to collect relevant and necessary information for our study, especially in the research developed where features of different natures executed in the learning models were combined. deep. We demonstrate how the essential information that has been fundamental for the subsequent analysis was obtained.

Features

Feature sets were used to capture the morphological aspects of complex words in the texts, while semantic aspects were coded based on the number of synonyms and hyponyms of each word, obtained through WordNet. The features computed from the data, including both morphological and semantic properties, are explained in more detail in Section 3.4.2, where their contribution to the assessment and prediction of linguistic complexity is described.

Evaluation Metrics

As mentioned by North, Zampieri, and Shardlow, 2023a, the performance of complexity prediction systems is evaluated using a variety of metrics, the choice of which depends on the nature of the task at hand. The most common tasks include: (1) binary classification, applied in Complex Word Identification (CWI) systems (Paetzold and Specia, 2016a), (Yimam et al., 2018a), and (2) regression, used in Lexical Complexity Prediction (LCP) systems (Shardlow et al., 2021a). The metrics we have used to perform the evaluations are described in Section 2.5.2.

Available resources

We specify the resources used, highlighting their importance to support the research and guarantee the success of the project. We have included a combination of sources, tools, and

materials necessary to carry out the project effectively. We have managed and used these resources efficiently to guarantee the success of the research and the validity of the results obtained.

Chapter IV: The results of the research are presented, starting with the resources created as a fundamental foundation for the development of experiments through the implementation of deep neural network language models. Specifically, those based on the Transformers architecture recognized and widely used in the field of Artificial Intelligence are addressed, along with the application of Fine-Tuning and Masked Language Model techniques.

Chapter V: Details the main conclusions, providing a clear perspective on the relevance and applicability of the research in the identification of complex words in the academic field in Spanish. A comprehensive view has been consolidated that coherently summarizes the key findings derived from a thorough analysis of the data collected during the study.

The discussion unfolds through a detailed analysis of the results obtained during this study. This analysis involves a thorough interpretation that clarifies the meaning of the results and their relationship to the hypotheses. At the same time, we have conducted an exhaustive examination of the results, which has allowed us to verify the achievement of the specific objectives set, demonstrating the success of our methodology and the strategies implemented throughout the project.

Another key aspect in the development of this chapter has been the incorporation of unique contributions, which are not limited to the results generated by the project. Additionally, a comparative analysis with the conclusions of other relevant research on complex words and lexical complexity prediction has been conducted, enriching the discussion and providing a broader and more grounded perspective.

As a complement, we have proposed areas that could benefit from additional research, based on the results obtained. This approach not only highlights the relevance of the current findings but also points out promising directions for future research on the topic of complex words in Spanish academia, thus contributing to the continued advancement of knowledge in the field of Lexical Simplification.

These conclusions are not only the logical conclusion of the research but also serve as a

platform for future research, pointing out areas of opportunity and highlighting the unique contribution of this study to the academic field in Spanish.

Chapter VII: We present, in Spanish, the most relevant points of our research.

Chapter I

The Lexical Complexity Problem

1.1 Introduction

Readability is defined as the quality that makes a text more accessible and easier to read (DuBay, 2004). One of the main difficulties in understanding a text is the lack of familiarity with the context and the terminology used (Di Nunzio et al., 2024). For many people, the way a text is written can be a barrier to understanding its content (Saggion et al., 2015a) due to the presence of infrequent or unknown words, and phrases with lexical and semantic complexity which drastically complicate the reader's understanding (North, Zampieri, and Shardlow, 2023b); this is especially evident in cases such as children, non-native speakers (Segura-Bedmar and Martinez, 2017) and people with various cognitive abilities or reading disabilities (Saggion et al., 2015b). The success or failure of understanding a text will depend on the reader's prior knowledge of the meaning of the words (Anula, 2008).

Zipf's Law implies that words that appear less frequently in a text tend to be longer, and are therefore likely to be considered more complex than words that are more frequent and shorter (Quijada and Medero, 2016). Additionally, words that have a high number of syllables tend to be difficult to pronounce, which can complicate the reading experience (Mukherjee et al., 2016). These factors provide solid, yet simple metrics for understanding the level of complexity associated with a word (Desai et al., 2021a).

In Figure 1.1, you can see an example of a text that is addressed to users of a public institution

in the city of Guayaquil (Ecuador) who should be able to understand its content. For example, terms such as “susesiones indivisas”, “constituidas”, “ejercicio fiscal” require a greater cognitive effort on the part of the reader, especially if they are not familiar with the specialized vocabulary, since the use of infrequent words can represent a significant barrier to understanding the text and, in this case, represent a difficulty in carrying out the bureaucratic procedure to be carried out. Source <https://www.sri.gob.ec/impuesto-renta#%C2%BFqu%C3%A9-es?>

¿Quién debe pagar?

La declaración de Impuesto a la Renta es obligatoria para todas las personas naturales, las sucesiones indivisas y las sociedades, nacionales o extranjeras, domiciliadas o no en el país, conforme los resultados de su actividad económica; aún cuando la totalidad de sus rentas estén constituidas por ingresos exentos, a excepción de:

1. Los contribuyentes domiciliados en el exterior, que no tengan representante en el país y que exclusivamente tengan ingresos sujetos a retención en la fuente.
2. Las personas naturales cuyos ingresos brutos durante el ejercicio fiscal no excedieren de la fracción básica no gravada.

La normativa tributaria define a estos contribuyentes como "sujetos pasivos".



FIGURE 1.1: Example of text with infrequent words addressed to general users.

Note: The original text is written in Spanish, so we have included the respective translation into the working language of this thesis.

¿Who must pay?

The Income Tax declaration is mandatory for all natural persons, undivided estates and companies, national or foreign, domiciled or not in the country, according to the results of their economic activity; even when all their income is made up of exempt income, except for:

1. *Taxpayers domiciled abroad, who do not have a representative in the country and who only have income subject to withholding at source.*
2. *Natural persons whose gross income during the fiscal year does not exceed the basic non-taxed fraction. The tax regulations define these taxpayers as “taxable persons.”*

Figure 1.2 shows an example of a text issued by a public institution in the city of Guayaquil (Ecuador), addressed to grandparents of children whose parents have passed away. The text includes uncommon and infrequently used terms, making it difficult for this target group to understand, especially considering their level of familiarity with formal or technical language. The complexity of the vocabulary can not only cause confusion, but can also prevent recipients from clearly understanding critical information, making it difficult to correctly interpret and access the resources or benefits mentioned in the message. Source <https://www.corporacionregistrocivil.gob.ec/pages/pfrecuentes.aspx>



6. Los padres fallecieron, ¿los abuelos pueden inscribirlos?

De conformidad al Art. 35 de la Ley Orgánica de Gestión de la Identidad y Datos Civiles, si la progenitora fallece al momento del parto se colocará la filiación materna con base al Informe Estadístico de Nacido Vivo.

“Art. 35.- Prueba de filiación. La filiación se probará con la comparecencia del padre o la madre o ambos. En caso de no tener vínculo matrimonial o unión de hecho registrada, la filiación se probará con la comparecencia de ambos. En caso de fallecimiento de la madre en el momento del parto, la filiación materna en la inscripción de nacimiento de su hija o hijo se probará mediante la presentación del Certificado Estadístico de Nacido Vivo y la historia clínica o su epicrisis debidamente legalizada”.

FIGURE 1.2: Example of a text posted by a public institution with rare words addressed to grandparents of children of deceased parents.

Note: The original text is written in Spanish, so we have included the respective translation into the working language of this thesis.

6. Parents have died, can grandparents register them?

According to Art. 35 of the Organic Law on Identity and Civil Data Management, if the mother dies at the time of birth, maternal filiation will be recorded based on the Statistical Report of Live Births.

“Art. 35.- Proof of filiation. The filiation will be proven with the appearance of the father or the mother or both. In the event of not having a marital bond or registered common-law union, the filiation will be proven with the appearance of both. In the event of the death of the mother

at the time of birth, the maternal filiation in the birth registration of her daughter or son will be proven by presenting the Statistical Certificate of Live Births and the medical history or its duly legalized epicrisis.”

Predicting which words present comprehension difficulties for a specific target population is an important step in many Natural Language Processing (NLP) applications. This is particularly essential in text simplification, an area that has traditionally directed its focus towards second language learners, native speakers with low levels of literacy, and people with linguistic difficulties in reading (Saggion et al., 2015b). The task of determining which words are more difficult for the reader is also known as Complex Word Identification (CWI).

CWI (Complex Word Identification) and LCP (Lexical Complexity Prediction) can be considered two complementary aspects of the same perspective. While CWI focuses on the identification of complex words through a classification task, LCP addresses the prediction of the degree of lexical complexity of words through a regression task. In essence, CWI allows us to identify words that could be considered complex, and LCP provides us with a quantitative measure of their complexity. Precise lexical complexity prediction can allow texts to be adapted to the needs of readers (Shardlow, Cooper, and Zampieri, 2020b).

Rico-Sulayes, 2020 defines CWI as the task of detecting in the content of documents the words that are difficult or complex to understand by the people of a certain group. The application of CWI and the subsequent substitution of those words identified as difficult can generate significant improvements in the readability and comprehension of a specific text (Zotova et al., 2020).

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 (Fernández et al., 2009), NAACL-HTL 2018 (Yimam et al., 2018b), the ALexS task at IberLEF 2020 (Ortiz-Zambrano and Montejo-Ráezb, 2020), the 15th edition of SemEval and the first Lexical Complexity Prediction task (Shardlow et al., 2021b), TSAR-2022 workshop on Text Simplification, Accessibility, and Readability (Saggion et al., 2023), CLEF 2023 SimpleText Track (Ermakova et al., 2023), LREC-COLING 2024 - DeTermIt! Workshop in Evaluating Text

Difficulty in a Multilingual Context (Nunzio et al., 2024), among other initiatives in this sense.

In summary, the study of LCP is of great interest for the reasons indicated. There is a great deal of work being done and evaluation campaigns, which require, above all, reference corpora. In the case of Spanish, the existing reference corpora are very scarce or nonexistent, and this hinders progress. In addition, the scarcity of these resources can negatively affect the quality of linguistic analyses, emphasizing the urgent need for initiatives that promote the creation of Spanish language corpora to significantly boost research in this field.

To improve the performance of any machine learning model, it is essential to focus on the data rather than continuously developing new algorithms, and this is precisely the goal of feature engineering. Focusing on the intelligent manipulation of data, this technique takes advantage of the intrinsic bias of learning algorithms to improve both the accuracy and interpretability of models. Feature engineering, commonly applied with simple techniques such as regression or decision trees, can also optimize more complex models such as neural networks. Its approach is based on fine-tuning existing features or extracting new meaningful features from various data sources such as transactions, networks, or time series. Today, feature engineering coexists with advanced solutions such as Deep Learning, where representations encoded in Transformer coding architecture embeddings are employed, and large generative language models. The combination of these methodologies offers a wide spectrum of possibilities to more effectively address the complexity associated with complex word identification (CWI), creating a synergy between traditional and innovative techniques and opening a promising horizon for future research (Verdonck et al., 2024).

This thesis provides an advance in the study of the LCP for Spanish, creating new data sets and exploring new solutions for the prediction of lexical complexity, with contributions that have represented a new state of the art in some cases. In addition, an exhaustive experimental validation has been carried out, highlighting the robustness and generalization of the proposals presented. These results not only strengthen recent research on lexical complexity in Spanish but also open the way to new perspectives and challenges for future research, thus marking a significant milestone in the evolution of this developing discipline.

1.2 Research Hypothesis

According to Tamayo et al., 2001:

A hypothesis is a proposition that can be tested to determine its validity. It always leads to empirical testing. It is a question formulated in such a way that an answer of some kind can be anticipated.

The use of artificial neural networks, in particular deep learning architectures, such as linguistic models based on the Transformers architecture, has proven to be highly effective in the execution of Natural Language Processing tasks, specifically in regression tasks where features of different natures are combined (Shardlow et al., 2021a). Based on this premise, we propose to apply these approaches in the analysis of academic texts in Spanish, with the aim of opening new lines of research for the automatic identification and evaluation of complex words in said language within the academic field, which will allow us to improve the understanding of the linguistic challenges that users face when interacting with specialized texts, thus facilitating the development of tools that optimize the accessibility and understanding of academic content in Spanish. Hereby, we define our hypotheses:

H1. *The implementation of large language models complemented with linguistic features, emerges as an beneficial solution to identify and predict words of high complexity in academic texts in Spanish.*

H2. *Generative models, like auto-regressive language models, can accomplish the task of CWI.*

Since our hypotheses focuses mainly on the study of lexical and grammatical complexity in the Spanish language, the creation of new resources, which are a direct product of our research, became essential. These resources are exhaustively explored and detailed in Chapter IV, which represent valuable contributions that enrich the knowledge in the research of complex words.

1.3 Justification For The Research

One of the main barriers in text understanding is unfamiliar context and terminology. Even in developed countries, up to 30% of the population can only comprehend texts written with a basic vocabulary (Di Nunzio et al., 2024). The presence of unfamiliar words in a text can have

a significant impact on the reader's ability to understand it. In some cases, the reader may misunderstand the text, overlook it, or even feel discouraged. A dedicated reader might choose to spend time looking up the definition of the unfamiliar word, thus enriching their vocabulary through understanding and learning the word. However, this approach leads to a disruption in the reading process, which in turn can affect the reader's concentration (Shardlow et al., 2021a). Therefore, complex words can impose serious difficulties, considering that their meaning is often dependent on their context and cannot be easily inferred (Zaharia, Cercel, and Dascalu, 2021). Providing students with texts that are accessible and well matched to their abilities helps to ensure that students better understand the text and, over time, can help readers improve their reading skills (Crossley et al., 2023).

Complexity within LCP is used as a "synonym for difficulty" (Malmasi and Zampieri, 2016). Many factors affect the prediction of lexical complexity, one of which is the context in which a word is found, which will affect its understandability. The meaning of a word can sometimes be inferred from the context of known words around it; in contrast, a word that is in the context of other unknown words is more difficult to understand. Similarly, the genre to which a text belongs can affect the reader's understanding (Shardlow, Cooper, and Zampieri, 2020b). The word's complexity changes slightly based on the complexity of the sentence (Yaseen et al., 2021). Lexical simplification strives to enhance text comprehensibility for a broad audience by substituting intricate vocabulary and phrases with simpler alternatives while retaining the initial intended significance (Di Nunzio et al., 2024).

According to the research carried out, there are currently no studies conducted to investigate the phenomenon of lexical complexity in the academic domain of the Spanish language. 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), which constitutes an important gap, and a significant research opportunity to develop new tools and approaches to support students in the comprehension of academic texts in Spanish.

Deep learning and its innovative technologies represent edge new technologies for various natural language processing (NLP) tasks (Singh and Mahmood, 2021). Here, vocabulary complexity prediction (LCP) is no exception (Nandy et al., 2021). 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 (Bender et al., 2021).

The lack of research on the identification of complex words in the university academic context in Spanish represents an important opportunity to address this need through the application of NLP techniques. The project constitutes a significant contribution to the field of higher education, by focusing on the solution of a specific problem that could significantly impact the quality of education and equality of opportunity for all students. The contribution of this project also has an impact on the social sphere since facilitating the identification of complex words can improve the understanding of academic texts, which benefits both students and academics by addressing possible linguistic barriers and improving access to information. Its contribution is not limited only to the specific student population, but can also be integrated into other higher education institutions as a valuable resource to support the learning of university students.

Addressing the identification of complex words in the academic field in Spanish can significantly enhance educational equity, information accessibility, and the development of resources that benefit students in Spanish-speaking contexts and the broader scientific community. Identifying complex words is fundamental for designing specific interventions and resources aimed at students with linguistic difficulties, disabilities, low levels of reading comprehension, or limited vocabulary construction, providing them with the necessary support to overcome academic challenges. The research results could contribute to the creation of specialized educational resources and pedagogical strategies that improve teaching and learning in academic environments. Furthermore, enhancing the identification of complex words can positively impact the clarity and accessibility of scientific communication, facilitating the effective transmission of knowledge and discoveries within the academic community.

1.4 Scientific Motivation

The scientific motivation that promotes this doctoral thesis originates from the prevailing need to deepen research on lexical complexity, specifically in the academic field of the Spanish language. This boost will be achieved through the exploration and application of deep learning algorithms, as well as the execution of various natural language processing (NLP) tasks focused

on the analysis of complex words. This choice is justified by the inherent potential of large language models based on the Transformer architecture, which can be efficiently trained using extensive data sets. This strategy will allow not only deeper research in the area of lexical complexity but also the identification of specific linguistic patterns in the academic context, thus boosting knowledge and practical application in this field of research. The results of this research can help the implementation of measures and practices that make information easily understandable for the university population. Improving learning during the development of their academic studies by understanding accessible academic texts not only benefits those with disabilities, but also for those who have varying levels of language ability, or low levels of literacy; but also contributes to creating more inclusive and equitable educational environments for all.

As a use case, the application of various NLP techniques gives us the ability to anticipate those complex words that often become obstacles to the reading comprehension of undergraduate students at the University of Guayaquil. The early identification of complex words, specifically in the texts used by higher education students during the learning process can have a significant impact on the improvement of reading comprehension and, therefore, on the academic performance of university students. By conducting this study, we are not only exploring innovative ways to support students, but we are also contributing to the advancement of research in the field of education and NLP in a specific academic context, which is the University of Guayaquil.

1.5 Research Objectives

1.5.1 General Objective

Applying advanced natural language processing techniques to identify and reduce the complexity of words in academic texts in Spanish, in order to improve the accessibility and understanding of said texts, and evaluate the effectiveness of the algorithms used through an exhaustive analysis of the results obtained.

1.5.2 Specific Objectives

1. To develop new benchmark corpora as essential resources for the development of research on linguistic complexity in academic texts in Spanish.

2. To explore advanced strategies for Complex Word Identification (CWI) and Lexical Complexity Prediction (LCP) in data sets, such as the application of artificial neural networks, in particular deep learning architectures such as Transformers-based language models, various Natural Language Processing (NLP) techniques, classic machine learning algorithms, and linguistic features calculated at the word level. Subsequently, evaluate the performance of the models by analyzing the generated results.
3. To integrate linguistic features into classic machine learning algorithms and deep neural network models for performance evaluation on different data sets in both Spanish and English.
4. To explore and apply unsupervised techniques, like auto-regressive models, for the analysis of Lexical Complexity Prediction (LCP) to examine the capabilities of generative pre-trained large language models to solve this task.

Chapter II

Theoretical framework

The presentation of this theoretical framework has an essential role in this research, as it establishes a solid foundation for a deeper understanding of the key concepts related to the topic of study. This encompasses precise definitions of relevant concepts and theories, bringing coherence and consistency to our research approach. Through this theoretical framework, our objective is to provide the essential theoretical basis that will allow a more complete understanding, an effective approach, and a contribution to scientific knowledge in the field of Lexical Simplification.

We aim to contextualize this study, exploring an underlying research problem and supporting its relevance through theoretical foundations. This will help both researchers and readers understand the importance and relevance of the challenge related to complex words in the context of the academic field in Spanish. The following sections provide definitions and introductions for the main concepts and areas involved in our research.

2.1 Natural Language Processing

Alsalman, [2018](#) states that *Natural Language Processing* (NLP) is also a branch of artificial intelligence that is responsible for making computers understand and process natural human languages, and NLP is frequently used in sentiment analysis as a preprocessing step. NLP is often used for processing textual data. NLP is used for lexical analysis tasks, such as Part of

Speech (POS) tagging, Named Entity Recognition (NER), and Topic Modelling (TM).

Alarcón, Moreno, and Martínez, 2021 expressed that *Natural Language Processing* is a discipline dedicated to developing technology capable of understanding natural language in a way similar to human beings. One area in which this could be applied is the development of technology that improves accessibility for individuals with disabilities.

According to Kublik and Saboo, 2022, in the real world, NLP has become central to many AI applications, such as:

Spam detection. Applying NLP, it is evaluated which emails appear to be suspicious. The process of filtering spam in the email inbox consists of assigning to the spam folder, a percentage of incoming emails.

Machine translation. Machine translation Google Translate, DeepL, and other machine translation programs use NLP to evaluate millions of sentences translated by human speakers from different language pairs.

Virtual assistants and chatbots. In this category are considered all voice assistants or digital assistants (virtual assistants) such as Alexas, Google, Siris, and the world's customer service chatbots. These applications use NLP to understand, analyze, and prioritize user questions and requests, helping them in various tasks by responding promptly and correctly.

Social media sentiment analysis. Posts about specific brands are collected on social media by marketers as conversation topics and keywords; After applying NLP, the state of satisfaction of the users is analyzed about how they feel about each topic individually and collectively. This process contributes to brands since it provides information about customer research, image evaluation, and the detection of social dynamics.

Text summarization. Summarizing a text involves reducing its size while keeping key information and the essential meaning. Some everyday examples of text summarization are news headlines, movie previews, newsletter production, financial research, legal contract analysis, and email summaries, as well as applications delivering news feeds, reports, and emails.

Semantic search. Searching data intelligently is successful thanks to semantic search that takes advantage of deep neural networks, for example, every time you search for something on Google. Indeed, semantic search is useful when searching for something based on context rather than using specific keywords.

2.2 Textual Simplification

Text simplification (TS) can be regarded as the process of converting input text into a more accessible form (Štajner et al., 2012). It is a branch of the NLP that continues to grow. It has been the subject of study for more than twenty years (Chandrasekar, Doran, and Bangalore, 1996) and is being applied to improve content accessibility (Saggion et al., 2011), although currently there are few tools for simplifying texts in Spanish (Bott et al., 2012). It has been shown that the techniques used in text simplification provide a shorter and easier-to-understand text as a result, that is, a text reduced in size compared to the original text.

TS is useful for advanced NLP applications such as machine translation, Q&A systems, or dependency parsers (Chandrasekar, Doran, and Bangalore, 1996). In either case, it is of prime importance to keep the meaning of the original text, or at least to try not to lose information. In the evaluation process, three aspects are considered essential to measure the objectives of the simplification system: reduction of the complexity of the text, preservation of meaning during simplification, and production of grammatical results (Saggion et al., 2015a).

Automated Text Simplification (ATS)

According to Aranzabe, Ilarraza, and Gonzalez-Dios, 2012 *Automatic Text Simplification* is an NLP task which aims to simplify texts so that they are more accessible to people who learn foreign languages (Petersen and Ostendorf, 2007); or people with disabilities (Carroll et al., 1998). It is a technology used to adapt the content of a text to the specific needs of individuals or a determined group to make said texts more readable and understandable by them (Saggion et al., 2015a).

Automatic Text Simplification is a technology used to adapt the content of a text to the specific needs of particular individuals or target populations in a way that the text becomes more readable and understandable for them (Saggion et al., 2015a) reducing their syntactic and lexical

complexity (Espinosa-Zaragoza et al., 2023).

ATS aims to transform complex texts into their simpler variants which are easier to understand to wider audiences and easier to process with natural language processing (NLP) tools (Štajner and Glavaš, 2017). According to the research by Espinosa-Zaragoza et al., 2023, it is established that (Chen et al., 2017) supports the claim that “ATS is composed of lexical, syntactic and discourse simplification levels”. A pre-requisite for TS to be useful is that it should convey information that is consistent with the meaning of the original text. However, current TS evaluation protocols assess system outputs for simplicity and meaning preservation without regard for the document context in which output sentences occur and for how people understand them (Agrawal and Carpuat, 2024).

2.2.1 Syntactic Simplification

Siddharthan, 2006 defines *Syntactic simplification* as the process of reducing the grammatical complexity of a text while retaining its information content and meaning. *Syntactic simplification* involves replacing particular syntactic constructs (like relative clauses, apposition, and conjunction) in sentences to make the text easier to read for some target groups (people with aphasia, deafness, or low reading ages have trouble understanding long sentences and complex grammar) or easier to process by some program (like parsers or machine translation systems).

The presence of several structures proposed in systems has been identified in systems aimed at simplifying syntax.

The work presented by Siddharthan, 2006 based on *syntactic simplification* was developed in three fundamental stages:

1. *Analysis stage*:

Output Specification for Analysis Stage:

- (a) The text should be segmented into sentences.
- (b) Words should be part-of-speech tagged.
- (c) Elementary noun phrases should be marked up and annotated with grammatical function information.

- (d) Boundaries and attachments should be marked up for the clauses and phrases to be simplified.
- (e) Pronouns should be co-referred to their antecedents.

2. *Transformation and Regeneration Stages*

The transformation stage receives as input the output of the analysis stage. The transformation stage implements an algorithm that recursively simplifies the analyzed text.

3. *Preserving Conjunctive Cohesion*

How the issues of sentence ordering, cue-word selection, and determiner choice can be resolved in a manner that preserves conjunctive cohesion and connectedness.

(a) Sentence ordering

- i. Constraint Based Text Planning: Formulation of the sentence ordering task as a constraint satisfaction problem.
 - ii. Local vs. Global Sentence Ordering: Global sentence ordering would involve deciding the relative order of the simplified sentences.
 - iii. Local Sentence Ordering and Recursive Transformation: In local sentence ordering as a constraint satisfaction problem, the variables represent the positions of the simplified sentences in the regenerated text, and the constraints are expressed in terms of the possible orderings of the two sentences generated by a transform. These constraints arise from the discourse relation between the two sentences, as well as from considerations of referential cohesion and connectedness.
 - iv. The Algorithm for Sentence-Ordering: for deciding sentence order.
- (b) Cue-Word Selection: To preserve the relation between conjoined clauses that have been simplified into separate sentences, it is necessary to introduce new cue-words in the simplified text.
- (c) Determiner choice: Simplification of relative clauses and appositives results in doubling a noun phrase. It becomes necessary to use a reference expression the second time. You also need to decide which determiners to use. This decision depends on the relationship between the extracted clause or phrase and the noun phrase to which it is attached.
- (d) Evaluation: Evaluating issues of conjunctive cohesion is non-trivial. One way to evaluate these regeneration issues is by using human judgments.

- i. Grammaticality: The evaluation results for grammaticality and meaning preservation are according to all judges.
- ii. Cohesion: The coherence is done according to the judgment of the judges, taking into consideration a class of interruptions.
- iii. Readability: Some objective readability measures are applied, such as the Flesch measure.

Figure 2.1 presents a planned architecture for syntactic simplification as proposed by Siddharthan, 2006.

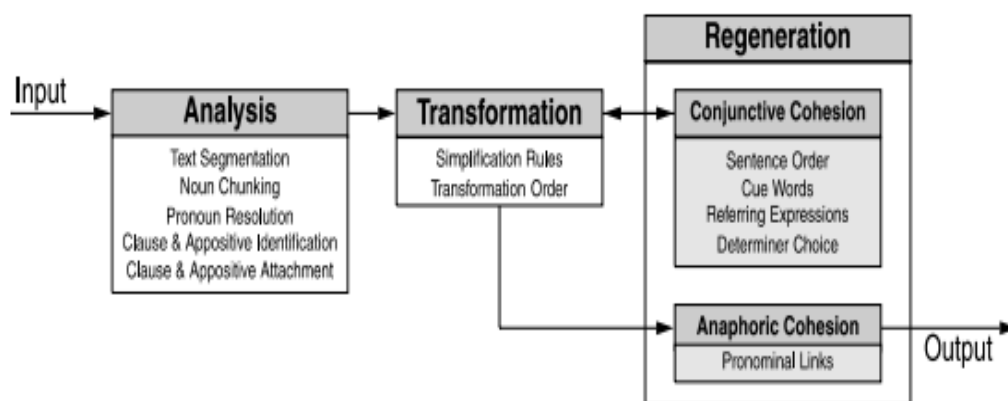


FIGURE 2.1: Artificial Intelligence domain areas by Rouhiainen, 2018.

In their work presented by Espinosa-Zaragoza et al., 2023 assert that this process includes:

- Sentence structure reordering,
- Splitting, and adjustment,
- The reduction of grammar complexity, and,
- The elision of unnecessary information.

2.2.2 Complex Words Identification

Complex Words are considered as mono-lexical units which are difficult to read (i.e. decode), especially for poor and dyslexic readers (Hmida et al., 2018).

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., 2018a).

Predicting which words are considered difficult to understand for a given target population is commonly known as *Complex Word Identification* (CWI) (Shardlow, Cooper, and Zampieri, 2020b).

To Rico-Sulayes, 2020, *Complex Word Identification* is the task of detecting words in the contents of texts that are difficult or complex for people in a certain group to understand. CWI aims to select the complex word candidates to be simplified in a given text (Alarcón, Moreno, and Martínez, 2021). CWI and the substitution of words identified as complex may significantly improve the readability and understandability of a given text (Zotova et al., 2020).

2.2.3 Lexical Simplification

Is an important subfield of *Text Simplification* (Alarcón, Moreno, and Martínez, 2021) and it involves replacing a word in a given context with a synonym that is easier to understand. It can also be solved by enriching or enhancing the text by providing a definition, image, or video, among others. Implicit in this step is the disambiguation task, which entails selecting the most prevalent meaning among the list of synonyms available (Espinosa-Zaragoza et al., 2023). The goal is to succeed not only in terms but also in confusing sentences with simple phrases; that is, a word substitution task, where the purpose is to find a synonym that is, in some sense, simpler than the original word (Saggion et al., 2011).

LS becomes a substantial task because a person to be able to understand a text at a basic level needs to know 95% of the vocabulary, therefore, this suggests that substituting words that are infrequent for a person can improve accessibility of certain texts (Alarcón, Moreno, and Martínez, 2021). *Automatic lexical simplification* can be an effective way to make text accessible to different audiences (Uluslu, 2023)

According to the conclusions of Saggion et al., 2023, research has revealed that there are various architectures proposed in lexical simplification systems. Several studies have suggested the following approach, also applied by Todirascu et al., 2022, where each process has been defined by Alarcón, Moreno, and Martínez, 2021:

1. Identification of complex terms (Complex Word Identification - CWI).

It is necessary to distinguish which words are complex and which are not for a given audience.

2. Generation of substitution words (Substitute Generation - SG).

In this step SG generates the substitution candidates for the complex words, taking into account all the contexts in which they may appear.

3. Selection of the substitutes that can fit in the context (Substitute selection - SS).

The SS stage takes the list of synonyms extracted from the previous step and selects the most suitable synonym according to its simplicity and context.

4. Taking substitutes by their simplicity (Substitute Ranking - SR).

The SR stage takes the list of synonyms extracted from the previous step and chooses which candidate that fits the context is the simplest, taking into account the target user.

5. Morphological generation and context adaptation.

2.3 Lexical Complexity Prediction

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 (Abramov, Ivanov, and Solovyev, 2023).

LCP is beneficial within the broader context of lexical simplification. It can play a role in identifying words that could be candidates for simplification, as well as in classifying words that could serve as substitutes. Additionally, the LCP has applications in the field of readability assessment, as it provides information on the proportion of complex words in a text, which contributes to the understanding of the overall complexity of the text. “Whereas CWI is particularly useful as a preprocessing step for Lexical simplification (identifying which words should be transformed), LCP may also be useful for readability assessment or as a rich feature in other downstream NLP tasks” (Shardlow et al., 2021a).

According to North, Zampieri, and Shardlow, 2023b, “*Monolingual LCP*” refers to the task of predicting the complexity values of words in a single language. *Multilingual LCP* refers to the task of creating an LCP system that can be trained on and used to predict the lexical complexities of multiple languages. *Cross-lingual LCP* refers to the task of training an LCP system on one or multiple languages and then using that system to predict the lexical complexities of a language previously unseen within the training set.

2.3.1 Lexical Complexity Metrics

A good indicator of writing quality is to use a measure of lexical complexity, referring to the size, variety, and quality of a student’s vocabulary (Crossley, Salsbury, and McNamara, 2012).

The analysis of the complexity of Spanish texts is very important in our research since it allows both the study of the texts (which are the object of our work) and the evaluation of a future system. To carry out this analysis, we have applied some of the metrics selected by Saggion et al., 2015a. These metrics consist of a set of formulas that allow evaluating the complexity of the text at various levels, including the lexical complexity index (LC), the sentence complexity index (SCI), which were proposed by Rebollo, 2008, and Spaulding’s Spanish Readability (SSR) by Spaulding, 1956. These values are calculated according to the following formulas:

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.

The formulas that correspond to the metrics are:

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

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

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

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

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

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

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

For better understanding, Table 2.1 shows the definition of the variables on which the formulas are based.

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 2.1: Definition of the basic variables used in complexity metrics

Some of these metrics are accompanied by an explicit interpretation of their possible values. For instance, for the interpretation of the results of the SSR formula, we can check Table 2.2.

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 2.2: Interpretation of the SSR metrics by Campos et al., 2020.

2.4 Artificial Intelligence and Machine Learning

In the context of his research, Chan, 2023 defined *Artificial Intelligence* (AI) based on Haenlein and Kaplan, 2019's work about as a system that accurately interprets and learns from external

data to achieve specific results through flexible adaptation. AI examines significantly more data, using neural networks with multiple hidden layers. The need for rich data to train deep learning models lies in their ability to learn directly from this data; The more data they are given, the greater their accuracy. AI achieves astonishing levels of precision through deep neural networks, achieving results previously considered unattainable. A clear example of this advancement is evident in the iterations with technologies such as Alexa, Google Search, and Google Photos, which are based on deep learning principles and improve their accuracy with each continuous use.

AI explores the possibilities of creating thinking machines based on bio-mathematical models capable of performing human-like tasks. It is used in fields such as medicine and economics for prediction and classification (Valbuena, 2021). Today, people are benefiting from notable improvements and experiencing greater efficiency in virtually all areas of life thanks to the use of artificial intelligence-based technologies. Unlike people, however, AI-based devices do not require rest and can process large amounts of information simultaneously. Furthermore, machines have a significantly lower error rate when performing tasks similar to those performed by humans (Rouhiainen, 2018).

Advances in artificial intelligence technologies have enabled remarkable progress in their ability to perform computer vision, speech recognition, and natural language processing tasks. See Figure Figure 2.2. These capabilities, previously the exclusive attributes of humans, are on the way to being acquired by machines and robots thanks to developments in artificial intelligence.

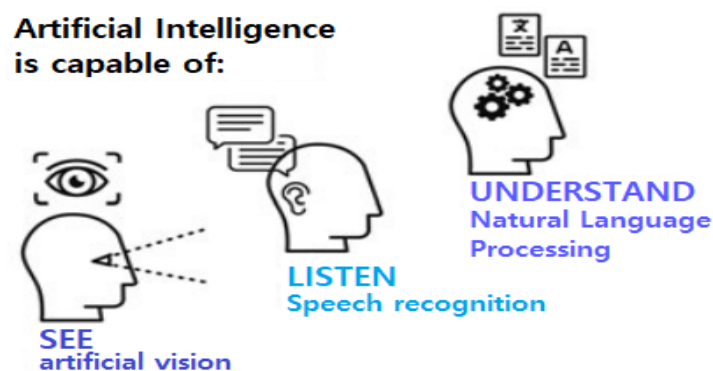


FIGURE 2.2: Artificial Intelligence domain areas by Rouhiainen, 2018.

Machine Learning (ML) is one of the main approaches to artificial intelligence (Rouhiainen,

2018). ML is a subfield of AI that deals with the study of machines capable of learning from experience and performing tasks without being explicitly programmed to do so Kublik and Saboo, 2022. According to Norman, 2019, “is the science of teaching computers to make predictions based on data”, as shown in Figure 2.3. In short, it is a field of computing in which computers or machines acquire the ability to learn without having to be explicitly programmed. A common result of this process is predictions in specific situations (Rouhiainen, 2018).

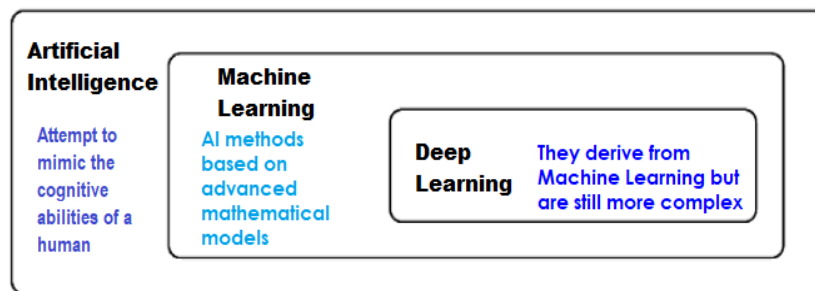


FIGURE 2.3: Scheme of the relationship between Artificial Intelligence, Machine Learning, and Deep Learning by Torra, 2011.

2.4.1 Machine Learning Taxonomy

In the vast field of machine learning (ML), there are various types of algorithms, two of the most prevalent being supervised and unsupervised learning algorithms. These different approaches are used to develop machine learning models for predictive purposes (Alsalman, 2018), and are visually detailed in Figure 2.4. According to Li, 2023, machine learning classification typically encompasses three fundamental categories: supervised learning, unsupervised learning, and reinforcement learning. On some occasions, semi-supervised learning and active learning are also incorporated.

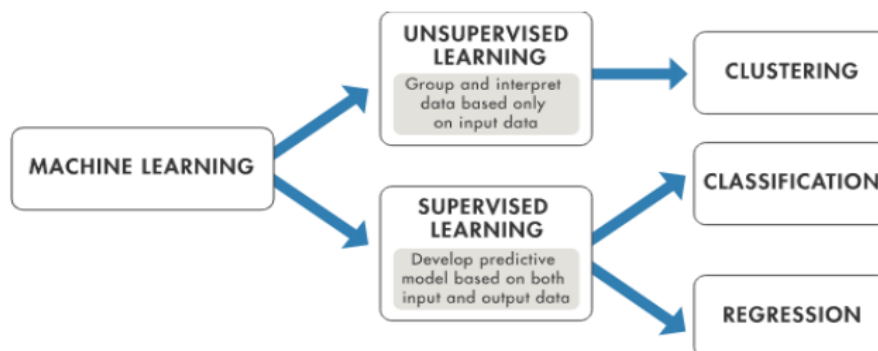


FIGURE 2.4: Machine learning types by Alsalman, 2018.

A dataset aggregates some necessary data to help the model perform better. It is fed to the ML algorithm to ensure how accurately the algorithm is interpreted (Ali et al., 2023). When you have limited, well-labeled data, the use of supervised learning is recommended. On the other hand, for large data sets, unsupervised learning generally provides more effective performances and results. If you have a sizable data set, it might be beneficial to explore deep learning techniques as an option to consider (Mahesh, 2020).

Supervised learning refers to a machine learning process in which a model learns to associate input data with corresponding output data using a set of input-output pairs provided as examples. In supervised machine learning algorithms, external guidance is required. Thus, the input data set is usually divided into two subsets: the training data set and the testing data set (Mahesh, 2020). The training data set, which has the particularity of being a labeled data set (Kochmar, 2022) contains the output variable that the model aims to predict or classify (Mahesh, 2020) with precision as many observations as possible (Dean, 2014). As Mahesh, 2020 points out, in this approach, there are clear measurements and evaluations to determine the quality of the model.

Concerning test data, this refers to a distinct portion of the data, other than the training set, where the algorithm is subjected to evaluation. A common practice is to split the data into 80% for training and 20% for testing. It is essential to emphasize that the two sets should not overlap. If an algorithm is trained and tested on the same data, determining its actual learning capabilities becomes challenging as it may end up memorizing patterns instead of genuinely learning, as Kochmar, 2022 highlights. The Figure 2.5 presents the Supervised learning workflow.

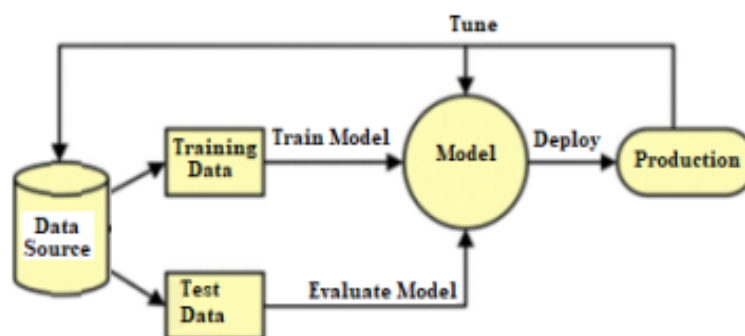


FIGURE 2.5: Supervised learning Workflow by Mahesh, 2020.

Unsupervised Learning allows you to examine existing data and predict future data (Li, 2023). It has no predefined goal and therefore lacks an objective way to measure model quality, that is,

models explore the data without labels or prior desired outputs, seeking to identify patterns or hidden structures in the data. Unlike the supervised learning mentioned above, there are no correct answers or a teacher to guide you. Algorithms alone face the challenge of discovering and revealing interesting structures in data. *Unsupervised learning* algorithms extract some features from the data. When new information is introduced, they use previously learned features to identify the category of the data (Mahesh, 2020).

The core of *Reinforcement learning* lies in acquiring the optimal sequential approach. This is a machine learning problem in which the intelligent system learns the optimal behavioral strategy through constant interaction with its environment. In case the interaction between the intelligent system and the environment follows the Markov decision process, the system can observe the sequence of data derived from its interaction with the environment. The nature of the Markov decision process in reinforcement learning implies that the state, reward, and sequence of actions are random elements (Li, 2023).

2.4.2 Other Concepts in Machine Learning

Predictive Modeling

As indicated by Dean, 2014, a *Predictive modeling* is a common activity within data mining. It involves using historical data to identify patterns using a specific methodology and then applying that model to make predictions about future events using new data. In short, it involves analyzing the past, building a model based on observed patterns, and using it to predict future events. Since modern data mining problems are often complex and large, it is advisable to follow the best practice of trying various modeling techniques and algorithms, as well as performing several attempts with different configurations and parameters on the same algorithm. This strategy is because a brute force approach is required to find the best solution. The number of iterations required to reach the best model can vary considerably depending on the domain and specific modeling challenge being addressed.

Classification Problems

Machine Learning Classification Problems are those in which one tries to predict whether a given input vector belongs to one of several classes (such as predicting the species of a plant

given several of its measures). Unlike regression problems, where there is a single neuron in the output layer, the output layer typically has one neuron for each possible class. The softmax activation function, in combination with the cross-entropy error function, is often used for output layer neurons, since the output value can be interpreted as the probability that a given input vector belongs to the class represented by each output neuron.

Regression Problems

Machine learning regression problems are those in which an attempt is made to predict a numerical value in response to a given input vector, that is, based on a set of input variables (for example, predicting a person's age based on your weight, height, and gender). Since we are looking to predict a single number, the output layer contains only one neuron. The identity activation function is commonly used for the output layer neuron, and the square of the difference between the network output and the desired value is often used as the error function. The goal in regression problems is to find a model that can accurately generalize and predict numerical values as a function of input variables, allowing decisions or forecasts to be made based on the available data.

The precision and generalizability of the model are key aspects of regression problems. The goal is to minimize the error between the model's predictions and the known actual values in the training dataset while ensuring that the model can be effectively applied to new data to make accurate predictions in real-world scenarios. In terms of machine learning algorithms, various techniques are used to address regression problems, such as Linear Regression, Polynomial Regression, Nearest Neighbor (KNN) Regression, Decision Tree Regression, and Support Vector Machine (SVM) Regression, among others. We will comment on the most used ones later on.

Figure 2.6 presents a classification of the classical regression and classification algorithms commonly used in engineering applications (Huertas Mora, 2020).

Overfitting

One of the main difficulties encountered when using historical data to develop a predictive model is the issue of *overfitting*. This arises when the model overfits a particular set of training observations, failing to generalize accurately or make accurate predictions on new data or in

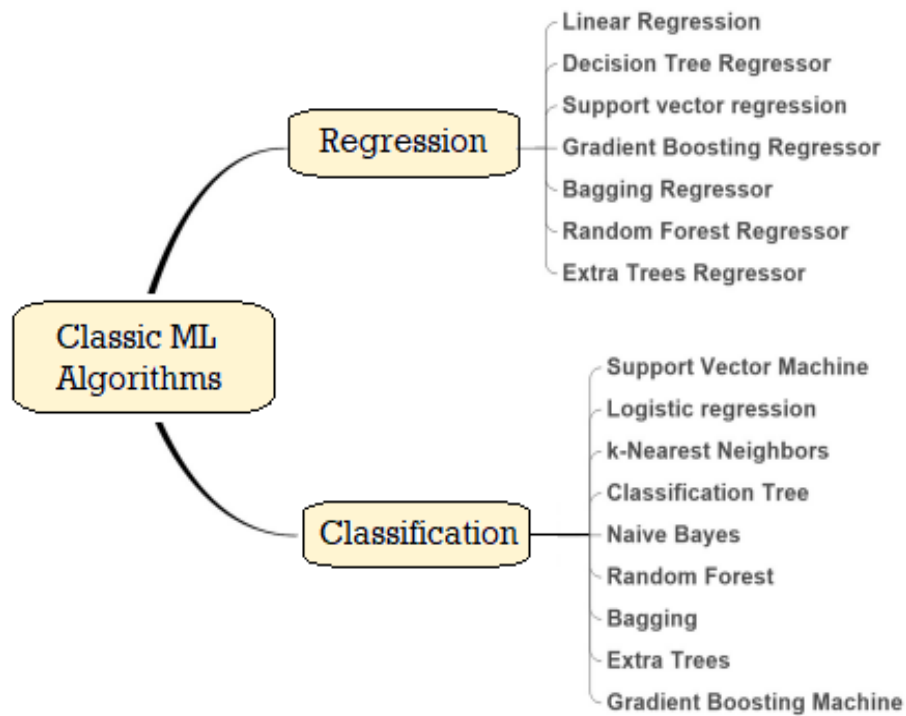


FIGURE 2.6: Classic regression and classification algorithms by Huertas Mora, 2020.

different scenarios (Dean, 2014). This phenomenon usually arises when the model has relied too much on patterns and trends present in the current data set, which might not be repeated under other circumstances (Kuhn and Johnson, 2019).

Overfitting represents a specific challenge in the realm of deep learning. Although this phenomenon can generate highly accurate classification results, they can be misleading. Consequently, if the problem of overfitting is not adequately addressed, systems that rely on accuracy in prediction or recognition could produce unreliable results (Kadhim, Abdullah, and Ghathwan, 2023).

Validation Data To Adjust For Overfitting

Overfitting occurs when a model is too complex and has too many parameters, leading to a poor generalization of new data (Ali et al., 2023). To avoid *overfitting*, it is recommended to use validation data during the model fitting process. An effective practice is to employ *cross validation* during training. This involves splitting the training data into a training subset and a validation subset. The validation subset must represent at least 25% of the total data available and it must be ensured that both subsets have a similar distribution of data. Ali et al., 2023

define K-fold cross validation is a commonly used technique to evaluate the performance of a machine learning model. It consists of dividing the data into k groups or “folds” randomly, where each group is used as a test set in an iteration, while the others are used to train the model. This process is repeated k times, with each group acting as a test set exactly once. The performance of the model is then averaged across all iterations to obtain an overall indicator of its effectiveness.

During training, only the training subset is provided to the algorithm. After each iteration of the training algorithm, the prediction error is calculated on both the training and validation subsets, using the current weights calculated by the training algorithm. This process allows you to evaluate model performance on both training and validation data, helping to identify and control for *overfitting*.

Cross-Validation

The cross-validation technique is used to deal with small data sets when evaluating models, avoiding the loss of data that would go into model testing. Generally speaking, this is done because of the persistent concern that test data are not being used effectively to estimate model parameters. However, setting aside data to understand how your machine learning model performs in production is more valuable than the minimal changes the data could bring to the model. If a 20% increase in data produces significant changes, this suggests that your model is not stable and ultimately lacks robustness. It is preferable to have a simpler model with fewer parameters that maintain constant performance with the available data (Duboue, 2020).

Ensemble Methods

An *Ensemble Model* represents a technique in machine learning that fuses various base models to create an optimal predictive model (Dietterich, 2000). It is formed by combining two or more models, using various strategies to amalgamate their results. A common and particularly favored practice is voting, in which each model casts its vote to rank an observation, and the final rank is determined democratically by selecting the option with the highest number of votes. Another strategy involves averaging the predictions from different methods, thus generating a more robust numerical prediction (Dean, 2014).

From the perspective of Sagi and Rokach, [2018](#), *ensemble methods* are considered the cutting edge in overcoming multiple challenges in machine learning. These methods improve predictive performance by training multiple models and synergistically combining their predictions.

Feature Engineering

In their investigation, Ozdemir and Susarla, [2018](#) define *features* as its most fundamental essence, a feature is defined as a data attribute relevant to machine learning. In most cases, when analyzing tabular data, we identify the columns that represent these characteristics, considering them simply as attributes. About this, Ozdemir and Susarla, [2018](#) also highlight the importance of noting that in the literature, a significant distinction is often drawn between the terms “*features*” and “*attributes*”. The term “*attribute*” typically refers to columns of data in a tabular structure, while the term “*feature*” is used specifically to describe those attributes that contribute to the successful performance of machine learning algorithms. In other words, some attributes may prove to be superfluous or even harmful to our machine-learning systems.

Ozdemir and Susarla, [2018](#) indicate that *feature engineering* is “The process of transforming data into features that better represent the underlying problem, resulting in better machine learning performance, in other words”, it is the process of creating representations of data that increase the effectiveness of a model (Kuhn and Johnson, [2019](#)). In summary, feature engineering converts raw data into features that can be used to build a prediction model with ML or statistical modeling. It aims to optimize ML model’s performance by preparing an input dataset that best fits the algorithm (Ali et al., [2023](#)).

2.4.3 Main Algorithms Used In Machine Learning

At the heart of machine learning lie algorithms, which are trained to transform into the machine learning models that drive some of the most impactful innovations in the world today. Machine learning algorithms are the essential foundations of machine learning models, from classification to regression. We present eight *traditional Machine Learning Classifiers* widely applied and also used in our research.

Traditional Machine Learning Classifiers:

1. AdaBoost — AB (Paetzold, [2021](#)).

2. Random Forest — RF Zaharia, Cercel, and Dascalu, [2021](#). (Desai et al., [2021a](#)).
3. Naïve Bayes — NB Mahesh, [2020](#).
4. Decision Tree — DT (Shardlow, Evans, and Zampieri, [2022](#)).
5. Support Vector Machines — SVM (Liebeskind, Elkayam, and Liebeskind, [2021](#)).
6. Nearest Neighbors — KNN (Liebeskind, Elkayam, and Liebeskind, [2021](#)).
7. Gradient Boosting — GB (Vettigli and Sorgente, [2021](#)).
8. Stochastic Gradient — SG (Bottou, [2010](#)).
9. Passive Aggressive — PA (Crammer et al., [2006](#)).

Below is the detail:

Random Forest — RF (Ali et al., [2023](#))

Random forest is a supervised learning method that can be utilized for data classification and prediction. Nonetheless, it is primarily employed to overcome classification issues. The random forest algorithm constructs decision trees from sample data, generates predictions from each one, and then conducts a vote to identify the optimal option.

AdaBoost — AB (Paetzold, [2021](#))

known as *Adaptive Boosting*, is an ensemble learning technique utilized for classification and regression tasks. For regression problems, it is referred to as AdaBoost Regressor. This widely used machine learning algorithm blends the predictions of several weak learners, typically shallow decision trees, to construct a robust ensemble model. The adaptive nature enables the model to enhance its performance through successive iterations.

Decision Tree — DT (Shardlow, Evans, and Zampieri, [2022](#))

Functions as a supervised machine learning technique utilized for regression purposes. Unlike classification challenges that seek to predict categorical labels, regression tasks focus on forecasting continuous numerical values. A decision tree takes the form of a tree-like model in which internal nodes indicate decisions based on specific features, while leaf nodes correspond to the predicted output values. In the case of the DTR, the value at each leaf node is derived from the average (or another measure) of the target values from the training samples that lead

to that particular lea.

Naïve Bayes — NB (Mahesh, [2020](#))

The Naïve Bayes classifier is a classification technique that is based on Bayes' theorem and operates under the assumption of independence between predictive features. In simple terms, this type of classifier assumes that the presence of a specific feature in a class is unrelated to the presence of any other feature. Naive Bayes is mainly used in the text classification industry and is mainly applied to clustering and classification tasks that depend on the conditional probability of occurrence.

Support Vector Machines — SVM (Liebeskind, Elkayam, and Liebeskind, [2021](#))

Another widely used cutting-edge machine learning technique is known as Support Vector Machines (SVM). In the field of machine learning, SVMs are supervised learning models that use specific algorithms to analyze data used in classification and regression analysis tasks. Apart from their ability to perform linear classifications, SVMs can effectively perform non-linear classifications by applying the so-called "kernel trick", which involves an implicit mapping of inputs to higher-dimensional feature spaces.

Nearest Neighbors — KNN (Liebeskind, Elkayam, and Liebeskind, [2021](#))

The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to implement and understand but has a major drawback of becoming significantly slow as the size of that data in use grows.

Gradient Boosting — GB (Vettigli and Sorgente, [2021](#))

Gradient boosting is an ambitious algorithm that can overfit a training dataset quickly. You can gain benefits from using regularization methods that penalize different aspects of the algorithm and generally improve its performance by reducing overfitting.

Stochastic Gradient — SG (Bottou, [2010](#))

An innovative idea regarding ensembles and random forests was the possibility of efficiently generating trees from subsets of the training data, which in turn decreased the correlation between trees within the collection. This concept can also be applied to mitigate the correlation between trees in the sequence of gradient-boosting models. This adapted version of the boost is

known as *stochastic gradient boost*.

Passive Aggressive — PA (Pal, Pranav, and Pradhan, 2023)

According to a study carried out by Pal, Pranav, and Pradhan, 2023, passive-aggressive algorithms, previously mentioned in research such as those of Granik and Mesyura, 2017, Tacchini et al., 2017 are online learning algorithms that share similarities with the perceptron algorithm. These algorithms maintain a passive approach when the classification is correct, but become aggressive in case of error, performing updates and adjustments. Its main objective is to make modifications that correct the loss, while minimizing alterations in the weight vector norm (Ahmad et al., 2020).

Unlike the batch learning approach, where the entire training data set is processed at once, online machine learning algorithms work with the input data sequentially and gradually update the machine learning model. This methodology is especially beneficial in situations where large volumes of data are handled, and training the entire set is computationally infeasible due to the large magnitude of the data, as mentioned in previous research (Wong, 2016).

2.4.4 Artificial Neural Networks

According to Valbuena, 2021, artificial neural networks have their roots in the latter part of the 20th century, being conceived from a fundamental mathematical representation of biological neurons, which play a important role in the learning process of living organisms. These “Artificial Neural Networks” (ANN) constitute a type of artificial intelligence system modeled after neurons, the most prominent cells in the human biological nervous system. Its design is conceived to simulate the processes through which the human brain processes information, learns, and retains knowledge.

An artificial neural network is defined as an interconnected structure of nodes that, by applying different weights and biases between layers, processes input values until obtaining an output value. This system comprises multiple layers of configured nodes, where each one, both in the hidden and output layers, incorporates its own classification mechanism. On the other hand, the nodes located in the visible layer are responsible for acquiring the input values, after which the processing process is activated (Nokeri, 2021).

Valbuena, 2021 cites the work of McCulloch and Pitts in 1943 by stating that “This model was initially conceptualized by neurophysiologist Warren McCulloch and logician Walter Pitts in 1943” McCulloch and Pitts, 1943. In their model, a neuron “fires” or generates an output of 1 when the cumulative excitatory inputs surpass a specific threshold, and there are no inhibitory inputs present. McCulloch and Pitts illustrated that networks constructed with these modeled neurons could effectively compute the values of intricate logical equations, with variables constrained to the binary values of 0 or 1 (Dean, 2014).

Rosenblatt’s Simple Perceptron Model Or Single-Layer Perceptron Simple ANN’s

In 1957, Frank Rosenblatt built a device called the *Perceptron* (Rosenblatt, 1958) which Valbuena, 2021 refers to as “a single neuron or a simple structure thereof, used to classify input data into one or more categories”. The neuron becomes the most fundamental element of the perceptrons. Rosenblatt implemented this neural network, also called *perceptron*, composed of neurons similar to those proposed by McCulloch and Pitts, 1943. This network consisted of electronic nodes capable of transmitting and receiving signals, whose operation resembled the nervous stimuli present in the neural interconnections of the brain, as described by Bronte-Stewart et al., 2009.

The *Perceptron learning algorithm* starts with a random initial set of weights and then iteratively adjusts those weights based on an error function that evaluates the difference between the actual neuron output and the desired neuron output (For Example the desired output will be 1 if the neuron represents the required option and 0 otherwise). This process is repeated for each neuron in the network. Below is an architectural model of the perceptron with a ***Hard-Limit transfer function***¹.

This training method is widely known as *supervised learning*, as it adjusts the weights of the network by looking at a series of examples where the desired output is known. Later, when a set of inputs is presented for which the correct output is not known, the “trained” network is expected to generate the correct result. See Figure 2.7.

¹Hard-Limit transfer function. According to Valbuena, 2021, “A Hard-Limit transfer function consists of a function that maps inputs greater than 0 and 1, or equal to these values. And, for any other value, it becomes zero. That is, the perceptron neuron will produce a 1 if the network input, to the transfer function, is equal to or greater than 0, otherwise it produces a 0”.

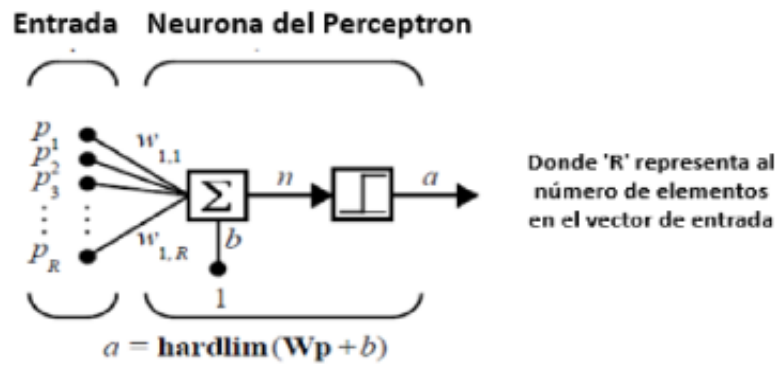


FIGURE 2.7: Perceptron Neuron Model with Hard-Limid Transfer Function by Valbuena, 2021.

The Multilayer Perceptron Model

In multilayer perceptron networks, neurons are organized integrally in layers, establishing connections with the preceding and following layers, but not with adjacent neurons in the same layer. This design guarantees a unidirectional flow of information, which is why these networks are also known as Feed-Through Networks (Valbuena, 2021). Below, the Figure 2.8 is an illustrative model:



FIGURE 2.8: Topology of a Multilayer Perceptron by Valbuena, 2021. Model taken and edited from Hopgood, 2021.

Deep Learning

Deep learning, which is a subfield of machine learning, has become one of the most impactful

and rapidly growing applications within the field of artificial intelligence. This technique is used to address extremely complex problems that often involve large volumes of data (Rouhiainen, 2018). Deep learning is presented as both a paradigm and a neural network model that has a multi-layer architecture of computational units, also known as neurons. Its effectiveness lies in its ability to learn from patterns stored in previous layers, increasing learning by applying gradients or scaling. This learning process follows an ordered, chronological sequence and specific functions. It is precisely this progression, from simple to complex, that gives it the name "Deep" (Valbuena, 2021).

As indicated by Rouhiainen, 2018, in recent times, the term *deep learning* has become popular to refer to machine learning methods that can learn more abstract concepts. These methods have proven to be especially effective in areas such as image recognition and text processing, outperforming previously available techniques. A notable example is the use of neural networks with three or more hidden layers, which have demonstrated remarkable performance in tasks such as handwritten digit recognition.

Rouhiainen, 2018 emphasizes that *Deep learning* is achieved through the use of neural networks, which are structured in layers to identify complex relationships and patterns in the data. This approach requires a vast amount of data and powerful computational processing power. Currently, it is used in various areas such as speech recognition, natural language processing, computer vision, and vehicle identification in driver assistance systems. A considerable drawback and barrier in the use of DL is the need for large datasets, which would serve as the input during the training procedure (Kamilaris and Prenafeta-Boldú, 2018)

Most Used Algorithms In Deep Learning

Gradient Descent

Torra, 2011 defines the *Gradient Descent* as an algorithm widely used in the field of artificial intelligence, especially in machine learning. It is an optimization technique that is used to adjust the parameters of a model or algorithm iteratively, to minimize a loss or error function. Is a key technique in the optimization of artificial intelligence models, allowing parameters to be adjusted iteratively to improve the performance and accuracy of the model for the specific task being addressed.

In very simple terms, Torra, 2011 expresses that, *Gradient Descent* seeks to discover the values of a model's parameters that minimize the difference between the model's predictions and the actual data values. The algorithm calculates the slope of the loss function concerning the parameters and makes adjustments in the opposite direction of the slope, allowing it to descend towards the minimum of the function. This process is repeated iteratively, updating the parameters at each step, until a point is reached where the loss function is minimized or a predefined stopping criterion is met.

Backpropagation Learning Algorithm

Backpropagation is an AI technique that allows the model to adjust, through training and added data when the first answer is not entirely correct (Adamssen, 2020). This algorithm is one of the most used throughout the history of Machine Learning and Deep Learning (Torra, 2011).

To Rumelhart et al., 1995 *backpropagation learning* is the most used method for training neural networks. Its popularity lies in its underlying simplicity and effectiveness. The effectiveness is because, unlike its predecessors, such as the perceptron learning rule or the Widrow-Hoff learning rule, this method can be used to train networks with nonlinear connections of any complexity. Because these networks are essential in real-world applications, the learning process becomes a fundamental component. However, what is equally notable is the ease with which this method can be applied.

This algorithm can calculate the error produced by each neuron in a Feed Forward network mapping process. The sum of these errors results in the total error of the network. The algorithm performs two passes: one in the input-to-output direction and one in the reverse direction. In the second pass, using the stored information, it adjusts the values of the synaptic weights of the affected neurons in reverse, which reduces the output error and improves the prediction accuracy (Valbuena, 2021).

***Large Language Models* (LLMs)**

A Large Language Model (LLM), also known as a large language model, is a type of machine learning model, which has been trained with a large amount of text data, usually on the order

of billions, with the objective to understand and produce human language effectively. These models are designed to identify and understand complex linguistic patterns, allowing them to generate coherent and relevant text (Jiménez Revuelta, 2023).

LLMs are experiencing a steady rise in popularity, both in academia and industry, thanks to their exceptional performance in various applications. As *LLMs* continue to play a critical role in both research and everyday life, evaluation of these models becomes increasingly critical. This assessment not only focuses on their capacity for specific tasks but also becomes relevant at a societal level, to better understand their potential risks and their implications on society (Chang et al., 2023).

LLM - Based On Transformers

According to Rothman and Gulli, 2022, *Transformers* are industrialized post-deep learning models designed to simultaneously execute a wide variety of tasks without requiring adjustments. These models can perform self-supervised learning processes on huge unlabeled data sets, using billions of parameters. These particular architectures that come after the deep learning era are referred to as foundational models. Artificial Intelligence as a whole, and especially in the context of Natural Language Processing for Industry 4.0 (I4.0), has advanced well beyond traditional software approaches from the past. See Figure 2.9.

The domain of natural language processing is being completely taken over by *Transformers*. These amazing models are pushing the boundaries of NLP and breaking numerous records. They are utilized in numerous applications, including conversational chatbots, machine translation, and even to fuel better search engines. *Transformers* are currently all the rage in deep learning; you may be familiar with models like BERT, GPT, and GPT-2. The paper that started all this transformation is *Attention is all you need* (Vaswani et al., 2017). A recommendation consists of the first need must understand the attention mechanism to comprehend Transformers (Chan, 2023).

Another revolution brought about by transformers in the field of Natural Language Processing (NLP) is thanks to their exceptional ability to process sequential data in a highly efficient manner. This has enabled parallelization and capturing long-term dependencies in text. An essential element of Large Language Models (LLM) lies in their emphasis on contextual learning

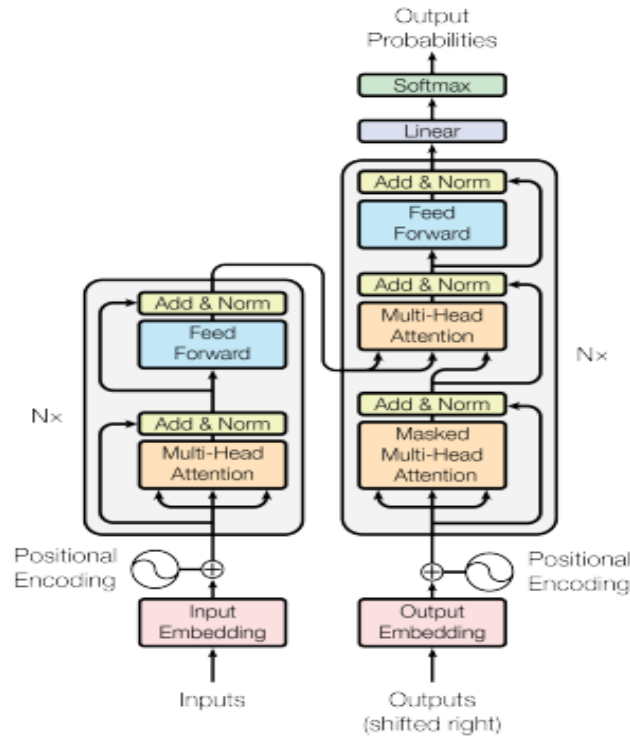


FIGURE 2.9: The Transformer - model architecture by Vaswani et al., 2017.

(Brown et al., 2020), where the model is trained to generate text based on a specific context or message.

Over the past two years, the field of Natural Language Processing (NLP) has seen astonishing progress, largely due to the influence of the transformative architecture presented in the work of Vaswani et al., 2017. This architecture has demonstrated its capacity when working with large volumes of unlabeled corpora, as evidenced in the study by Canete et al., 2020. Thanks to the growth of transfer learning and the proliferation of pre-trained language models, deep learning approaches have seen significant improvements over shallow machine learning models. In particular, BERT (Devlin et al., 2019) and XLM-RoBERTa (Liu et al., 2019), GPT-3 (Brown et al., 2020) stand out as state-of-the-art models pre-trained on NLP tasks in the study by Yaseen et al., 2021.

BERT - Bidirectional Encoder For Transformers

Is a deep neural network developed by Devlin et al., 2019 and his team, which was presented in 2018. The most well-known Transformer-based model is BERT (Kenton and Toutanova, 2019).

According Liebeskind, Elkayam, and Liebeskind, 2021:

“BERT is a bidirectional transformer pre-trained on a large corpus containing Toronto Book Corpus and Wikipedia using a combination of masked language modeling objective and next sentence prediction”.

Since its introduction, it has become a ubiquitous baseline in natural language-processing research and it has achieved state-of-the-art results in numerous benchmarks and continues to be considered a vital foundation (Rogers, Kovaleva, and Rumshisky, 2021).

In the original article by Devlin et al., 2019, two BERT models were presented: BERT-large and BERT-base, both in English. These models were pre-trained from extremely large corpora obtained from the Internet, making them demanding in terms of computational resources and memory. As BERT attracted increasing attention from the research and practitioner community, additional variants of BERT models were developed and introduced.

“There are two main approaches for using BERT: feature extraction and finetuning. In feature extraction, the architecture of the BERT model is preserved, i.e., the model’s parameters are frozen. Features are extracted from the pre-trained BERT model and then fed into a classifier model to solve a given task. In finetuning, the model’s parameters are finetuned by adding extra layers to the original BERT architecture. These new layers are used to train the model on the downstream tasks” (Alammary, 2022).

The pre-trained BERT model in its base version known as BERT base generates 768-dimensional vector representations for each token of an input text sequence, also for its fine-tuned model reaching a total of 1536 encodings. These representations are known as embeddings or encodings. For the largest version of BERT, that is, BERT Large, the numerical representations are 1024 dimensions (Devlin et al., 2018).

The process of computing input representations is carried out as follows: initially, each word in the input is tokenized into word fragments, as described in the work of Wu et al. (2016). Subsequently, three embedding layers are combined, addressing aspects of tokenization, position, and segmentation, to generate a vector of constant length. The special token [CLS] is used for classification predictions, while [SEP] is used to delimit and separate input segments

(Rogers, Kovaleva, and Rumshisky, 2021).

According to Rothman, 2021, CLS is a binary classification token that is added to the beginning of the first sequence to predict whether the second sequence follows the first. A positive example usually consists of a pair of consecutive sentences taken from a data set. Instead, a negative example is generated using sequences from different documents. SEP is a separation token that signals the end of a sequence.

To exemplify their argument, Rothman, 2021 propose the following situation:

For example, sentences taken from a book as input could be the following:

The cat slept on the rug. It likes sleeping all day

These two sentences were merged into a complete input sequence:

[CLS] the cat slept on the rug [SEP] it likes sleep ##ing all day[SEP]

This approach implies the need for additional coding data to differentiate between sequences A and B.

Putting the entire embedding process together produces the following result:

Input	[CLS]	The	cat	slept	on	the	rug	[SEP]	it	likes	sleep	##ing	[SEP]
Token	$E_{[CLS]}$	$E_{[The]}$	$E_{[cat]}$	$E_{[slept]}$	$E_{[on]}$	$E_{[the]}$	$E_{[rug]}$	$E_{[SEP]}$	$E_{[it]}$	$E_{[likes]}$	$E_{[sleep]}$	$E_{[##ing]}$	$E_{[SEP]}$
Embeddings													
	+	+	+	+	+	+	+	+	+	+	+	+	+
Sentence	$E_{[A]}$	$E_{[A]}$	$E_{[A]}$	$E_{[A]}$	$E_{[A]}$	$E_{[A]}$	$E_{[A]}$	$E_{[A]}$	$E_{[B]}$	$E_{[B]}$	$E_{[B]}$	$E_{[B]}$	$E_{[B]}$
Embeddings													
	+	+	+	+	+	+	+	+	+	+	+	+	+
Positional	$E_{[0]}$	$E_{[1]}$	$E_{[2]}$	$E_{[3]}$	$E_{[4]}$	$E_{[5]}$	$E_{[6]}$	$E_{[7]}$	$E_{[8]}$	$E_{[9]}$	$E_{[10]}$	$E_{[11]}$	$E_{[12]}$
encoding													

FIGURE 2.10: Input embeddings by Rothman, 2021.

About this, Rothman, [2021](#) also highlights *Input embeddings* are generated by the sum of token embeddings, segment embeddings (which indicate the sentence, phrase, or word to which the token belongs), and positional encoding embeddings.

The input embedding and positional encoding sub-layer of a BERT model can be summed up as follows:

- A sequence of words is broken down into WordPiece tokens.
- A [MASK] token will randomly replace the initial word tokens for masked language modeling training.
- For classification purposes [CLS] classification token is inserted at the beginning of a sequence.
- For NSP (Next Sentence Prediction) training, a [SEP] token separates two sentences (segments, phrases).
- *Sentence embedding* is added to token embedding, so that sentence A has a different sentence embedding value than sentence B.

RoBERTa - Robustly Optimized BERT Approach

RoBERTa is proposed by Liu et al., [2019](#) - Robustly Optimized BERT Approach has been trained by its creators using a 160-gigabyte corpus that includes data from sources such as Wikipedia, Common Crawl News, Book Corpus, and Webtext Corpus. This corpus is ten times larger than the one used to train the original BERT model. The training process involved 40 epochs of the corpus, for which 1024 NVIDIA Tesla V100 graphics cards were used over 24 hours (Rosado Fernández, [2020](#)).

Briskilal and Subalalitha, [2022](#) argue that RoBERTa is an improvement of the BERT model developed by Facebook AI Research (FAIR). Identified shortcomings in BERT were addressed by training RoBERTa with larger data batches and longer sequences, increasing accuracy on final tasks compared to BERT. RoBERTa implements dynamic masking, which multiplies the training data and applies various masking strategies at each step. This is in contrast to BERT, which uses static masking during data preprocessing.

Malik et al., 2023 emphasize what is highlighted by Mujahid et al., 2023 stating that RoBERTa is a model that can automatically identify various context-based features, resulting in the effective capture of real language context in specific Natural Language Processing (NLP) tasks; and emphasizes the research of Höfer and Mottahedin, 2023 arguing that RoBERTa in its multilingual version has already proven to be a benchmark in multilingual and cross-lingual comprehension performance in several languages.

Malik et al., 2023's findings indicate that there are two stages necessary to make the data suitable for insertion into the transformer model. The first involves tokenization, which consists of dividing text into smaller units. The second refers to how the information is adequately represented. In the case of RoBERTa, the tokens [CLS] are added at the beginning of each sentence and [SEP] at the end of each sentence to mark the beginning and end of the sentences. This approach allows the creation of a single vector representing all text and has become a standard pattern for the RoBERTa classifier.

For English texts, RoBERTaTokenizer is used (Liu et al., 2019), while for multilingual texts XLM-RoBERTa Tokenizer is used (Mehta and Varma, 2023). Ultimately, each token is mapped to an index based on RoBERTa's pre-trained vocabulary. RoBERTa model for multi-lingual tasks is a pre-trained uncased base model and it is used with the following parameters: hidden_size = 768, hidden layers = 12, attention_heads = 12, etc. The same parameters are used for XLM-RoBERTa.

GPT-3 - Generative Pre-Trained Transformer 3

OpenAI introduced GPT-3 in May 2020 as its newest language model, within its series of Transformer-based models for Natural Language Processing (NLP). The transformer is a type of neural network architecture that employs layers of encoders and decoders to simultaneously process a sequence of words instead of relying on memorizing entire sequences, as was done in less efficient but previously recurrent neural networks (RNNs) considered next generation (Chan, 2023).

Last generation GPT-3, or Generative Pre-Trained Transformer 3, is a transformer-based large language model developed by OpenAI². It contains 175 billion parameters. Due to its extensive language knowledge, enormous processing power, and potential to learn from vast amounts of

²<https://openai.com/>

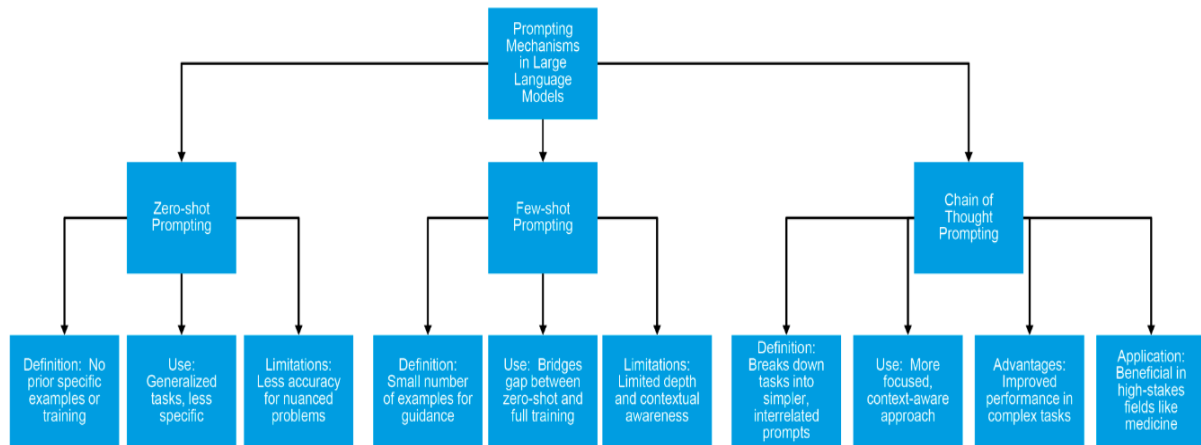


FIGURE 2.11: Prompting mechanisms in large language models by Miao et al., 2024.

online text data, GPT-3 can perform a wide variety of general natural language-based tasks with unprecedented ease such as the generation and classification of the text (Kublik and Saboo, 2022).

The enormous scale of the model allows the generation of results with quality, precision, and diversity of the generated content. This breakthrough has sparked considerable interest and concern in the NLP field, the machine learning industry in general, the media, the AI ethics communities, and civil society (Chan, 2023).

GPT-4 - Generative Pre-Trained Transformer 4

GPT-4, was developed by OpenAI, is an advanced language model presented in March 2023. GPT-4 is accessible through the OpenAI API and ChatGPT More. This model has 1.76 billion parameters. OpenAI has released two versions of GPT-4, with context lengths of 8192 and 32768 tokens, representing a significant improvement over GPT-3.5 and GPT-3, which were limited to 4096 and 2048 tokens respectively. GPT-4, unlike its predecessors, this model has the ability to process both images and text as input, becoming a multimodal intelligence model (Jiménez Revuelta, 2023).

According to Miao et al., 2024, LLMs like GPT-4 have the versatility of be significantly enhanced using specialized prompting methods. Broadly, these prompting methods fall into three categories: zero-shot, few-shot, and chain-of-thought prompting. See figure 2.11 presents the prompting mechanisms in large language models.

Prompting Zero-Shot and Few-Shot

Kublik and Saboo, 2022 emphasize, that the GPT-3 model, although primarily known for its generative capabilities, can employ several approaches to classifying text. These approaches range from zero-shot classification, where no examples are provided to the model, to one- or few-shot classification, where a few examples are shown to the model.

Miao et al., 2024 argue that, zero-shot prompting refers to a strategy employed with machine learning models, particularly Large Language Models (LLMs), wherein the model is tasked with executing a task without any preceding, explicit examples or dedicated training for that particular task. Essentially, the model leverages its pre-existing knowledge, garnered through comprehensive training on diverse datasets, to deduce the approach to tackle the novel task at hand. This capability empowers the model to formulate responses or address challenges in scenarios it hasn't undergone specific training for, thereby demonstrating its prowess in generalization. Recent studies have affirmed the efficacy of zero-shot learning across various conventional natural language comprehension tasks.

In few-shot learning, the prompt is composed including some examples of the task to be solved. The ability to learn with few attempts, never before seen in Natural Language Processing (NLP) models, is a prominent feature of GPT-3 (Chan, 2023).

Chain-of-Thought (CoT) Prompting

(Wei et al., 2022) define:

A Chain-of-Thought is a series of intermediate natural language reasoning steps that lead to the final output, and we refer to this approach as chain-of-thought prompting.

Chain-of-Thought prompting enhances the abilities of large language models (LLMs) significantly. Advancements in this methodology are poised to position it as a pivotal component in the future landscape of AI, imbuing these technologies with greater purpose, accuracy, and ethical considerations due to its adeptness in handling intricate data, coupled with its logical and sequential reasoning capabilities. This not only refines the LLMs to be more precise and contextually aware but also extends its impact across the broader spectrum of artificial intelligence (AI). Chain-of-Thought prompting expanded the usability and efficiency of AI, bringing

it into closer alignment with human cognition and decision-making patterns (Miao et al., 2024).

Wei et al., 2022 argument that *Chain-of-Thought* presents several attractive qualities as an approach to improve the reasoning process in language models:

First, the Chain-of-Thought provides the ability for models to break down complex problems into intermediate stages, meaning that additional calculations can be applied to problems that require more detailed reasoning.

Second, a Chain-of-Thought offers an interpretable view of the model's behavior, indicating how it might have arrived at a specific answer and providing opportunities for troubleshooting when the reasoning path deviates (although how the responses are characterized remains to be fully answered computations of a model that supports an answer).

Third, the Chain-of-Thought reasoning method can be applied in various tasks, such as solving math word problems, reasoning based on common sense and symbol manipulation, and theoretically, it is adaptable (at least in theory) to any task that human beings can solve using language.

Finally, the Chain-of-Thought reasoning process can be easily implemented in sufficiently broad and commercially available language models, simply by embedding examples of chain-of-thought sequences into the exemplars of few-shot prompting.

The example in Figure 2.12 illustrates how stimulating the chain of thought improves performance on various tasks of arithmetic, common sense reasoning, and symbolic reasoning.(Wei et al., 2023).

2.5 Evaluation

In the process of implementing an effective machine learning model, one of the most essential steps is to evaluate the performance of the model. Multiple metrics are used to determine its efficiency or level of quality (Sagu et al., 2023). To North, Zampieri, and Shardlow, 2023a the effectiveness of complexity prediction systems is evaluated through various performance

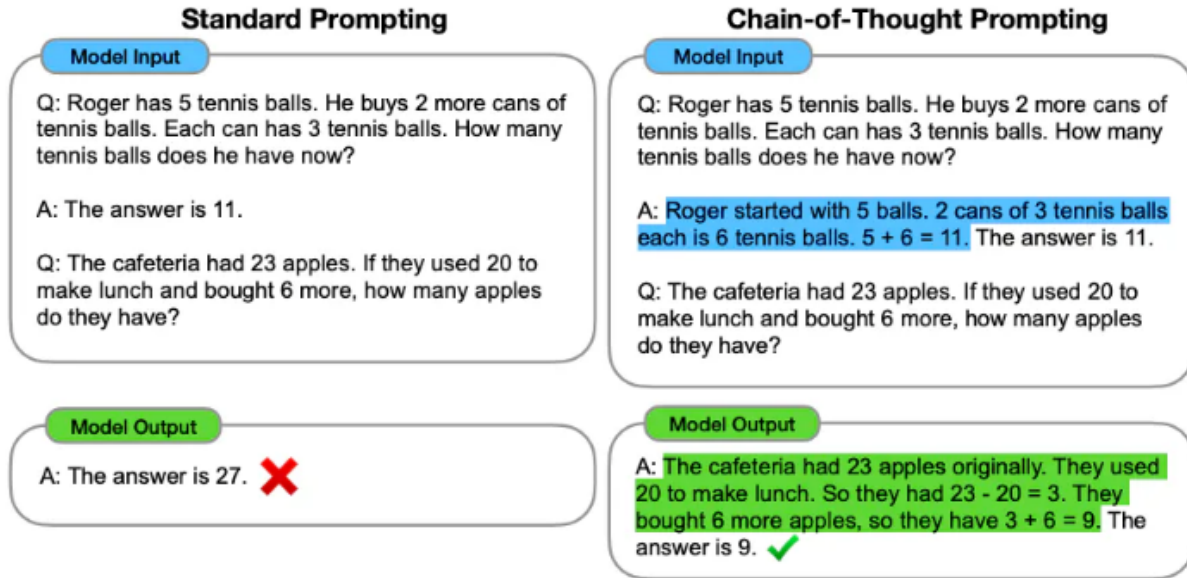


FIGURE 2.12: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted (Wei et al., 2023).

metrics, which vary from task to task.

2.5.1 Evaluation Of Complex Word Identification

According to Hackeling, 2017: “The effectiveness of binary CWI systems used to be assessed by metrics such as *Accuracy*, *Precision*, *Recall*, *F1-score*”. Kamilaris and Prenafeta-Boldú, 2018 emphasize that, these measures are widely used in deep learning implementations.

Accuracy

Is calculated by comparing the number of correct predictions to the total number of predictions, and is considered one of the most important classification metrics (Sagu et al., 2023).

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (8)$$

Precision

Is the fraction of true positives (TP, correct predictions) from the total amount of relevant results, i.e. the sum of TP and false positives (FP). Is the fraction of positive predictions made that are correct.

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

Recall

Is the fraction of TP from the total amount of TP and false negatives (FN), that is to say, is the fraction of the truly positive instances that the classifier recognizes.

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

F1-score

Is the harmonic average of the accuracy and recall scores.

$$F1 = 2 * \frac{TP * FP}{TP + FP} \quad (11)$$

2.5.2 Evaluation Of Lexical Complexity Prediction

In the evaluation of the latest LCP (Lexical Complexity Prediction) systems aimed at predicting complexity values in continuous rather than binary form, metrics such as mean absolute error, mean square error, Pearson's correlation, and Spearman's correlation coefficient are frequently used (North, Zampieri, and Shardlow, 2023a). Below is a detailed description of each metric used in the (Nokeri, 2021) and (Shardlow et al., 2021a) investigations:

- **MAE - Mean Absolute Error**

An absolute error is a difference between the estimated value and the true value. Mean Absolute Error (MAE) represents the average magnitude of error in estimates without considering the direction. Typically used for the evaluation of regression tasks, it indicates how close the predicted labels are to the gold labels of the task.

$$MAE = \frac{1}{n} \sum_{k=1}^n |y_k - \hat{y}_k| \quad (12)$$

Where n represents the number of error.

- **MSE - Mean Square Error**

Mean Squared Error (MSE) represents the variability explained by the model about the data after considering a regression relationship. There is little difference in the calculation of MSE versus MAE, however in the tasks also include this metric for completeness.

$$MSE = \frac{1}{n} \sum_{k=1}^n (y_k - \hat{y}_k)^2 \quad (13)$$

- **RMSE - Root Mean square Error**

Mean Square Error (RMSE) indicates that it represents the variability explained without considering a regression relationship. To obtain an RMSE score, calculate the square root of the MSE.

$$RMSE = \sqrt{\frac{1}{n} \sum_{k=1}^n (y_k - \hat{y}_k)^2} \quad (14)$$

- **Pearson - Pearson's Correlation Coefficient**

Pearson correlation: This is a well-known and understood metric, especially in the context of evaluating systems with continuous outcomes. Pearson correlation is robust to changes in scale and measures how the input variables change relative to each other.

$$R_{X,Y} = \frac{cov(X,Y)}{\sigma_X \cdot \sigma_Y} \quad (15)$$

- **Spearman's Rank**

This metric does not consider the values generated by a system, nor in the test labels, only the order of those labels. In research, it is usually chosen as a secondary metric because it is more robust against outliers than the Pearson correlation.

$$p = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (16)$$

- **R-Square (R^2)**

R^2 represents the variability explained by the model about the data. The metric comprises values that range from 0 to 1, where 0 shows that the regression model poorly explains the variability in the data, and 1 shows that the regression model best explains the data. This measures the proportion of variance of the original labels captured by the predicted labels. It is possible to do well on all the other metrics, yet do poorly on R^2 if a system produces annotations with a different distribution than those in the original labels.

$$R^2 = 1 - \frac{\sum_{k=1}^n (y_k - \hat{y}_k)^2}{\sum_{k=1}^n (y_k - \bar{y})^2} \quad (17)$$

According (Pilicita and Barra, [2023](#)):

- **Validation Loss:** Evaluate the model error on the training data set .
- **Training Loss:** Evaluate the performance of a deep learning model on the test set.

2.6 Previous Work

The study of LCP has aroused considerable interest due to several fundamental reasons. Although a considerable amount of work has been done and evaluation campaigns carried

out, reference corpora are largely lacking, representing a significant challenge to progress in this field. Particularly in the case of the Spanish language, reference corpora are scarce or even non-existent, which has hindered progress in the understanding and evaluation of lexical complexity. In the field of algorithms and techniques used, various approaches coexist. On the one hand, we find solutions based on feature engineering together with classic machine learning algorithms. On the other hand, we observe the emergence of solutions based on Deep Learning, where classification techniques using encoded representations of texts in embeddings of transformer encoding architectures stand out. In addition, there is a promising use of large language models (generators), which has still been very little explored.

This thesis represents a significant advance in the study of lexical complexity for the Spanish language. New data sets have been created and new solutions have been explored to predict lexical complexity, resulting in contributions that have marked a new milestone in the current state of the art in some specific cases.

In recent times, there has been a growing research interest that seeks to advance the study of CWI. The primary objective of these applications is to streamline the assessment of word complexity by analyzing various indicative features, as highlighted in the work conducted by Shardlow, Cooper, and Zampieri (2020b). This study introduces the CompLex: A New Corpus for Lexical Complexity Prediction from Likert Scale Data introduces a novel dataset aimed at advancing lexical complexity prediction (LCP). The dataset, called CompLex, is annotated at the word level using a Likert scale ranging from 1 to 5, capturing the complexity of words as perceived by human annotators. This nuanced annotation approach allows for more detailed predictions of word complexity compared to traditional binary classification tasks. The dataset is drawn from a range of text genres, including Wikipedia, biomedical texts, and European Parliament debates, offering a rich and diverse set of linguistic contexts.

CompLex uses data from three different sources: the Bible, Europarl, and biomedic texts (see Table 1). It is a set of multidomain English data made up of sentences, the targeted token, and its respective level of complexity as described. See Table 2.3.

In their experiments, the authors applied several machine learning models to predict lexical complexity. These included traditional regression models like linear regression and support vector regression, as well as more advanced deep learning models. Transformer-based models,

	Contexts	Unique Words	Median Annota- tors	Mean Complexity
All	9476/7974/1500	5166/3903/1263	7.0/7.0/7.0	0.394/0.385/0.442
Europarl	3496/2896/600	2194/1693/501	7.0/7.0/7.5	0.390/0.381/0.433
Biomed	2960/2480/480	1670/1250/420	7.0/7.0/7.0	0.407/0.395/0.470
Bible	3020/2600/420	1705/1362/343	7.0/7.0/8.0	0.385/0.379/0.422

TABLE 2.3: The statistics for CompLex. Each cell shows three values, which are split according to the statistics for ‘All’ / ‘Single Words’ / ‘Multi Words’.

particularly BERT (Bidirectional Encoder Representations from Transformers), were used to generate contextual embeddings of words, capturing both lexical and contextual features. Additionally, the authors experimented with ensemble methods, combining the outputs of different models to improve performance.

As a classifier, they chose a linear regression. The authors also used Glove embeddings and infersent. They may find that using embeddings which adapt to the context, such as in BERT and a neural network for prediction would yield stronger results. However, in this work they have only aimed to give an understanding of what types of features can be useful for predicting the values in our corpus, not to produce a state of the art system for the prediction of lexical complexity. They can see that there are significant differences in the mean absolute error for each sub-corpus. Whereas the mean absolute error was lower for Europarl (0.0801), it was higher for the Bible and Biomed, indicating that the type of language in these two corpora was more difficult to model. This is reflected across different feature subsets, indicating it is a feature of the dataset and not a random fluctuation of our model.

The CompLex dataset, along with these techniques and results, provides a strong foundation for further research in lexical complexity prediction, with applications in improving readability and text accessibility for non-native speakers and individuals with varying levels of literacy.

The last few years have indeed witnessed remarkable progress in Natural Language Processing (NLP), and the Transformer architecture, introduced by Vaswani et al. (2017), has played a pivotal role in driving this advancement. Transformers have proven to be highly effective, especially when leveraging large amounts of unlabeled text corpus, as highlighted by the work of Canete et al. (2020). This architecture’s attention mechanism has allowed models to capture complex contextual relationships in language, leading to significant improvements in various NLP tasks.

Deep learning models, particularly with the advent of transfer learning and pre-trained language models, have surpassed the capabilities of shallow machine learning models. BERT and XLM-RoBERTa, as mentioned in Yaseen et al., 2021, stand out as state-of-the-art pre-trained language models, demonstrating exceptional performance across a wide range of NLP tasks. Their ability to capture intricate contextual information during training and apply it to various downstream tasks has marked a paradigm shift in the field. The success of these models underscores the importance of contextual embeddings and transfer learning in addressing the complexities of natural language, contributing to the ongoing evolution of NLP methodologies and applications.

In their study, Uluslu, 2023 addressed the automatic simplification of lexical text in Turkish, a morphologically rich language, which presents a particular challenge in the field of natural language processing. The research is significant for expanding the scope of text simplification in languages with complex grammatical structures and multiple affixes, such as Turkish. In his research, he points out the need to simplify Turkish texts to improve accessibility, especially for people with reading comprehension difficulties, such as students, elderly people, or people with cognitive disabilities. The research is also relevant for applications such as language teaching and web content adaptation. The main focus of this work is lexical simplification through a process of replacing complex words with simpler words. To do so, the authors use machine learning techniques and statistical methods such as:

Word Embeddings: The authors implement word vector models, such as Word2Vec and FastText, to capture the semantic relationships between words in the Turkish language.

Substitution Selection Model: To determine the most appropriate substitutions, the authors employ a classifier that combines features of lexical frequency, contextual synonyms, and semantic distance between words.

Lexical Complexity Filter: A filter is used to identify words that are more likely to be complex for readers, using metrics such as frequency of use and word length.

To train and evaluate the simplification models, the authors use a specific corpus of the Turkish language. This corpus was created by extracting texts from different domains and annotating the lexical difficulty of the words, highlighting the importance of working with

human-annotated texts to improve the accuracy of the system, see Table 2.4. The best results were obtained using a combination of lexical embeddings and a classification system that filters out complex words with greater accuracy. The authors manage to reduce text complexity while maintaining semantic coherence and fluency, which is essential in simplifying languages with rich morphological structures.

Total Dataset	Training Data	Test Data
3k Sentences	2,650 Sentences	350 Sentences

TABLE 2.4: Turkish CWI Dataset.

Metrics such as fluency, meaning preservation, and complexity reduction are used to evaluate the performance of the proposed systems. Optimal results: The system achieved high accuracy in lexical simplification, especially in selecting simpler words in given contexts, outperforming previous models. This work expands the knowledge on lexical simplification in morphologically complex languages, providing an innovative approach that could be adapted to other non-analytic languages. The methods used in this paper highlight the importance of lexical embeddings and semantic classifiers to identify and replace complex words, contributing to the development of more accessible systems in natural language processing.

2.6.1 ALexS 2020 - Lexical Analysis in SEPLN 2020 Lexical

Linguistic complexity assessment (LCP) has been the focus of attention in several international competitions, known as shared tasks. Below is a brief description of the architecture, development, and evolution of the systems participating in these shared tasks: The ALexS (Lexical Analysis in SEPLN 2020 Lexical) Analysis Workshop at SEPLN, was part of IberLEF 2020³ - Iberian Languages Evaluation Forum, hosted by and co-located with the International Conference of the Spanish Society on Natural Language Processing - SEPLN 2020⁴. IberLEF is the result of the association of some workshops in the Natural Language Processing (NLP) domain for Spanish and other languages spoken in the Iberian peninsula, and the aim of joining forces of different NLP research communities to provide a common forum for assessing NLP systems and interchanging research ideas, issues, challenges, and experiences.

³<https://sites.google.com/view/iberlef2020/>

⁴<http://sepln2020.sepln.org/index.php/iberlef/>

In September 2020, the first edition of the ALexS workshop (Lexical Analysis Task in SE-PLN) was held in Malaga, Spain as part of the second edition of IberLEF (Iberian Language Evaluation Forum), which added the efforts of IberEval and TASS. Only one task was proposed: Complex Word Identification (CWI) (Ortiz-Zambrano and Montejo-Ráezb, 2020). More than seven teams joined the campaign, but only three of them finally presented the results and a description of their systems. The difficulty of the task, due to the lack of labeled data for the participants, has forced interesting approaches to address the CWI problem in an unsupervised or semi-supervised manner.

The dataset provided to participants was the VYTEDU-CW (Ortiz-Zambrano et al., 2019) corpus as a variant of the original VYTEDU corpus (Ortiz-Zambrano and Montejo-Ráez, 2017). The collection contains 55 texts which correspond to transcripts of academic videos in Spanish made within the classrooms of the different courses of the University of Guayaquil. The VYTEDU-CW corpus (Videos and transcripts in the educational field - Complex words) is made up of more than 1,200 words per transcript on average, it has a total of 9,175 different words. Its dataset contains 723 words annotated as complex (difficult) terms that are present in the different documents, these difficult words were identified and tagged by 430 annotators (students), 250 students tagged words that other users had not selected, that is, that did not match those annotated by other students.

The ALexS 2020 workshop focused on two main tasks: Lexical Complexity Prediction: Identifying words or phrases that may be difficult for users to understand in different contexts. Automatic Language Simplification: Developing methods to simplify text, preserving meaning but making it more accessible. The workshop presented several advanced approaches and techniques to address these problems, including:

Machine Learning and Deep Learning Models: Workshop participants employed a variety of machine learning and deep learning methods for identifying complex words. These include models such as Random Forests, Support Vector Machines (SVMs), and neural network-based techniques such as LSTM and BERT.

Word Embeddings: As with many modern NLP approaches, participants used word representations using embeddings such as Word2Vec and FastText, as well as cutting-edge embeddings such as BERT and ELMo, which capture the contextual meaning of words.

Lexical Complexity Metrics: Several proposals used complexity metrics based on factors such as word frequency, word length, and morphological structure. Other approaches also considered cognitive aspects, such as the processing effort an average reader might need to understand a specific word.

Table 2.5 shows the results obtained on the ALexS task for complex word identification on the VYTEDU-CW dataset. UDLAP team obtained the overall best results in terms of F-score (0.2725) and precision (0.3415) (Rico-Sulayes, 2020). Vicomtech system performed well on recall (0.6881) but at the price of a very low precision (Zotova et al., 2020). Although up to three different runs were allowed, only the UDLAP and Vicomtech submitted three annotations, while HULAT team only submitted one annotation (Alarcón, Moreno, and Martínez, 2020).

Team	Run	Macro F1	Macro Precision	Macro Recall
UDLAP	Method 2	0.272528	0.341598	0.226691
UDLAP	Method 1	0.266004	0.334722	0.220696
UDLAP	Method 3	0.261395	0.328729	0.216955
Vicomtech	Method 1	0.166623	0.096827	0.596863
Vicomtech	Method 3	0.176915	0.103961	0.593173
Vicomtech	Method 2	0.162051	0.091838	0.688192
HULAT	Method 1	0.164521	0.093712	0.673171

TABLE 2.5: ALexS - Official results from (Ortiz-Zambrano and Montejo-Ráezb, 2020).

The best results of the workshop came from deep learning models that used advanced contextual representations such as BERT and RoBERTa, significantly outperforming traditional rule-based and classical machine learning methods. The best techniques were able to accurately identify complex words in different contexts. Approaches based on BERT and other contextual embedding models showed greater generalization capacity. The results in this task showed that transformer-based techniques were more effective in suggesting simpler words, while maintaining the original meaning of the text.

The main contributions of the workshop include the creation and evaluation of new methodologies to predict lexical complexity in Spanish, providing new insights and algorithms applicable to other languages with complex structures. The promotion of collaboration between researchers from different areas of NLP and the impulse to develop annotated lexical corpora

in Spanish. The advances in the use of pre-trained language models such as BERT to improve both complexity prediction and automatic text simplification. The approaches and data sets introduced in the workshop have laid the groundwork for future research in the field of lexical analysis and text simplification, opening opportunities to apply these techniques to other languages and domains.

2.6.2 LCP-2021 at SemEval 2021

The task of Lexical Complexity Prediction was carried out at SemEval-2021⁵ - 15th International Workshop on Semantic Evaluation, SemEval and SemEval2021 (Shardlow et al., 2021a). The competition provided participants with an augmented version of the CompLex Corpus (Shardlow, Cooper, and Zampieri, 2020b). CompLex is a multidomain English corpus in which words and multiword expressions (MWE) were scored concerning their complexity using a five-point Likert scale. SemEval2021 Task 1 included two subtasks: Subtask 1 focused on individual words and Subtask 2 focused on MWE. The competition attracted 198 teams in total, of which 54 teams submitted official analyses of the test data for Subtask 1 and 37 for Subtask 2.

The dataset contains in total 5,617 unique words are divided into 10,800 contexts, with an average complexity across our entire data set of 0.321. Each genre has 3,600 contexts, each divided between 3,000 individual words and 600 multi-word expressions. While single words are slightly below the average complexity of the data set at 0.302, multiword utterances are much more complex at 0.419, indicating that annotators found them more difficult to understand. There were 55 teams participating in the 2 subtasks, 19 participating only in Subtask 1, 1 participating only in Subtask 2, and 36 participating in both subtasks. Pearson correlation was applied for the final classification of participants, but other metrics that are appropriate for evaluating continuous and classified data were also included, and secondary classifications of these were provided.

Subtask 1 asked participants to assign complexity values to each of the individual word instances in our corpus. Subtask 2, consisted of participants submitting results for both individual words and MWE. Participants were not classified based on MWE presentations alone due to the relatively small number of MWEs in the corpus (184 in the test set). The metrics that the

⁵SemEval 2021 - The 15th International Workshop on Semantic Evaluation.
Available in <https://semeval.github.io/SemEval2021/>

organizers chose for the classification were the following: Pearson correlation, Spearman rank, Mean Absolute Error (MAE), Mean Square Error (MSE), R^2 .

The analysis of the systems involved in the task revealed that there was little variation between deep-learning approaches and feature-based approaches. However, despite this similarity, the Deep Learning approaches ultimately achieved the highest scores in the data. However, both strategies remain relevant in the context of Lexical Complexity Prediction (LCP). Interestingly, systems that claimed not to use context underperformed in overall evaluations like the work done by Zambrano and Montejó-Ráez, 2021 who developed a machine learning approach focusing on 15 word-level linguistic features.

Their methodology involved training a supervised random forest regression algorithm using a specific set of features. Several runs of the algorithm, each with different parameter values, were performed to systematically evaluate its performance. The best results of the algorithm achieved MAE = 0.07347, MSE = 0.00938, and RMSE = 0.09687. These metrics are indicative of the algorithm's accuracy in predicting word complexity based on selected features. Low MAE, MSE, and RMSE values provided a high level of prediction accuracy, demonstrating the effectiveness of their machine learning approach for this particular task.

The Table 2.6 shows the final results of the Lexical Complexity Prediction task on the single words dataset according to MAE.

#	Team Name	MAE	MSE	R^2
1	JUST_Blue	0.0609	0.0062	0.6172
2	DeepBlueAI	0.0610	0.0061	0.6210
3	Alejandro M.	0.0619	0.0064	0.6062

TABLE 2.6: Final results of the LCP task on the single words dataset according to MAE.

2.6.3 SimpleText at CLEF 2023

CLEF 2023⁶ held the 14th CLEF conference continuing the popular CLEF campaigns which have run since 2000 contributing to the systematic evaluation of information access systems,

⁶CLEF 2023 - Conference and Labs of the Evaluation Forum
Available in <https://clef2023.clef-initiative.eu/>

primarily through experimentation on shared tasks.

Building on the format first introduced in 2010, CLEF 2023 consists of an independent peer-reviewed conference on a broad range of issues in the fields of multilingual and multimodal information access evaluation and a set of labs and workshops designed to test different aspects of mono and cross-language information retrieval systems. Together, the conference and the lab series maintained and expanded upon the CLEF tradition of community-based evaluation and discussion on evaluation issues.

SimpleText⁷ track at CLEF 2023 faces the challenges of text simplification methods to promote access to scientific information. Provided a corpus of scientific literature abstracts and popular science requests. Three tasks were proposed. First, the content selection task (what is in, or out?) poses a challenge to systems, as they must choose which passages to include in a simplified summary in response to a query, deciding what information should be present, and what information should be omitted. Second, the goal of complexity detection (what is unclear?) is to identify what is unclear in a passage and a query. It seeks to classify the terms or concepts that require additional explanation to understand the passage, including definitions, context, and applications. Third, the text simplification task (rewrite this!) is to simplify passages of scientific abstracts in response to a query, while keeping the essential content intact Ermakova et al., 2023.

CLEF 2023 SimpleText Task 2: Difficult Concept Identification and Explanation

“Task 2: What is unclear?” Identification and explanation of difficult concepts” in the context of the Automatic Simplification of Scientific Texts (SimpleText) laboratory, executed as part of CLEF 2023. Concerning this task, a set of 39 proposals in total were received, and presented by 12 different teams. Task 2 focuses on identifying complexity in scientific texts (passages). Consequently, its purpose is directed toward detecting terms or concepts that require specific prior knowledge to understand the passage, evaluating its level of complexity for those who are not experts, and providing explanations for difficult concepts that are identified (Ermakova et al., 2023).

⁷SimpleText is a track organised as a part of CLEF 2023 conference, initiated by CLEF initiative. Available in <http://simpletext-project.com/2023/clef/>

The datasets for Task 2.1 was sent to the participants receiving a training set comprising 203 pairs of sentences along with their corresponding scientific terms, annotated real data, and difficulty scores on a scale of 0 to 2. The test set for Task 2.1 included a total of 116,763 sentences extracted from DBLP summaries. A subset of 1,262 unique sentences was manually evaluated to evaluate the performance of various models in identifying challenging terms and assigning appropriate difficulty scores. To ensure comprehensive evaluation, a clustering mechanism was implemented, resulting in the annotation of 5,142 different pairs of sentence term combinations. The metrics applied in this task were: the correctness of detected term span (limits) and difficulty scores:

In the datasets for Task 2.2, the assessment procedure for this task with approximately 800 terms accompanied by genuine definitions, served as a performance benchmark for the participants' systems. The performance of the systems is assessed by comparing their results with the genuine definitions. Adequate annotated samples have been included to assess the majority of runs. In total, the test set comprises 15,056 sentences, each containing at least one annotated term, which is employed to appraise the runs. The ultimate test set encompasses 38,416 sentences, each containing at least one abbreviation, and it is used for the final evaluation of this subtask. The metrics applied in this task were: BLEU score, ROUGE L F-measure, Semantic match, Exact match, and Partial match.

A total of 12 different teams participated in the task, presenting a total of 39 executions that covered a wide variety of approaches to address the challenges posed. These approaches range from conventional statistical methods to the use of Large Language Models (LLM). It was noted that even when using similar models, results can vary significantly due to factors such as the specific implementation, adjustment techniques, and the choice of cues used during the simplification process. LLMs demonstrated comparable results to unsupervised methods in detecting difficult terms and providing definitions. This observation highlights the importance of the balance between efficiency and effectiveness in text simplification, where the choice of approach must consider both aspects to obtain optimal results.

In the results of CLEF 2023's SimpleText Track, this track revealed significant progress in text simplification, focused on simplifying structure and syntax, making them more accessible and understandable. However, despite advances in syntax, there remains a significant challenge related to lexical complexity, which remains a major obstacle to understanding scientific texts.

These results indicate that there is still room for improvement in this area.

Results of Task 2.1: Difficult Term Spotting

Among all runs of Task 2.1, the SINAI_task_2.1_PRM_ZS_TASK2_1_V1 run achieved the highest number of correctly detected terms and scores (Ortiz-Zambrano, Espin-Riofrio, and Montejó-Ráez, 2023b). The authors investigated zero-shot and few-shot learning strategies over the auto-regressive model GPT-3, and in particular effective prompt engineering. SINAI achieved the highest score in task 2.1 with the application of its first approach applying Zero Shot learning using the GPT3 model. The performance of different runs on both the train and test sets (only in properly constrained terms) is presented in Table 2.7.

	Total	Test		Train	
		Evaluated	Score	Evaluated	Score
SINAI_task_2.1_PRM_ZS_TASK2_1_V1	11081	1185	507	94	56
UAMS_Task_2_RareIDF	675090	1145	241	40	21
SINAI_task_2.1_PRM_FS_TASK2_1_V1	10768	1122	405	81	40

TABLE 2.7: SimpleText Task 2.1: Results for the official runs on train and test sets. The evaluation is done based on the terms with their limits correctly detected.

Results of Task 2.2: Difficult Term Explanation

For Task 2.2, ten teams conducted 29 runs, showing strong performance in terms of semantic similarity. They achieved high BLEU metric scores, as illustrated in Table 2.8. These teams were proficient in providing explanations for challenging terms that closely matched the reference definitions. Runs utilizing Wikipedia as a knowledge source displayed the highest similarity to ground truth definitions, highlighting the advantage of leveraging Wikipedia as a resource.

Run	Evaluated	BLEU	ROUGE	Semantic
UBO_task_2.1_FirstPhrases_Wikipedia	393	29.73	0.41	0.80
Croland_task_2_PKE_Wiki	43	33.68	0.46	0.70
MiCroGerk_task_2.2_GPT-3_Wikipedia	932	26.38	0.41	0.75

TABLE 2.8: SimpleText Task 2.2: Results for the official runs on the test set.

An interesting observation is that test set scores were higher than training set scores, as seen in Table 2.9, which outlines the performance of runs in the abbreviation expansion task. The “MiCroGerk_task_2.2_GPT-3” run demonstrated the best performance, providing 326 expansions identical to real ones and 185 partially correct expansions. Overall, Language Models (LLMs) like BLOOMz and GPT-3 excelled in abbreviation expansion.

Run	Evaluated	BLEU	ROUGE	Semantic
MiCroGerk_task_2.2_GPT-3_BLOOMZ	854	13.87	0.68	0.76
MiCroGerk_task_2.2_GPT-3_GPT-3	855	11.86	0.64	0.73
MiCroGerk_task_2.2_GPT-3_Wikipedia	855	4.68	0.43	0.60

TABLE 2.9: SimpleText Task 2.2: Results for the official runs on the abbreviation expansion task.

2.6.4 DeTermIt! at LREC-COLING 2024

DeTermIt! Workshop Evaluating Text Difficulty in a Multilingual Context. Co-Located with the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation LREC-COLING 2024⁸ Co-Located with the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation.

The DeTermIt! Evaluating Text Difficulty in a Multilingual Context workshop explores the theoretical and practical perspectives surrounding the evaluation of text difficulty in a multilingual context. In today’s interconnected world, where information dissemination knows no linguistic bounds, it is mandatory to ensure that knowledge is accessible to diverse audiences, regardless of their language proficiency. Automatic Text Simplification (ATS) is the process that involves the reduction of linguistic complexity within a text to enhance its comprehensibility and readability. ATS plays a fundamental role in enhancing content and conveying clear, unambiguous information (comprehensibility and readability) for diverse audiences (Di Nunzio et al., 2024). The DeTermIt! workshop builds upon the recent achievements of several initiatives that addressed specific areas within our realm of interest. It is aligned with the CLEF SimpleText Track which provides appropriate reusable data and benchmarks for scientific text summarization and simplification. DeTermIt! aims to bring together researchers and practitioners in the field of text simplification, with a particular focus on the intersection of lexicography,

⁸LREC-COLING 2024 - DeTermIt! Workshop on Evaluating Text Difficulty in a Multilingual. Available in <https://determinit2024.dei.unipd.it/#conference>

terminology, and keyword extraction. This workshop will explore the theoretical and practical perspectives surrounding the evaluation of text difficulty in a multilingual context, and it will serve as a platform for discussing advancements, methodologies, and applications in simplification techniques that target different linguistic nuances and audiences.

(Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024b) described an experiment to evaluate the ability of the GPT-3 language model to classify terms regarding their lexical complexity. This was achieved through the creation and evaluation of different versions of the model: text-Davinci-002 y text-Davinci-003 and prompts for few-shot learning to determine the complexity of the words. They used the *CompLex* corpus proposed by Shardlow, Cooper, and Zampieri, 2020b. *CompLex*⁹ is the first English multiple domain dataset, where words are scored in a context concerning their complexity using a five-point Likert scale to label complex words in texts of three sources/domains: the Bible, Europarl and Biomedical texts. The corpus is split into single-word and multiple-word annotations. The corpus contains a total of 9,476 sentences, each annotated by approximately 7 annotators, see Table 2.10.

	All	Single words	Multiple words
Europarl	3,496	2,896	600
Biomed	2,960	2,480	480
Bible	3,020	2,600	420
All	9,476	7,974	1,500

TABLE 2.10: Volumetric information for the single and multiple words in each subcollection of the *CompLex* dataset.

Each entry contains the name of the source corpus, the sentence representing the context of the word, the targeted word to classify, and a score (in training data) that is an average of the scores given by different annotators, taking into account that each class has a predefined weight between 0 (very easy) and 1 (very difficult) in a linear distribution over possible classes, see Table 4.58. Sample entries are shown in Table 2.12. This resource was used as a benchmark collection for SemEval 2021 Task 1: Prediction of Lexical Complexity of the 15th International Workshop on Semantic Evaluation¹⁰ (Shardlow et al., 2021a).

The results obtained on the *CompLex* dataset achieve a minimum average error of 0.0856. Although this is not better than the state of the art (which is 0.0609), it is a performing and

⁹CompLex: Is available at <https://github.com/MMU-TDMLab/CompLex>

¹⁰Semeval 2021 - LCP SHARED TASK 2021 - <https://sites.google.com/view/lcpsharedtask2021>

Scale	Description	Complexity
<i>Very Easy</i>	Words which were very familiar to an annotator.	0
<i>Easy</i>	Words with which an annotator was aware of the meaning.	0.01 - 0.25
<i>Neutral</i>	A word which was neither difficult nor easy.	0.26 - 0.50
<i>Difficult</i>	Words in which an annotator was unclear of the meaning, but may have been able to infer the meaning from the sentence.	0.51 - 0.75
<i>Very Difficult</i>	Words that an annotator had never seen before, or were very unclear.	0.76 - 1.00

TABLE 2.11: Categories on the Likert scale proposed by (Shardlow, Cooper, and Zampieri, 2020b).

Corpus	Sentence	Token	LCP score
Bible	<i>He sees the place of stones.</i>	stones	0.3421
Bible	<i>But I will stay at Ephesus until Pentecost,</i>	Pentecost	0.6250
Bible	<i>These are the families of the Levites.</i>	families	0.2205
Bible	<i>The seeds rot under their clods.</i>	clods	0.6250
Biomed	<i>p150CAF-1 knockdown in ES cells was quantified.</i>	ES	0.6944
Biomed	<i>The 2P unique region (Region I) contains an hg</i>	hg	0.7500
Biomed	<i>on behalf of the PPE Group.</i>	Group	0.1527
Europarl	<i>We have taken note of your comment, Mr Helmer.</i>	comment	0.0499
Europarl	<i>Country Strategy Papers - Malaysia, Brazil</i>	Strategy	0.2894
Europarl	<i>Documents received: see Minutes</i>	Documents	0.2000
Europarl	<i>Situation in Darfur (vote)</i>	Situation	0.2115

TABLE 2.12: Examples form the CompLex dataset where the complex word is highlighted in bold.

promising approach to lexical complexity prediction without the need for model fine-tuning.

2.6.5 SimpleText at CLEF 2024

SimpleText¹¹ is a track organised as a part of CLEF 2024¹² conference. Co-Located CLEF 2024 Conference and Labs of the Evaluation Forum Information Access Evaluation meets Multilinguality, Multimodality, and Visualization.

CLEF 2024 is the 15th CLEF conference continuing the popular CLEF campaigns which have run since 2000 contributing to the systematic evaluation of information access systems, primarily through experimentation on shared tasks. was organized by the University of Grenoble Alpes, Grenoble, France, from 9 to 12 September 2024. CLEF 2024 was the 15th year of the CLEF Conference and the 25th year of the CLEF initiative as a forum for IR Evaluation, so it marked an important anniversary for CLEF. The SimpleText project focuses on improving accessibility to scientific texts for a broad audience, including students and experts, by tackling technical and evaluation challenges. It provides reusable data and benchmarks for scientific text summarization and simplification. The project includes four main tasks: retrieving relevant passages from academic abstracts to help explain topics in popular science articles, identifying and explaining difficult concepts within scientific abstracts, simplifying sentences extracted from these abstracts, and developing systems to track the state-of-the-art in AI publications by recognizing and extracting key metrics from papers. Together, these tasks aim to enhance understanding and access to complex scientific information for a wide audience (Goeuriot, 2024).

Task description

To generate a definition and an explanation for each difficult term. The goal of this task is to identify key concepts that need to be contextualized with a definition, example or use case, and provide useful and understandable explanations for them. In this task, for each term that refers to a difficult concept (those that have been evaluated with the highest level of difficulty), participants should provide the definition and explanation which will be evaluated both from a qualitative point of view (manual review by terminologists) and from a quantitative point of

¹¹SimpleText - Is a track organised as a part of CLEF 2024 conference, initiated by CLEF initiative.

Available in <https://simpletext-project.com/2024/en/>

¹²CLEF 2024 - Conference and Labs of the Evaluation Forum Information Access Evaluation meets Multilinguality, Multimodality, and Visualization.

Available in <https://clef2024.clef-initiative.eu/index.php>

view (overlapping text measures, for example, BLEU score).

Data

The corpus of Task 2 is based on the sentences in high-ranked abstracts to the requests of Task 1. and collected in 2023 (Ermakova et al., 2023). A total of 175 documents and 1,077 sentences were used to generate the training and test data. In particular, we had 115 documents and 576 sentences for building the training set and 60 documents and 501 sentences for building the test set.

Team and Approach Description

Table 2.13 below presents the main approaches of each participant who submitted at least one execution of a model to be evaluated on the test set.

Results

The results are products of the execution of various techniques and strategies implemented by the competing teams to provide a possible solution to the proposed task. Each team applied innovative approaches, using cutting-edge language models, natural language processing, and prompt engineering techniques. The results presented reflect both the diversity of the approaches and their effectiveness in addressing the specific challenges of the task. The results on the test set are summarized for each run (Faggioli et al., 2024a).

The results of the test set are summarized in Table 2.14 For each run, the following were considered:

- the recall of all the terms, independently from the level of difficulty;
- the precision of all the terms, independently from the level of difficulty;
- the F1 score of all the terms, independently from the level of difficulty;
- the recall of the difficult terms;
- the precision of the difficult terms;
- the F1 score of the difficult terms;

Summary of Team and Approach Description in Task 2: Identify and Explain Difficult Concepts

Team / Approach Description
AB&DPV - Varadi and Bartulović, 2024 submitted one run, employing natural language processing techniques to identify difficult terms within passages. They generated definitions for these terms or retrieved them from sources like Wikipedia.
AIIRLab - Largey et al., 2024 submitted three runs, utilizing LLaMA3 and Mistral language models. Their approach included prompt engineering and reinforcement learning with human feedback to enhance the quality of outputs generated by the LLaMA model.
Dajana&Kathy - Seng and Simunovic, 2024 submitted one run, using the LLAMA-2 13B model. No further information about their approach was provided.
Frane&Andrea - Zečević and Doljanin, 2024 submitted one run. No additional details about their methodology were given.
Sharingans - Ali et al., 2024 submitted one run, fine-tuning the GPT-3.5 turbo model for selecting difficult terms and generating definitions and explanations. They employed prompt-engineering techniques to create specific prompts that guided the model in producing accurate and contextually relevant definitions.
SINAI - Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024 submitted three runs, applying learning cues without prior examples to the GPT-4-Turbo model. They used the OpenAI API in Python to interact with the model, facilitating the integration of GPT-4-Turbo into their workflow.
team1_Petra_and_Regina - Elagina and Vučić, 2024 submitted one run, combining named entity recognition (NER) techniques with rule-based approaches to identify and extract entities such as proteins, genes, and chemical compounds. They utilized spaCy for NER and developed custom rules for entity extraction.
Tomislav&Rowan - Mann and Mikulandric, 2024 submitted two runs. They created prompts for the LLAMA-2 13B model to extract three scientific terms from each source sentence and then prompted the model to return a difficulty rating. Definitions for the difficult terms were retrieved from Wikipedia.
UAms - Bakker, Yüksel, and Kamps, 2024 submitted three runs. They employed an idf-based term weighting to identify the rarest terms for Task 2.1. For Task 2.3, they developed a method to rank definitions or explanations for given sentence-term pairs by examining the textual similarity of the provided sentences.
UBO - Vendeville, Ermakova, and Loor, 2024 submitted one run, using the Small Language Model, Phi3 mini, without fine-tuning, employing a one-shot prompt approach.
UNIPD - Di Nunzio, Gallina, and Vezzani, 2024 submitted three runs, focusing on identifying and explaining difficult content using Large Language Models (LLMs) to enhance text simplification. They iteratively experimented with various prompting strategies to optimize model performance for this task.

TABLE 2.13: Summary of team approaches and models used in the task.

- the BLEU score computed for bigrams (ngrams from $n = 1$ to $n = 4$).

For recall, precision, and F1, both “overall” scores calculated by summing all data for all retrieved and predicted terms for all sentences, and “average” scores calculated by averaging the recall and precision of predicted terms per sentence. In Table 2.14, the results are for all terms; in Table 2.15, we show the results for difficult terms only; in Table 2.16, we present the results for BLEU scores.

RunId	recall overall	precision overall	f1 overall	recall average	precision average	f1 average
AIIRLab_Task2.2_LLaMA	0.301	0.525	0.383	0.297	0.693	0.383
AIIRLab_Task2.2_LLaMAFT	0.008	0.989	0.016	0.005	0.217	0.016
AIIRLab_Task2.2_Mistral	0.469	0.559	0.510	0.436	0.494	0.510
Dajana&Kathy_SimpleText_Task2.2_LLAMA2_13B_CHAT	0.007	0.585	0.015	0.011	0.591	0.015
FRANE_AND_ANDREA_SimpleText_Task2.2_LLAMA2	0.005	0.645	0.010	0.007	0.458	0.010
team1_Petra_and_Regina_Task2_ST	0.003	0.500	0.005	0.004	0.500	0.005
Sharingans_Task2.2_GPT	0.485	0.428	0.455	0.524	0.650	0.455
SINAI_task_2_PRM_ZS_TASK2_V1	0.090	0.739	0.160	0.086	0.671	0.160
SINAI_task_2_PRM_ZS_TASK2_V2	0.167	0.672	0.268	0.161	0.499	0.268
SINAI_task_2_PRM_ZS_TASK2_V3	0.101	0.790	0.180	0.103	0.442	0.180
Tomislav&Rowan_Task2.2_LLAMA2_13B_CHAT_1	0.005	0.613	0.010	0.008	0.519	0.010
Tomislav&Rowan_Task2.2_LLAMA2_13B_CHAT	0.004	0.333	0.009	0.004	0.667	0.009
UAms_Task2-1_RareIDF	0.082	0.959	0.152	0.093	0.326	0.152
UboNLP_Task2.1_phi3-oneshot	0.582	0.527	0.553	0.555	0.563	0.553
unipd_t21t22_chatgpt	0.116	0.562	0.192	0.138	0.703	0.192
unipd_t21t22_chatgpt_mod1	0.227	0.398	0.289	0.234	0.764	0.289
unipd_t21t22_chatgpt_mod2	0.331	0.338	0.334	0.311	0.780	0.334

TABLE 2.14: Results for CLEF 2024 SimpleText Task 2 for all terms.

Analysis

The results indicate that the use of large language models for term extraction, difficulty assessment, and definition generation is still in its early stages, presenting new opportunities in Automatic Term Extraction. While recall and precision values are promising, they remain suboptimal compared to state-of-the-art models, especially as this is the first application of this new dataset by participants. Key findings include the stability of submitted runs in recall-precision performance across all terms and difficult ones, with participant models achieving

RunId	recall overall (difficult)	precision overall (difficult)	f1 overall (difficult)	recall average (difficult)	precision average (difficult)	f1 average (difficult)
AIIRLab_Task2.2_LLaMA	0.299	0.681	0.415	0.307	0.950	0.465
AIIRLab_Task2.2_LLaMAFT	0.006	1.000	0.012	0.007	1.000	0.014
AIIRLab_Task2.2_Mistral	0.212	0.485	0.295	0.199	0.892	0.326
Dajana&Kathy_SimpleText_Task2.2_LLAMA2_13B_CHAT	0.000	0.000	0.000	0.000	0.989	0.000
FRANE_AND_ANDREA_SimpleText_Task2.2_LLAMA2_13B_CHAT	0.006	0.364	0.012	0.010	0.981	0.020
team1_Petra_and_Regina_Task2_ST	0.000	0.000	0.000	0.000	0.995	0.000
Sharingans_Task2.2_GPT	0.565	0.587	0.576	0.583	0.854	0.693
SINAI_task_2_PRM_ZS_TASK2_V1	0.105	0.538	0.176	0.092	0.935	0.167
SINAI_task_2_PRM_ZS_TASK2_V2	0.149	0.806	0.251	0.134	0.978	0.236
SINAI_task_2_PRM_ZS_TASK2_V3	0.053	0.857	0.101	0.047	0.995	0.090
Tomislav&Rowan_Task2.2_LLAMA2_13B_CHAT_1	0.000	0.000	0.000	0.000	1.000	0.000
Tomislav&Rowan_Task2.2_LLAMA2_13B_CHAT	0.000	0.000	0.000	0.000	1.000	0.000
UAms_Task2-1_RareIDF	0.025	0.091	0.040	0.034	0.780	0.066
UboNLP_Task2.1_phi3-oneshot	0.351	0.387	0.368	0.332	0.737	0.457
unipd_t21t22_chatgpt	0.077	0.612	0.137	0.087	0.979	0.160
unipd_t21t22_chatgpt_mod1	0.226	0.591	0.327	0.234	0.979	0.378
unipd_t21t22_chatgpt_mod2	0.385	0.682	0.492	0.324	0.986	0.488

TABLE 2.15: Results for CLEF 2024 SimpleText Task 2 for difficult terms.

Run ID	BLEU (n1)	BLEU (n2)	BLEU (n3)	BLEU (n4)
AIIRLab_Task2.2_LLaMA	0.286	0.150	0.047	0.018
AIIRLab_Task2.2_LLaMAFT	0.240	0.117	0.000	0.000
AIIRLab_Task2.2_Mistral	0.259	0.133	0.041	0.014
Dajana&Kathy_SimpleText_Task2.2_LLAMA2_13B_CHAT	0.000	0.000	0.000	0.000
FRANE_AND_ANDREA_SimpleText_Task2.2_LLAMA2_13B_CHAT	0.000	0.000	0.000	0.000
team1_Petra_and_Regina_Task2_ST	0.000	0.000	0.000	0.000
Sharingans_Task2.2_GPT	0.227	0.106	0.031	0.016
SINAI_task_2_PRM_ZS_TASK2_V1	0.252	0.157	0.082	0.060
SINAI_task_2_PRM_ZS_TASK2_V2	0.276	0.159	0.067	0.049
SINAI_task_2_PRM_ZS_TASK2_V3	0.216	0.112	0.039	0.025
Tomislav&Rowan_Task2.2_LLAMA2_13B_CHAT_1	0.000	0.000	0.000	0.000
Tomislav&Rowan_Task2.2_LLAMA2_13B_CHAT	0.000	0.000	0.000	0.000
UAms_Task2-1_RareIDF	0.000	0.000	0.000	0.000
UboNLP_Task2.1_phi3-oneshot	0.001	0.000	0.000	0.000
unipd_t21t22_chatgpt	0.309	0.185	0.089	0.049
unipd_t21t22_chatgpt_mod1	0.311	0.181	0.082	0.045
unipd_t21t22_chatgpt_mod2	0.030	0.007	0.003	0.000

TABLE 2.16: Results for CLEF 2024 SimpleText Task 2 for the quality of definitions

precision above 0.50 and average recall between 0.3 and 0.5. Notably, the highest average precision (0.7604) was achieved through manual cleaning of ChatGPT's output. When focusing on difficult terms, recall decreases while precision increases, indicating fewer terms detected but with higher accuracy. Additionally, BLEU scores for generated definitions show stability, ranging from 0.2 to 0.3 for $n = 1$ and 0.1 to 0.2 for $n = 2$ across various recall values.

Chapter III

Methodological Framework

3.1 Research Approach and Design

3.1.1 Research Approach

In this work, a *quantitative* approach with *Correlational* scope is applied due to the nature of the research, and the stages of this process have been carefully followed.

(Hernández, Fernández, and Baptista, [2014](#)) defines:

The *quantitative approach* as a series of processes, it is sequential and evident. Each step precedes the next and we cannot “jump” or avoid steps. The order is strict, although of course, we can redefine some stages. It starts from an idea that is narrowed down and, once defined, the objectives and research questions are derived, the literature is reviewed and a framework or theoretical perspective is established. From the questions, hypotheses are formulated and variables are determined; a plan is made to test them (design); variables are measured in a specific context; the measurements obtained are analyzed using statistical methods and a series of conclusions are drawn in relation to the hypothesis.

3.1.2 Research Design

Arias, 2012, as cited in Armadas, 2018

Research designs are classified into documentary, field and experimental.

Considering the information provided by Armadas, 2018, the research design is mainly based on an *experimental design*. However, we have complemented this methodology with exhaustive *documentary research*. The latter has allowed us to explore the current state of the art in the field of study, with the aim of collecting relevant data that enriches our analysis. Specifically, we have delved into previous research related to our area of interest, focusing on the identification and understanding of complex words. This in-depth analysis will provide us with a solid basis and advance knowledge within this specific area.

3.2 Research Procedure

We have relied on the Natural Language Processing (NLP) timeline proposed by Kochmar, 2022 as a reference framework to represent the sequence of systematic and organized activities carried out in our research, guaranteeing the obtaining of reliable results, as well as as the formulation of valid conclusions. The timeline provides insight into the increasing level of sophistication of Natural Language Processing (NLP) algorithms, and in our study we have explored the three stages mentioned in the timeline. For a visual representation, see Figure 3.1.

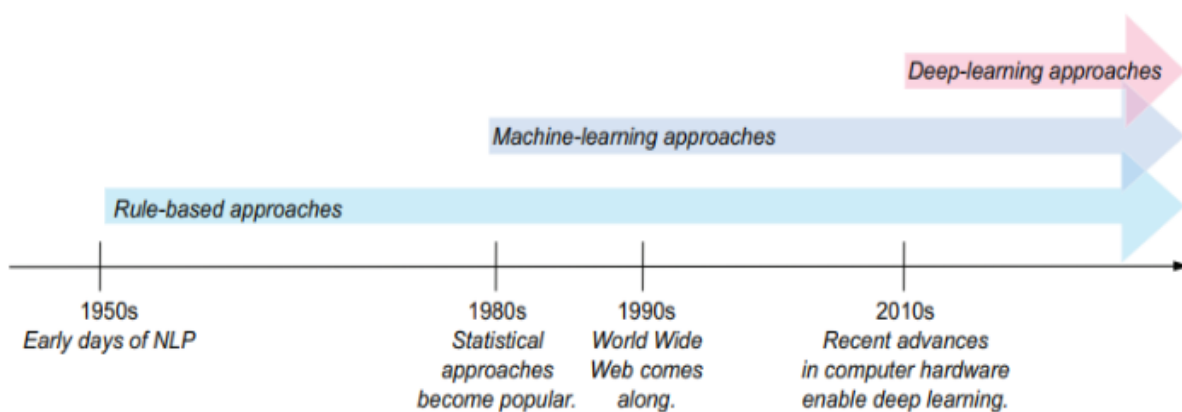


FIGURE 3.1: NLP timeline showing three different types of approaches by Kochmar, 2022.

Below we describe the stages that the project will follow to ensure the achievement of its objectives. These stages have been carefully designed to guide the research and development

process, facilitating the achievement of its results at each stage. We aim to maximize the efficiency and impact of the project, ensuring significant contributions to the field.

1. *SOTA - State Of The Art in Complex Word Identification - CWI and Lexical Complexity Prediction - LCP.*
2. *Creation of the corpora - Construction of new corpora and resources.*
3. *LF - Identification of relevant linguistic features in order to improve the performance of the algorithms through the use of the extracted linguistic features, and measure the impact of these (of the extracted linguistic features) in comparison with models based on deep learning (Gooding and Kochmar, 2018; Gooding and Kochmar, 2019), (North, Zampieri, and Shardlow, 2023a), (Aziz et al., 2023).*
4. *Rule-based approaches - The design of evaluation systems focused on evaluating the complexity of words according to various criteria such as the frequency of words, which was based on CREA¹ (Saggion et al., 2015a).*
5. *Evaluation of different models and combinations of linguistic features for CWI by evaluating machine and deep learning algorithms (Abramov, Ivanov, and Solovyev, 2023), (North et al., 2024).*
 - (a) *Machine-learning approaches - Research and design of a solution and its evaluation by executing classic supervised learning algorithms on data sets in Spanish and English.*
 - (b) *Deep-learning approaches - Research and design of a solution and its evaluation by executing the models based on the Transformer architecture in the Spanish and English data sets.*

The details of each stage are set out below:

3.2.1 SOTA - State Of The Art in Complex Word Identification - CWI and Lexical Complexity Prediction - LCP

This part of the research has been of vital importance in the exploration and analysis of the relevant documentary sources necessary to support the theories and formulas that underlay the study of the complexity of words in the Spanish language and its application from

¹CREA - Reference Corpus of Current Spanish - Frequency list <https://corpus.rae.es/lfrecuencias.html>

various approaches such as that based on word frequency, machine learning and deep learning approaches. This study has allowed us to reflect on the problem and identify the frontiers from where we explored our approaches to go beyond and contribute to further knowledge on the LCP task.

3.2.2 Creation Of The Corpora - The Creation Of Resources

These resources constitute the fundamental basis for carrying out tasks in our study of the 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 (Bott and Saggion, 2012).

The corpora developed as the initial basis of our project will be subject to a detailed compilation process, which will include the collection, organization and structuring of large volumes of texts relevant to each specific domain. This process will involve a rigorous selection of sources, data normalization and detailed annotation of key linguistic features. The proper structuring of these corpora will be essential to guarantee the quality and representativeness of the data, ensuring its optimal use in the subsequent phases of training and evaluation of Natural Language Processing (NLP) models. These corpora constitute the essential pillar for the implementation and execution of algorithms designed for specific tasks, playing a crucial role in the accuracy and effectiveness of the predictions and analysis carried out in the experiments (The details in the in chapter IV, and in annex A you will find the information corresponding to the scientific events).

3.2.3 Identification of Relevant Linguistic Features

After reviewing research on lexical complexity and the applications of different technologies, linguistic features have played a key role in the prediction of complex words. Therefore, these features will be part of the design and compilation of the Spanish data sets, using the resources created for this purpose. Likewise, they will be integrated into the English data sets, which will allow the development of more robust and multilingual proposals. These proposals will be sent to the International Semantic Evaluation Workshops, with the aim of contributing to the improvement of evaluation standards, validating our methodologies in a competitive context

and sharing our advances with the scientific community to promote international collaboration and encourage new research in the field of lexical complexity prediction.

3.2.4 Baseline With CREA

We will begin by exploring the simplest solutions. This approach is based on the frequency of use of a word according to Saggion et al., 2015a. A general resource such as CREA serves as a basis for identifying whether the word is complex or not. At this stage, an evaluation is proposed through the creation of systems that allow the detection of complex words by applying the calculation of low-frequency words, these being those words whose frequency is less than 1,000 (Rebollo, 2008).

3.2.5 Machine Learning Approaches

At this point, we propose to carry out an evaluation of solutions based on classical machine learning algorithms, performing an analysis of different linguistic features and their combination with pre-trained (non-fine-tuned) models, such as some BERT-based models. In these works we will apply these approaches to both the English and Spanish datasets.

3.2.6 Deep Learning Approaches - The Implementation Of The Models Based On The Transformer Architecture

In this phase, we will conduct an in-depth exploration and application of Transformer-based models such as BERT, RoBERTa, the GPT language generator model and their most recent versions. Our recent findings have been obtained through the implementation of GPT-4-turbo and GPT-4o. The analysis and evaluation of the performance of these models will be performed on both English and Spanish datasets, with the aim of comparing their performance, paying special attention to the Spanish approach, which is the core of our study.

3.3 Methodology

According to what was established by Abreu, 2014:

“The research method describes in good detail how the research has been carried out. This allows explaining the property of the methods used and the validity of the results, including the relevant information to understand and demonstrate the ability to replicate the research results.”.

Two fundamental aspects have been taken into account in the development of the research:

1. Creation of resources, and
2. The proposal of solutions for the detection of difficult words.

Regarding the “*creation of resources*”, the detail about the creation and evaluation of each of the corpora is found in chapter [IV](#).

About “*The proposal of solutions to the detection of difficult words*”, for its possible solution, we have explored the two strategies detailed below:

1. **Difficult Word Extraction**, where difficult words are located in the text, that is, the tagging of complex words that the reader identifies in the texts. During this research, the tagging process will be developed as a fundamental phase in the development of the corpora. The details of each of the corpora are presented in Chapter [IV](#).

Once a complex word has been identified, it is necessary to subject it to an assessment process. This process is divided into two types, depending on the situation or context: categorical or numerical, that is:

- **Categorical Assessment**

Through this assessment, the complex word can be labeled, for example, as “difficult” or “not difficult,” or assigned a level on a complexity scale. This assessment can take two approaches: the *binary* or the *multiclass* assessment.

- (a) **Binary Categorical Assessment**, where the word identified as complex can be assigned a label, such as: *difficult* or *not difficult*.

- (b) **Multiclass Categorical Assessment**, which can present a set of possibilities that can be assigned to the word identified as complex, such as: *very easy, easy, neutral, difficult or very difficult* (Shardlow, Cooper, and Zampieri, 2020b).
 - **Numerical Assessment**, is closely related to the assignment of a numerical value based on the level of complexity of the word, as described in more detail below.
2. **Evaluating Word Difficulty** involves providing a context and a word, and then rating the complexity of the latter, so we would face a *regression problem*. In this approach, a numerical value is assigned that represents an estimate of the difficulty level of the word identified as complex.

In order to carry out this type of strategy, by building a regression system, we first have to model the problem, as follows:

- (a) We take the texts and convert them into numerical vectors that are subsequently subjected to a conventional classifier. There are various methodologies to carry out this coding:
 - Taking a set of linguistic features that are part of our data sets and that are described extensively in the 3.4.2 section, and passing them to the classifier.
 - Another option is to use embeddings, which consists of encoding the text using a pre-trained language model and then entering it into the classifier.
 - Another alternative is to combine both approaches, that is, linguistic features with embeddings, to obtain a more complete representation.
- (b) Furthermore, it is possible to address the issue through fine-tuning the pre-trained model, which involves not only applying the previous approach, but also fine-tuning the encoder itself. This is relevant because the encoder is a neural network that can be tailored specifically for the problem at hand. In this way, we achieve a coding that is more adjusted to the problem domain.

- (c) Using unsupervised learning, the unsupervised learning approach we explore is based on the use of large language models and their “prompting”.

In this approach, there is no need to model the input, that is, there is no need to provide specific examples of difficult words (annotations, extractions or evaluations). Instead, a text or “token” is posed as a solution and the language model is expected to understand the problem and generate answers automatically. The objective is to ensure that the language model understands the problem and generates responses in an unsupervised manner through prompting.

3.4 Variables

According to what was established by León and González, [2020](#):

“The research variables are factors that have the characteristic of being able to be measured and manipulated. The variables constitute a concept of vital importance within a project since they are the specific statements that make up a hypothesis”.

Amiel Pérez, [2007](#) manifests:

“I consider the variable as the fundamental unit of the Scientific Research Method. Indeed, from it we formulate our hypotheses and to demonstrate it we design the experiments using operationalized variables; but we can detect them from observation, the formulation of the problem or by specifying the theoretical framework”.

“A variable is anything that has its own characteristics – that distinguishes it from the rest – that is susceptible to change or modification and we can study, control or measure it in research”.

In effect, based on the other study variables, we have formulated the hypotheses presented in Chapter I and section [1.2](#).

Following the above, other variables are:

- *Complexity*, which can take binary, categorical or numerical values, although categorical values can be represented as discrete integer numerical values, as they are a scale.

- The variables that represent the texts:
 - Embeddings.
 - Linguistic Features (some of them are complexity metrics that are seen in the next point).
- Complexity Metrics.
- Evaluation Metrics.

Regarding *Lexical Complexity Metrics*, we have implemented in our research the seven lexical complexity metrics, which we describe in detail in Chapter II section 2.3.1. These metrics allowed us to evaluate the systems designed to predict the lexical complexity of Spanish texts at various levels, that is, sentence level and word level.

The metrics are:

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

Regarding the *Evaluation metrics* for the prediction of lexical complexity, we will use the common error metrics in regression tasks, which are presented in detail in Chapter II section 2.5.2. By running these measures, we will be able to analyze and compare the performance of the applied models not only on Spanish data sets but also on the English language, thus managing to select the best solution for our object of study. The selected evaluation metrics are the following:

- Mean Absolute Error - *MAE*.
- Mean Square Error - *MSE*.
- Root Mean square Error - *RMSE*.

- Pearson’s correlation coefficient - *Pearson*.
- Coefficient of determination - R^2 .
- Precision, Recall y F-score - for the binary and multiclass cases.

3.4.1 Evaluation Metrics

North, Zampieri, and Shardlow, 2023a manifest that, the evaluation of the performance of complexity prediction systems is carried out using various evaluation indicators, which vary depending on the task performed. Previous CWI systems (Specia, Jauhar, and Mihalcea, 2012), (Paetzold and Specia, 2016a), (Yimam et al., 2018a) applied binary classification which is the most common task; regression performed by LCP systems (Shardlow et al., 2021a). The details can be found in the Chapter II in section 2.5.1 and section 2.5.2.

3.4.2 Features Sets

Next, we will describe how the features of our model encode the semantic properties of the sentence, in addition to the linguistic and statistical properties. These features are grouped into three different sets: *linguistic*, *statistical* and *semantic*, following an approach similar to that of the research previously carried out by Liebeskind, Elkayam, and Liebeskind, 2021.

- **Linguistic and Statistical Features**

In accordance with Liebeskind, Elkayam, and Liebeskind, 2021, the properties contemplated in the *linguistic features* include two families of properties: morphological and syntactical. The *linguistic features* are computed as classic features in computational, and the *semantic features* are those that result from the word-level and sentence-level *embeddings* described in section 3.4.2.

Linguistic Features have been widely used in research related to the prediction of lexical complexity, as discussed in the work by: Liebeskind, Elkayam, and Liebeskind, 2021; Desai et al., 2021a; Vettigli and Sorgente, 2021; Gooding and Kochmar, 2018; Paetzold and Specia, 2016b; Ronzano, Anke, Saggion, et al., 2016.

We consider the *en_core_news_sm* model of the Spacy POS statistical tagger² to extract the set of grammatical tagging features made up of eight tags from the universal POS tags:

²<https://spacy.io/>

PROP, AUX, VERB, ADP, NOUN, NN, SYM and NUM, which will allow us to calculate some morphological aspects of the text. The details are set out below:

Morphological Features:

1. *Part of Speech (POS)*: the Part of Speech category.
2. *PROP*: Number of pronouns within the sentence.
3. *AUX*: Number of auxiliaries within the sentence.
4. *VERB*: Number of verbs within the sentence.
5. *ADP*: Number of adverbs within the sentence.
6. *NOUN*: Number of nouns within the sentence.
7. *NN*: Number of nouns, singular or massive.
8. *SYM*: Number of symbols within the sentence.
9. *NUM*: Number of numbers within the sentence.

Statistical features:

Liebeskind, Elkayam, and Liebeskind, 2021 defined the features based on frequency:

10. *Word Length*: the number of characters of the token. The length of the word is calculated as the number of its characters. It is often the case that longer words are more difficult to process and can therefore be considered Shallow, 2013; Paetzold and Specia, 2016b; Mosquera, 2021a. We also highlight some linguistic features applied in research work in the identification of complex words. Below is the detail.
11. *Absolute Frequency*: the absolute frequency. The frequency of words is a measure that serves as an indicator of lexical complexity. If, in common parlance, a word occurs frequently, it is more likely to be recognized (Rayner and Duffy, 1986; Shallow, Cooper, and Zampieri, 2020b).
12. *Relative Frequency*: the relative frequency of the target word.
13. *Number of Syllables*: the number of syllables (Shallow, 2013; Ronzano, Anke, Saggion, et al., 2016; Shallow, Cooper, and Zampieri, 2020b; Paetzold and Specia, 2016b).
14. *Number of Words in the Sentence*: number of words in the sentence (Shallow, 2013; Ronzano, Anke, Saggion, et al., 2016).

Others features based on the work proposed by (Ronzano, Anke, Saggion, et al., 2016):

15. *Relative Frequency of the Previous Token*: the relative frequency of the word before the token.
16. *Relative Frequency of the Word After the Token*: the relative frequency of the word after the token.
17. *Length of the Previous Word*: the number of characters in the word before the token.
18. *Length of the Following Word*: the number of characters in the word after the token.
19. *Target Word Position*: the position of the target word in the sentence (Shardlow, 2013; Ronzano, Anke, Saggion, et al., 2016).
20. *Lexical Diversity - MTDL*: the lexical diversity of the target word in the sentence (Shiroyama, 2022).

WordNet Features were also considered for each target word, as in the work carried out by Gooding and Kochmar, 2018; Mosquera, 2021a:

21. *Number of Synonyms*.
22. *Number of Hyponyms*.
23. *Number of Hyperonyms*.

- **Semantic Features**

Regarding semantic features, Liebeskind, Elkayam, and Liebeskind, 2021 in their work of the Lexical Complexity prediction, manifest:

We will represent the meaning of the surrounding context of the target word by vectors in the same semantic space. We use the BERT semantic space. BERT contextualization vectors are used to represent the semantic meaning of the sentence by averaging the BERT vectors of seven words, the target word and three words before and after it. Therefore, our semantic representation adds 768 features (the BERT output layer size).

- **Word Embeddings**

Word embeddings are vector representations of words that are beneficial for various tasks related to NLP. These *embeddings* are an approach to distributional semantics that represents

words as vectors of real numbers. In short, words are converted into vectors containing actual numerical values, with each value reflecting a particular facet of the word's meaning. This means that words that have a similar meaning present similar vectors. In simple terms, each dimension of these vectors represents an aspect of meaning, and the numerical value in each dimension reflects the closeness of the word's relationship to that aspect. In conclusion, *word embeddings* capture the meaning of words, and translate them into an encoding that can be used as input for all types of neural networks. This has caused their use to spread rapidly and they are currently a fundamental piece in the architecture of all types of models that perform NLP tasks (García Ferrero, 2018).

In several tasks of our research, we will combine all the features corresponding to the 768-dimensional vector of the initial [CLS] token as *sentence embeddings*, the embeddings corresponding to the target token, and the linguistics features.

To further illustrate the above, Figure 3.2 shows the approach we have defined to apply in our study, based on *combining encodings* of pre-trained language models with different linguistic features to produce a final vector that is fed into some classical machine learning algorithms. As can be seen, from a given text, two feature vectors are generated: the embeddings of the transformation model and the linguistic features computed from the text (normalized with a Z-score transformation). These vectors are then combined and serve as input to the classification algorithm.

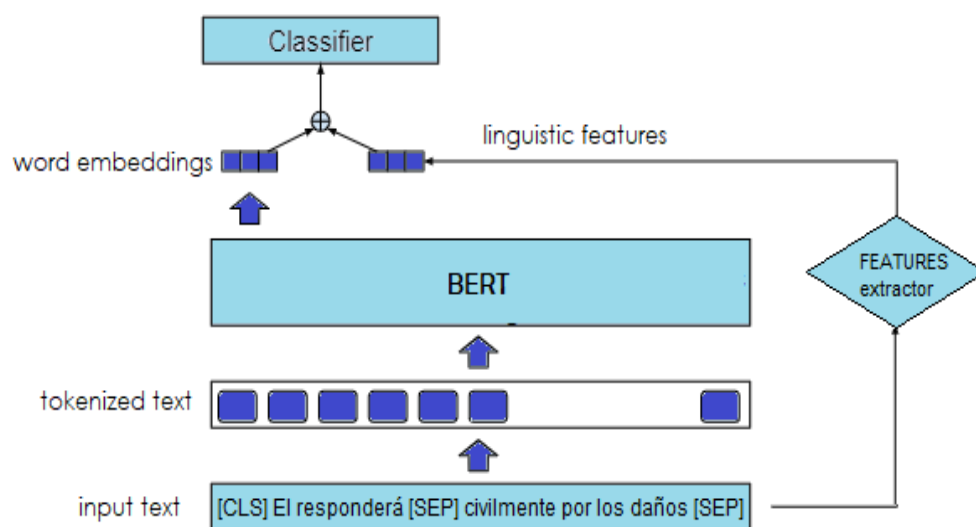


FIGURE 3.2: Combining encodings of pre-trained language models with different linguistic features.

3.5 Available Resources

We have used key resources, such as information sources, databases and other relevant materials to achieve the expected results in the research process. These resources have been essential to obtain information that supports this study and serves as a basis for the collection, analysis of data and generation of knowledge. Below are the details:

- **Documentary Information**

Armadas, [2018](#) define a la *Documentary research* as:

Documentary research is one that seeks to collect information with the objective of state the theories that support the study of phenomena and processes. This type of information is extracted through an exploratory investigative level. In order to familiarize ourselves with the existing knowledge within the field to which the object of research study belongs.

In accordance with Tancara, [1993](#), *documentary research* is a set of methods and techniques for searching, processing and storing the information contained in documents in the first place, and the systematic, coherent and sufficiently justified presentation of new information in a scientific document in the second place. In this way, documentary research should not be understood or exhausted as a simple search for documents on a subject.

For the systematic search and selection of information we will use one of the scientific dissemination engines such as Google Scholar, which is focused and specialized in the search for scientific-academic content for the exploration and filtering of scientific material related to the object of study in the field of Lexical Simplification. This material included articles, doctoral theses and publications, from indexed scientific databases such as SCOPUS, WEB of Science, among others (Ortíz et al., [2024](#)).

- **Primary Sources of Information**

In accordance with Armadas, [2018](#):

All documents provide us with some information, but it is not always relevant so it is necessary to prioritize them; Primary documents are those original documents that provide first-hand data. These documents are those that record or corroborate immediate

knowledge of the investigation; such as: Books, Periodical publications, Reports from international organizations, Scientific and Technical reports, Conference and symposium proceedings, Theses, monographs, Results of application of instruments among others.

As indicated by Armadas, 2018, below, we detail the primary sources that we have used to support the execution of this project:

*** Scientific Competitions in computational linguistics**

The contribution of information sources from digital platforms of scientific conferences such as conferences and workshops in computational linguistics has been very relevant, as it has provided us with essential and necessary data not only for the development of our research, but also so that our results can be reviewed and published at scientific events within the area. All of this has allowed us to generate our own results that we intend to constitute a contribution to the scientific community in the area of lexical simplification in the academic field in Spanish. Below are several scientific events in brief detail:

ACM³ - Association for Computing Machinery

Provides scientific information from academic publications, conferences and activities related to computing.

ACL⁴ - Association for Computing Machinery

Is a leading international organization in the field of computational linguistics, its main objective being to advance in the area of natural language processing by promoting research in technologies related to human language and automated communication.

CLEF⁵ - Conference and Labs of the Evaluation Forum

The CLEF Initiative promotes research, innovation and development of information access systems with an emphasis on multilingual and multimodal information with various levels of structure. CLEF promotes research and development workshops to discuss and

³ACM - Association for Computing Machinery. Available in <https://dl.acm.org/>

⁴ACL - Association for Computing Machinery. Available in <https://www.aclweb.org/portal/>

⁵CLEF - Conference and Labs of the Evaluation Forum. Available in <https://www.clef-initiative.eu/>

pilot innovative evaluation activities.

EMNLP⁶ - Conference on Empirical Methods in Natural Language Processing.

The EMNLP 2022 tutorials session provides an in depth coverage of a variety of topics reflecting recent advances in Natural Language Processing methods and applications, presented by experts from academia and ranging from introductory to cutting-edge.

IberLEF⁷ - Iberian Languages Evaluation Forum

IberLEF is a shared evaluation campaign of Natural Language Processing systems in Spanish and other Iberian languages that has been organized since 2019, and is held as part of the annual conference of the Spanish Society for Natural Language Processing (SEPLN). It aims to promote research in text processing, understanding and generation tasks in at least one of the following Iberian languages: Spanish, Portuguese, Catalan, Basque or Galician.

RanLP⁸ - Recent Advances in Natural Language Processing

The international conference RANLP is offering an opportunity to promote the latest advances in the Natural Language Processing (NLP) field, including the employment of latest deep learning technology amongst a host of other solutions.

SemEval⁹ - International Workshop on Semantic Evaluation

Is a series of international Natural Language Processing (NLP) research workshops whose mission is to advance the current state of the art in semantic analysis and help create high-quality annotated datasets on a variety of increasingly challenging problems in the semantics of natural language.

TSAR¹⁰ - Text Simplification, Accessibility, and Readability

⁶EMNLP - Conference on Empirical Methods in Natural Language Processing. Available in <https://aclanthology.org/venues/emnlp/>

⁷IberLEF - Iberian Languages Evaluation Forum. Available in <https://sites.google.com/view/iberlef-2019/>

⁸RANLP - Recent Advances in Natural Language Processing. Available IN <https://aclanthology.org/venues/ranlp/>

⁹SemEval - International Workshop on Semantic Evaluation. Available in <https://semeval.github.io/>

¹⁰TSAR - Text Simplification, Accessibility, and Readability. Available in [assds](#)

The TSAR workshop aims to foster collaboration among all parties interested in making information more accessible to all people discussing recent trends and developments in the area of automatic text simplification, text accessibility, automatic readability assessment, language resources and evaluation for text simplification, etc.

** Books*

For the development of this study, several books concerning the field of Artificial Intelligence, language models based on Transformers architecture, also books about GPT-3, Natural Language Processing, languages such as Python, Deep Learning and Machine Learning have been reviewed, among other related things.

This source of information has been a fundamental resource in the research, since it provided a solid basis on the topics that concern this study offering deep and detailed information.

• Tools and Resources

The variety of NLP-related tools and resources that Hugging Face offers is very wide such as: tokenizers, functions, lines of code, evaluation metrics, and more.

** Free Resources: Hugging Face Platform¹¹*

Several tools were used, provided by the open source development platform *Hugging Face* which specializes in Natural Language Processing and Machine Learning models. Thanks to the advantages offered by this platform, users can access advanced pre-trained NLP models and also have access to a wide range of natural language processing tools. Some of the most notable resources from this platform used in the various experiments are:

** Python Natural Language Toolkit - NLTK*

NLTK is the Natural Language Toolkit, a comprehensive Python library for natural language processing and text analytics. Originally designed for teaching, it has been adopted in the industry for research and development due to its usefulness and breadth of coverage

¹¹Hugging Face - Available <https://huggingface.co/>

(Perkins, 2010).

** Machine Learning libraries - spaCy¹²*

A free, open-source library accurate for advanced Natural Language Processing (NLP) via Python. It comprehends and delineate s the text (either small or large) by processing the same. Moreover, spacy provides wide range of in-built features which makes it an efficient tool for Text Processing and Language Modelling (Amade et al., 2024), and was designed from day one to be used in real products.

** Scikit-learn¹³*

Scikit-learn, one of the most frequently used libraries in machine learning (Zhang and Chen, 2024). It has a large amount of information available, among which the documentation stands out with its detailed guide, the tutorials, the examples and the notebooks provided are a valuable resource that has been very useful to us in knowing how to approach machine learning tasks in the Identification of complex words.

** Machine Learning libraries - Pandas*

The *pandas* library is the pretty much ubiquitous. It is built on the top of the *NumPy* library, which allow as to perform mathematical operations on arrays of single-type data of efficiently. *Pandas* expands this to *dataframes*, which can be thought of as tables of data (Molin, 2019).

** Machine Learning libraries - NumPy*

NumPy library is used to perform the statistical operations on the dataset (Raheja et al., 2022).

** Machine Learning libraries - Matplotlib plotting library*

Matplotlib is a Python package for 2D plotting that generates production-quality graphs. It support interactive and non-interactive plotting, and can save image and several output formats (PNG, ps, and others).It can use multiple window toolkit (GTK+, wxWidgets, Qt

¹²spaCy - Available in <https://spacy.io/>

¹³Scikit-learn - Available in <https://scikit-learn.org>

and so on) and it provides a wide variety of plot types (lines, bars, pie, charts, histograms, and many more) (Tosi, 2009)

** Transformers Library*

This library has been widely used during the development of experiments in text processing aplicando modelos basados en la arquitectura Transformers como BERT, RoBERTa y sus variantes XLM-RoBERTa, RoBERTa base, RoBERTa-large-BNE.

** Machine Learning library - PyTorch*

Deep learning frameworks have often focused on either usability or speed, but not both. PyTorch is a machine learning library that shows that these two goals are in fact compatible: it was designed from first principles to support an imperative and Pythonic programming style that supports code as a model, makes debugging easy and is consistent with other popular scientific computing libraries, while remaining efficient and supporting hardware accelerators such as GPUs (Paszke et al., 2019).

** OpenAI library*

OpenAI is an AI research and implementation company. Its overall mission is to harness the power of artificial general intelligence for the betterment of humanity, ensuring its benefits are accessible to all¹⁴.

- **Pre-Trained Models**

Hugging Face contains state-of-the-art language models and is freely accessible. The pre-trained natural language models based on the Transformer architecture that have been applied are: BERT, RoBERTa and their variants. Therefore Hugging face is a very good source to find models with great impact in academia and industry (García Subies, 2021).

In the experiments carried out, we applied the Fine-Tuning technique to refine the model. This action made it possible to adapt the neural network to the specific problem of the task to be solved, thus ensuring a tighter and more effective approach to solving the problem.

¹⁴OpenAI - Available on <https://openai.com/about>

- **Technological Infrastructure**

The Center for Advanced Studies of Information and Communication (CEATIC¹⁵) has contributed to the technological infrastructure by providing the computer equipment where the code was developed and the subsequent execution of the applications for the experiments during the research process. Its cluster for supercomputing tasks is well equipped with GPU units needed for training deep neural networks.

3.6 Information Collection And Analysis Techniques

According to what was established by Hernández Sampieri, Fernández Collado, Baptista Lucio, et al., 2018, *“Data collection techniques, as in quantitative research, can be multiple, these are: interviews, projective tests, closed questionnaires, group sessions, analysis of episodes, biographies, cases, audio or video recordings, records, file review, observation, etc.”*.

During the research process, several techniques were applied for the collection and analysis of information, which included:

Questionnaires

The design of questionnaires with closed questions is very useful to obtain precise and relevant information. In our case, we will apply it to students with the purpose of knowing their appreciation about the importance of incorporating new technologies that can improve their learning process, with special attention to accessible technologies aimed at supporting groups with learning difficulties. To analyze the responses to the questionnaires, it will be necessary to select a representative sample of students with active enrollment in the various academic offerings of the University of Guayaquil, which is our case study. The data collected will be processed and analyzed using statistical software.

Video Recordings

To analyze the responses to the questionnaires, it will be necessary to select a representative sample of students with active enrollment in the various academic offerings of the University of Guayaquil, which is our case study. The data collected will be processed and analyzed using

¹⁵CEATIC - Available on <https://www.ujaen.es/centros/ceatic/>

statistical software. The use of this data collection technique is essential for the development of the research, since it contributes significantly to the creation of essential primary sources in this process. The main objective is to build a corpus of academic texts that serves as a solid basis for the beginning of the research. In addition, the use of video recordings is contemplated as an additional source of data, considering the presence of online education. These videos will be collected and transcribed using specialized software, which will allow obtaining the corresponding transcripts of the classes taught virtually by the teachers.

Records

It is essential to keep a detailed record of the students enrolled in the different academic courses at the University of Guayaquil. This record is key to ensure a careful selection of participants, which is essential for the validity of the study. The record of actively enrolled students will allow us to identify those who best fit the research criteria. These criteria will include aspects such as demographic diversity, place of residence, and academic performance, ensuring that the selected sample adequately represents the student population. Precision in this process will ensure that the results obtained are generalizable and relevant to the academic context of the university and will better delve into the topic of identifying complex words in the academic domain in Spanish.

Chapter IV

Resources, Experiments and Results

4.1 Introduction

This chapter offers a broad exposition that covers everything from the conceptualization and development of the resources used to the implementation of the experiments, providing a comprehensive view of the methodology used and the results obtained; this is a fundamental component of our research, as it delves into the construction and application of the resources, as well as the evaluation of the experiments. The proposed structure in the timeline by Kochmar, [2022](#) has allowed us to identify the various tasks that our research encompasses, from the exploration of studies related to the identification of complex words to the implementation of deep learning models. To organize these activities effectively, we have adapted our schedule within the Natural Language Processing (NLP) timeline, incorporating the phases that encompass the initial activities of our project as shown in [Figure 4.1](#).

The results obtained throughout our research focused on the identification of complex words in the academic environment of the Spanish language are presented and were exposed at the Doctoral Symposium on Natural Language Processing from the ILENIA Project (PLN-DS-2023) held as part of the XXXIX edition of the International Conference of the Spanish Society for Natural Language Processing (SEPLN 2023) (Ortiz-Zambrano, [2023](#)). [Figure 4.2](#) shows a general diagram that summarizes the methodology used in the development of the research.

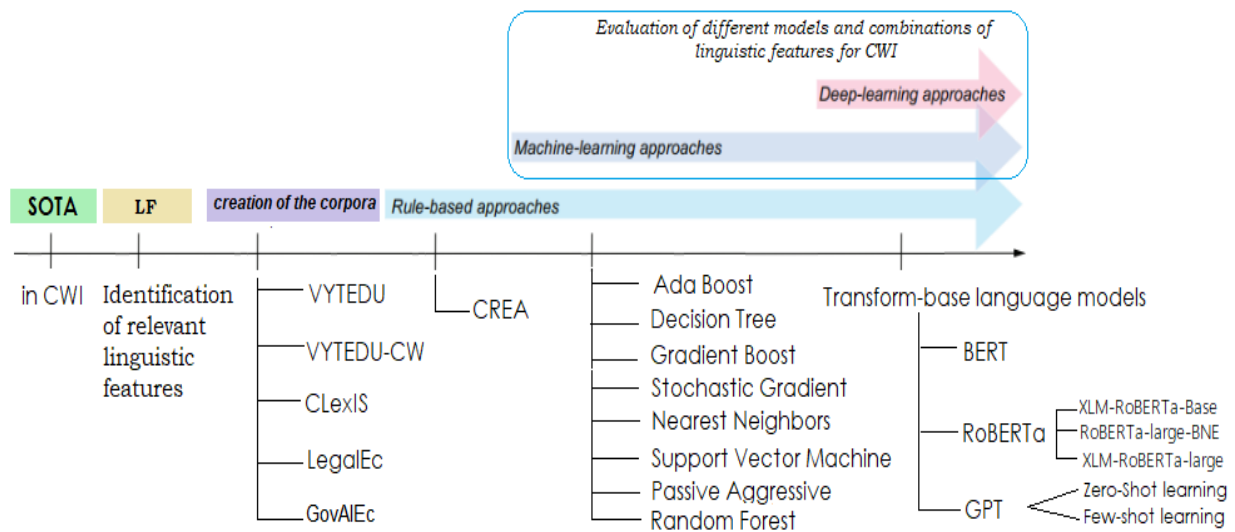


FIGURE 4.1: The tasks covered by our research based on the proposed structure in the timeline of Kochmar, 2022.

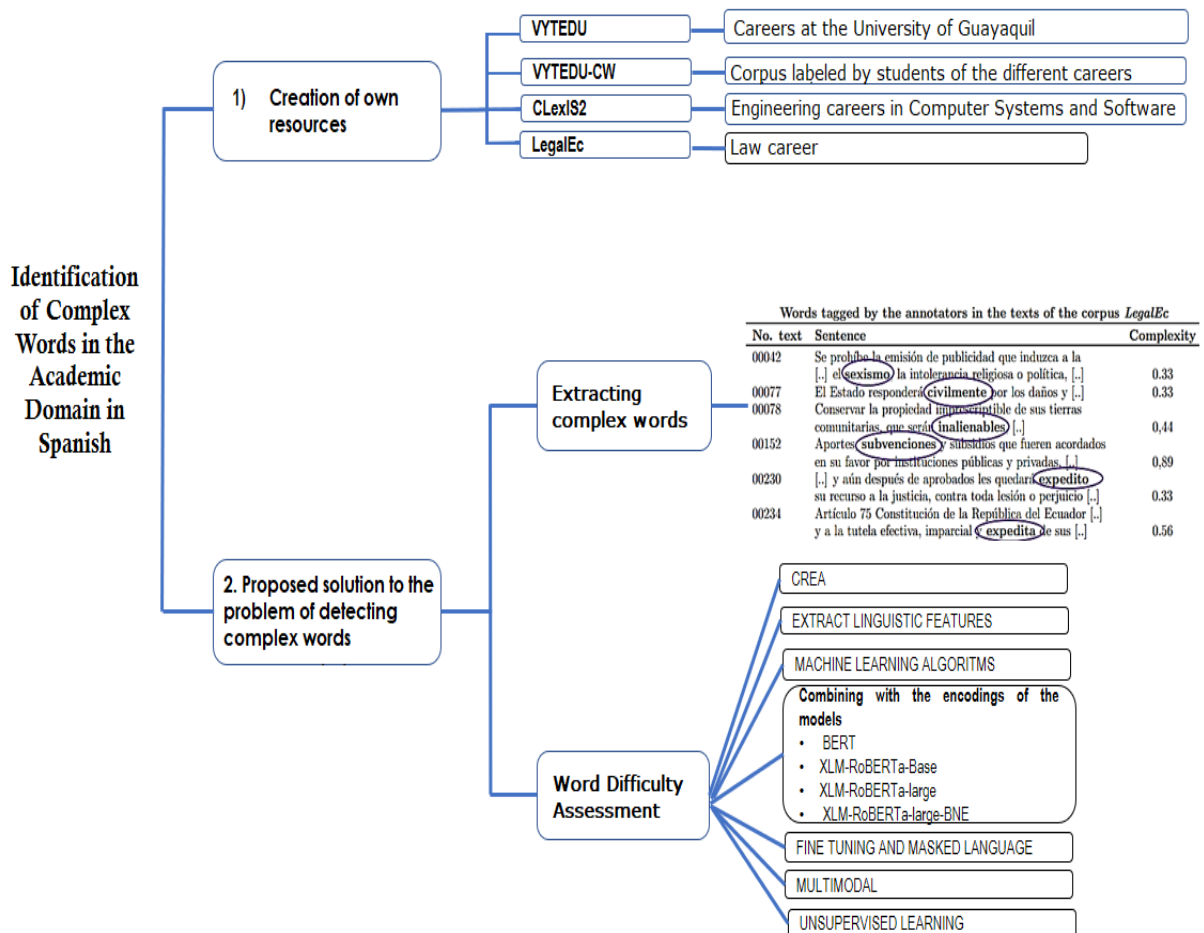


FIGURE 4.2: The methodological process carried out during the development of the research on the identification of complex words in the academic domain in Spanish.

The visual representation provided in Figure 4.3 offers an alternative perspective to understand the various phases that we have addressed during the development of this study. We started by creating our own resources, which has become the cornerstone of our development. Then, we have advanced from the application of simple techniques to the implementation of classic machine learning algorithms. The results have been disseminated in high-impact international scientific workshops, conferences, and magazines that promote all types of activities related to the study of Natural Language Processing (NLP) and Artificial Intelligence, such as:

1. **SEPLN 2017**¹ - Spanish Society for Natural Language Processing.

Natural Language Processing, Magazine No. 59, March 2017 - September 2017. Available on <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/issue/view/234>

Paper: VYTEDU: A corpus of videos and their transcriptions for research in the educational field (Ortiz-Zambrano and Montejo-Ráez, 2017). Available on <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5512>.

2. **ICCCE 2018**² : Computer and Communication Engineering.

First International Conference, ICCCE 2018, Guayaquil, Ecuador, October 25–27, 2018, Proceedings available on <https://link.springer.com/book/10.1007/978-3-030-12018-4>.

Paper: Reading Comprehension in University Texts: Lexical Complexity Metrics in corpus analysis in Spanish (Ortiz-Zambrano and Varela Tapia, 2018). Available on https://link.springer.com/chapter/10.1007/978-3-030-12018-4_9.

Published in the International Conference on Computer and Communication Engineering (pp. 111-123). Springer, Cham.

3. **ICAETT 2019**³: The International Conference on Advances in Emerging Trends and Technologies.

1st International Conference on Advances in Emerging Trends and Technologies (ICAETT 2019), held in Quito, Ecuador, on 29–31 May 2019.

¹SEPLN 2017 - Spanish Society for Natural Language Processing.
Available on <https://sepln2017.um.es/>

²ICCCE 2018 Conference: Computer and Communications Engineering pp. 111–123.
Available at <https://link.springer.com/book/10.1007/978-3-030-12018-4>

³ICAETT 2019 Conference: The International Conference on Advances in Emerging Trends and Technologies pp. 167-176.
Available at <https://link.springer.com/book/10.1007/978-3-030-32022-5>

Paper: VYTEDU-CW: Difficult Words as a Barrier in the Reading Comprehension of University Students (Ortiz-Zambrano et al., 2019). https://link.springer.com/chapter/10.1007/978-3-030-32022-5_16.

Springer Nature Switzerland AG 2020. ICAETT 2019, AISC 1066, pp. 167–176, 2020.

4. **IJASEIT**⁴: International Journal on Advanced Science, Engineering and Information Technology.

Paper: Barriers in Reading Comprehension of University Students: Analysis of the Complicated Words Annotated in the VYTEDU-CW Corpus (Ortiz-Zambrano and Montejo-Raéz, 2020). Available on <https://ijaseit.insightsociety.org/index.php/ijaseit/issue/view/65>. Vol. 10 No. 5 (2020).

5. The proposal of a task called ALexS 2020 for **IberLEF 2020**⁵.

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.

Paper: Overview of ALexS 2020: First Workshop on Lexical Analysis at SEPLN (Ortiz-Zambrano and Montejo-Ráezb, 2020) - Available on https://ceur-ws.org/Vol-2664/alexs_overview.pdf.

6. **SemEval-2021**⁶ - The 15th International Workshop on Semantic Evaluation.

Participation in Task 1: Lexical Complexity Prediction.

Paper: Complex words identification using word-level features for SemEval-2020 Task 1 (Zambrano and Montejo-Ráez, 2021). Available on <https://aclanthology.org/2021.semeval-1.11/>

⁴IJASEIT 2020 - International Journal on Advanced Science, Engineering and Information Technology is indexed in Elsevier Scopus database.

Available on <https://ijaseit.insightsociety.org/index.php/ijaseit>

⁵IberLEF 2020: Iberian Language Evaluation Forum.

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

⁶SemEval-2021 - The 15th International Workshop on Semantic Evaluation.

Available on <https://semeval.github.io/SemEval2021/>

item **RaNLP 2021**⁷ - Recent Advances in Natural Language Processing.

Deep Learning for Natural Language Processing Methods and Applications - Proceedings available on <https://ranlp.org/ranlp2021/proceedings-20Sep.pdf>.

Paper: CLeXIS²: A New Corpus for Complex Word Identification Research in Computing Studies (Zambrano and Montejo-Ráez, 2021). Available on <https://ranlp.org/ranlp2021/proceedings-20Sep.pdf>.

7. **SEPLN 2022**⁸ - Spanish Society for Natural Language Processing.

Natural Language Processing, Magazine No. 69, September 2022. Available on <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/issue/view/286>.

Paper: Transformers for Lexical Complexity Prediction in Spanish Language (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2022b) . Available on <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6438>.

8. **Electronics**⁹ - open access journal.

Paper: Combining Transformer Embeddings with Linguistic Features for Complex Word Identification (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2022a). Available on <https://www.mdpi.com/search?authors=ortiz+zambrano&journal=electronics>.

9. **Springer Journals** - Journal "Neural Computing & Applications".

Our study for the English and Spanish datasets have been detailed in our work titled "Deep Encodings vs. Linguistic Features in Lexical Complexity Prediction", which has been accepted for publication in Springer's journal Neural Computing and Applications.

10. **SEPLN 2023**¹⁰ - Spanish Society for Natural Language Processing.

Procesamiento del Lenguaje Natural, Revista n° 71, septiembre de 2023.

Available on <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/issue/view/292>.

⁷RaNLP 2021 - Recent Advances in Natural Language Processing.
Available on <https://ranlp.org/ranlp2021/start.php>

⁸SEPLN 2022 - Spanish Society for Natural Language Processing.
Available on <http://sepln2022.grupolys.org/>

⁹Electronics - open access journal.
Available on <https://www.mdpi.com/journal/electronics>

¹⁰SEPLN 2023 - Spanish Society for Natural Language Processing.
Available on <http://35.157.106.15/>

Paper: LegalEc: A New Corpus for Complex Word Identification Research in Law Studies in Ecuadorian Spanish (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023a).

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

11. **SimpleText@CLEF**¹¹ at **CLEF 2023**¹² - Automatic Simplification of Scientific Texts.

Task 2: Complexity Spotting: Identifying and explaining difficult concepts for general audience.

Overview of the CLEF 2023 SimpleText Lab: Automatic Simplification of Scientific Texts.

Available on https://link.springer.com/chapter/10.1007/978-3-031-42448-9_30.

Paper: SINAI participation in SimpleText Task 2 at CLEF 2023: GPT-3 in Lexical Complexity Prediction for general audience (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023b). Available on <https://ceur-ws.org/Vol-3497/paper-239.pdf>.

12. **DeTermIt!**¹³ at **LREC 2024**¹⁴ - Workshop Evaluating Text Difficulty in a Multilingual Context.

Co-Located with the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation LREC-COLING 2024. Available on <https://lrec-coling-2024.org/>.

Paper: Enhancing Lexical Complexity Prediction Through Few-Shot Learning with GPT-3 (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024b). Available on <https://aclanthology.org/2024.determin-1.7/>.

13. **SimpleText@CLEF**¹⁵ at **CLEF 2024**¹⁶ - Automatic Simplification of Scientific Texts.

Task 2: Identifying and explaining difficult concepts.

¹¹SimpleText - CLEF 2023 SimpleText Track

Available on <https://simpletext-project.com/2023/clef/>

¹²CLEF 2023 - Conference and Labs of the Evaluation Forum.

Available on <https://clef2023.clef-initiative.eu/>

¹³DeTermIt! - DeTermIt! Workshop Evaluating Text Difficulty in a Multilingual Context.

Available on <https://determin2024.dei.unipd.it/>

¹⁴LREC 2024 - The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation.

Available on <https://lrec-coling-2024.org/>

¹⁵SimpleText - CLEF 2024 SimpleText Track

Available on <https://simpletext-project.com/2024/en/>

¹⁶CLEF 2024 - Conference and Labs of the Evaluation Forum Information Access Evaluation meets Multilinguality, Multimodality, and visualization.

Available on <https://clef2024.clef-initiative.eu/index.php>

Paper: SINAI Participation in SimpleText Task 2 at CLEF 2024: Zero-shot Prompting on GPT-4-Turbo for Lexical Complexity Prediction (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024b). Available on <https://ceur-ws.org/Vol-3740/paper-319.pdf>.

14. Paper: *GovAIEc*: A Lexical Complexity Corpus for Spanish in Ecuadorian Public Administration.

Paper under revision.

The Figure 4.3 presents the scientific production carried out during the development of the research.

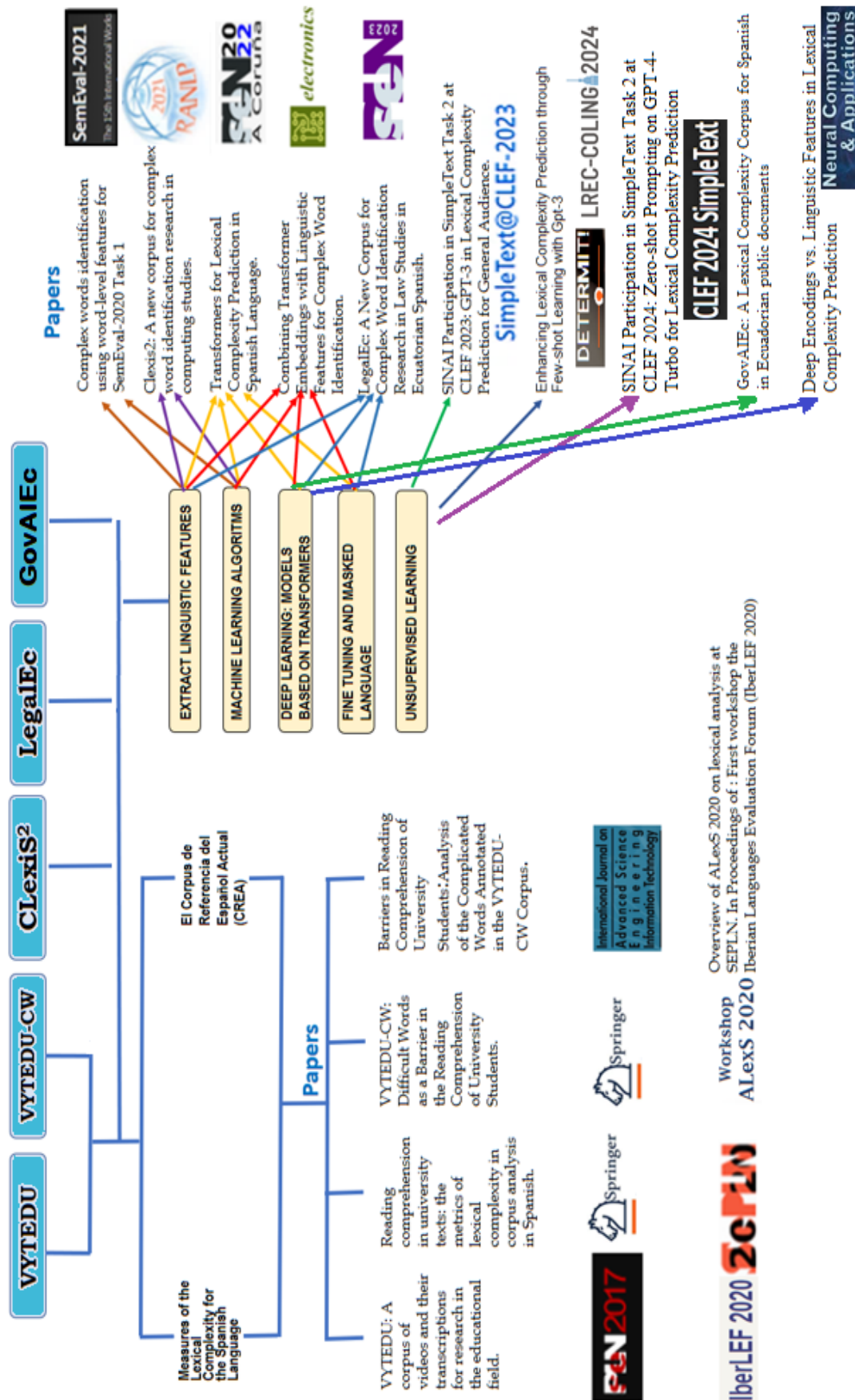


FIGURE 4.3: Scientific production carried out during the development of the research.

Below, each of the phases and how they have been developed are presented in detail, following the scheme presented in the Chapter III in section 3.3.

4.2 Creation Of Resources

Given that we currently face limitations in terms of the resources available to carry out research on complex words in Spanish in the academic field, we have recognized the need to create our own resources. Figure 4.4 presents a diagram of the resources created and the origin from which the texts come.

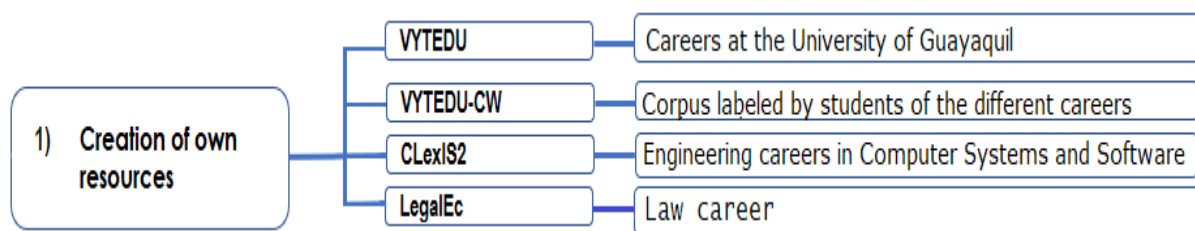


FIGURE 4.4: The own resources created by authors

We have created a comprehensive set of four corpora in the academic field, covering various disciplines of study:

- VYTEDU – (Ortiz-Zambrano and Montejo-Ráez, 2017).
- VYTEDU-CW – (Ortiz-Zambrano et al., 2019).
- CLexIS² – (Ortiz-Zambrano and Montejo-Ráez, 2021).
- LegalEc – (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023a).
- GovAIEc – Finished article to be sent to journal.

A comprehensive understanding of complex vocabulary within diverse academic disciplines in Spanish necessitates access to robust resources. Given the limited availability of resources tailored specifically to Spanish language studies, particularly within the realm of education, the significance of these resources cannot be overstated. The scarcity of parallel corpora and lexical references poses a considerable obstacle to advancements in Lexical Simplification, a challenge that is particularly pronounced within the Spanish language context (Bott and Saggion, 2012).

In general terms, we implement various strategies to collect the texts that make up each corpus. Initially, we chose to record our teachers' in-person classes, capturing their teaching in the classroom environment. We also took the opportunity to collect material from virtual classes, especially during the coronavirus pandemic. To further enrich our data set, we decided to include the final dissertations of law students, as well as several articles from the Constitution of the Republic of Ecuador. These resources not only enriched our knowledge but also contributed significantly to the advancement of research in this specific field. Below is a detailed description of each corpus.

4.2.1 VYTEDU - A Corpus of Videos and Transcriptions for Research in the Education Domain

The starting point for the development of our work originated from the conception of the VYTEDU corpus. This corpus, which consists of a collection of 55 videos along with their transcriptions, was proposed with the purpose of supporting research in the university educational field. This resource is presented as an essential source of data for the progress of our research. Although there are various collections of videos and transcripts used in research, such as the AMI corpus on videoconferencing (Carletta, 2007), which is used in the annotation of semantic roles (Sapru and Boulard, 2015). In the context of Spanish, the corpus created by Rosas, Mihalcea, and Morency, 2013 for sentiment analysis stands out, which consists of 105 videos extracted from the popular YouTube service.

In this context, the VYTEDU corpus stands as a fundamental and unique contribution, providing a data set specifically designed to support research in the educational field in Spanish. VYTEDU was presented through a demonstration at the SEPLN 2017¹⁷ congress held in the city of Murcia - Spain in 2017 (Ortiz-Zambrano and Montejo-Ráez, 2017). The analysis of the VYTEDU corpus was published in the article "Reading Comprehension in University Texts: Lexical Complexity Metrics in corpus analysis in Spanish" (Ortiz-Zambrano and Varela Tapia, 2018) published in the International Conference on Computer and Communication Engineering (pp. 111-123). Springer, Cham¹⁸.

¹⁷SEPLN 2017 - Available in <https://sepln2017.um.es/>

¹⁸ICCCE 2018 Conference: Computer and Communications Engineering pp. 111-123. Available at https://link.springer.com/chapter/10.1007/978-3-030-12018-4_9

This resource was provided for our ALexS 2020 workshop proposal - First Workshop on Lexical Analysis at SEPLN - Lexical Analysis Task at SEPLN 2020 as part of the second edition of IberLEF 2020¹⁹ - Iberian Language Evaluation Forum that added the efforts of the IberEval and TASS workshops where participants applied interesting approaches to address the CWI problem in an unsupervised or semi-supervised manner (Ortiz-Zambrano and Montejo-Ráezb, 2020).

4.2.1.1 The Methodology

The creation of the VYTEDU corpus was developed in three phases: recording, transcription, and dataset creation, which are detailed below:

Phase: Recording

The process for creating the corpus was carried out by applying a procedure that involved recording the videos in the classrooms of the different careers of the different faculties of the State University of Guayaquil. The work lasted three months, during which time it was used to request authorization from the authorities to make a video in the classroom and thus record the teacher's master class. In this stage, 10 first-semester students of the Computer Systems Engineering degree from the Faculty of Mathematical and Physical Sciences collaborated.

Thanks to the support of the collaborating students, the different careers were visited in their different study modalities: morning, afternoon and night. A Samsung Galaxy S4 phone was used to record the videos while the teacher was teaching his class, using the device's own software. Some of the main characteristics that needed to be taken into account in this equipment to achieve quality results in both audio and video are shown in Figure 4.5.

The videos consisted of a recording lasting between 5 and 15 minutes since the goal was that when the videos were transcribed, the maximum number of paragraphs possible would be reached. Several videos were recorded in a longer amount of time than estimated because the pace at which some teachers taught their classes was varied. The minimum recording time of one video is 00:05:01 and the maximum recording time of another video reaches 00:21:08 (hh:mm:ss format).

¹⁹IberLEF 2020 - Iberian Language Evaluation Forum. Available at <https://ceur-ws.org/Vol-2664/>

MEMORY	
RAM	2 GB
Internal memory	64 GB
MicroSD expansion	up to 64 megabytes
MAIN CAMARA	
Resolution	13 megapixels
Flash	Flash LED
Video	Full HD 1080p
Battery life of 17 hours	
Processor	
Speed Processor	1.9 GHz Qualcomm APQ8064T Snapdragon 600
cores	4

FIGURE 4.5: Main hardware and performance characteristics of the Samsung S4 (mobile) device.

Phase: Transcription

Ten students from different courses were chosen, the material was divided among them and the files were identified according to the associated degree, as well as a numbering that indicated the order. The transcription process was manual, that is, no automatic transcribers were used. Once this process was completed, it was possible to obtain the number of paragraphs that each text contained and, at the same time, obtain the number of words that made up it. The data corresponding to each video along with its respective contents were recorded to generate corpus statistics, as seen in Figure 4.6.

#	Video name	Faculty	Career	video file size	recording time	video recording	# of words in the text	# paragraphs
1	Video-01-Licenciatura en Sistemas de I	Ingeniería Industrial	Licenciatura en Sistemas	66,126.00 Kb	9:47	2/14/2017	1318	11
2	Video-02-Ingeniería en Sistemas.mp4	Ciencias Matemáticas y Físicas	Ingeniería en Sistemas	2,084,939.00 Kb	16:39	3/11/2017	1894	16
3	Video-03-Ingeniería en Sistemas.mp4	Ciencias Matemáticas y Físicas	Ingeniería en Sistemas	2,049,791.00 Kb	16:22	3/11/2017	1307	11
4	Video-04-Ingeniería en Sistemas.mp4	Ciencias Matemáticas y Físicas	Ingeniería en Sistemas	1,822,849.00 Kb	14:33	3/11/2017	1619	12
5	Video-05-Ingeniería en Sistemas.mp4	Ciencias Matemáticas y Físicas	Ingeniería en Sistemas	2,239,387.00 Kb	17:52	3/11/2017	894	9
6	Video-06-Ingeniería en Sistemas.mp4	Ciencias Matemáticas y Físicas	Ingeniería en Sistemas	2,249,970.00 Kb	17:57	3/11/2017	2125	15
7	Video-07-Odontología.mp4	Medicina	Odontología	219,172.00 Kb	9:00	2/15/2017	1386	11
8	Video-08-Odontología.mp4	Medicina	Odontología	270,989.00 Kb	10:21	2/13/2017	1057	12
9	Video-09-Odontología.mp4	Medicina	Odontología	270,192.00 Kb	10:19	2/13/2017	989	10
10	Video-010-Licenciatura en Sistemas de I	Ingeniería Industrial	Licenciatura en Sistemas	1,472,198.00 Kb	10:00	2/14/2017	2646	18
11	Video-011-Ingeniería Agronómica.mp4	Ciencias Agrarias	Ingeniería Agronómica	28,116.00 Kb	10:35	2/16/2017	1046	12
12	Video-012-Economía.mp4	Facultad de Ciencias Económicas	Economía	162,948.00 Kb	6:55	2/16/2017	900	10

FIGURE 4.6: The VYTEDU corpus data model.

Phase: Dataset Creation

According to the data model of the corpus, the data that corresponded to each of the videos was registered, that is: a sequential order, name of the video, associated faculty and degree, file size, recording time in minutes, video format, date and time of recording, number of words and number of paragraphs of the transcript, as summarized in Figure 4.7.

The videos contain their recording of various topics that correspond to the different subjects of academic programs, such as: Virtual Library (Information Systems), Biomechanical principles of dental preparations (Dentistry), Botany (Agronomic Engineering), The problem of debt as a development problem (Economy), Market Economy (Authorized Public Accounting), Information Systems (Telecomputing Engineering), Educational Psychology (Psychology), Hydraulics (Civil Engineering), Strategic Administration (Commercial Engineering), LAN Networks (Networking), Reiteration (Environmental Engineering), Processes to direct and manage the execution of the process (Industrial Engineering), Entered and Reentered Construction Processes (Architecture), and some more. See Figure 4.7.

Video-02-Ingenieria en sistemas.mp4	11/03/2017 11:56	MP4 File	2.084.939 ...
Video-03-Ingenieria en sistemas.mp4	11/03/2017 15:28	MP4 File	2.049.791 ...
Video-04-Ingenieria en sistemas.mp4	11/03/2017 9:38	MP4 File	1.822.849 ...
Video-05-Ingenieria en sistemas.mp4	11/03/2017 9:21	MP4 File	2.239.387 ...
Video-06-Ingenieria en sistemas.mp4	11/03/2017 11:25	MP4 File	2.249.970 ...
Video-07-Odontologia.mp4	15/02/2017 1:22	MP4 File	219.172 KB
Video-08-Odontologia.mp4	13/02/2017 18:16	MP4 File	270.989 KB
Video-09-Odontologia.mp4	13/02/2017 17:50	MP4 File	270.192 KB
Video-10-Licenciatura en sistemas de Informacion.mp4	14/02/2017 22:06	MP4 File	1.478.198 ...
Video-11-Ingenieria Agronómica.mp4	16/02/2017 3:38	MP4 File	28.116 KB
Video-12-Economia.mp4	16/02/2017 17:20	MP4 File	162.948 KB
Video-13-Economia.mp4	16/02/2017 17:10	MP4 File	133.874 KB
Video-16-Publicidad y Mercadotecnia.mp4	16/02/2017 8:37	MP4 File	64.059 KB
Video-17-Publicidad y Mercadotecnia.MP4	15/02/2017 14:36	MP4 File	254.012 KB
Video-18-Ingenieria Comercial.mp4	16/02/2017 15:33	MP4 File	503.368 KB
Video-19-Ingenieria en Teleinformatica.mp4	15/02/2017 0:18	MP4 File	19.799 KB
Video-20-Licenciatura en Educacion Basica.mp4	15/02/2017 17:06	MP4 File	156.209 KB
Video-21-Psicologia.mp4	16/02/2017 9:10	MP4 File	49.930 KB

FIGURE 4.7: Examples of video files from the VYTEDU corpus.

The following are some examples of the transcribed texts. It should be noted that, in the cases of the examples from the corpus texts, they are presented in the official language of the country (Ecuador), but we have added the translation into English to facilitate understanding and analysis.

Note: The original text is written in Spanish, so we have included the respective translation into the working language of this thesis.

File name: Video-51 - Ingeniería en TeleInformática

Subject: "Clasificación de los Sistemas"

"Para que un sistema muestre un comportamiento oscilante es necesario que tenga al menos dos niveles que son elementos del sistema en los que se producen acumulaciones. En ocasiones se observa un comportamiento oscilante como algo natural en todos los procesos ejemplos: al verano le sigue el invierno, al calor el frío, la noche el día y siempre vuelve al estado inicial, entonces tengo un sistema oscilante por ejemplo cuando sabemos que llegó el verano, a continuación llega el invierno y así sucesivamente, en conclusión si el estado actual del sistema no nos gusta o no es el correcto, no es necesario hacer nada ya que todo parece ser cíclico y volverá a la normalidad por sí solo".

File name: Video-51 - Telecomputing Engineering

Subject: "System Classification"

For a system to show oscillating behavior, it must have at least two levels that are elements of the system in which accumulations occur. Sometimes an oscillating behavior is observed as something natural in all example processes: summer is followed by winter, heat is followed by cold, night is followed by day and it always returns to the initial state, so I have an oscillating system, for example when we know that it has arrived summer, then winter comes and so on, in conclusion if we do not like the current state of the system or it is not correct, it is not necessary to do anything since everything seems to be cyclical and will return to normal on its own.

File name: Video-40 - Ingeniería Ambiental

Subject: "La Reiteración"

"¿Qué pasa si las personas incumplen? Lo primero es que en el caso de que tengan ya sus dos no-conformidades, tienen que pagar multas impuestas que vamos a ver que van desde veinte hasta doscientos sueldos mínimos, entonces tenemos: pago de multas impuestas, ejecución inmediata de correctivos de la no conformidad (eso lo tenemos como en el caso de la gente de Daule inmediatamente tuvimos que contratar empresa mediadoras para que san guiarán el área

de hecho fue muy interesante porque utilizaron incluso de a un profesor de ingeniería química de la universidad de Guayaquil ellos generaron un polímero y ese polímero lo dispersaron en la zona y le pegaron palos y se hizo como un plástico y en ese plástico se pegaron todos los hidrocarburos fue bien interesante porque contrataron a varias compañías y una de las compañías lo que hizo fue llevar plumas de aves y ciertos compuestos para que se adhirieran los aceites pero la más interesante fue esa de ingeniería química porque realmente se vio cómo es un muy buen absorbente de las grasas”.

File name: Video-40 - Environmental Engineering.

Subject: “The Reiteration”

“What happens if people default? The first thing is that in the event that they already have their two non-conformities, they have to pay imposed fines that we will see range from twenty to two hundred minimum wages, then we have: payment of imposed fines, immediate execution of corrective measures of the non-compliance. conformity (we have that as in the case of the people of Daule. We immediately had to hire a mediating company to guide the area. In fact, it was very interesting because they even used a chemical engineering professor from the University of Guayaquil. They generated a polymer And they dispersed that polymer in the area and they glued sticks to it and it became like a plastic and all the hydrocarbons stuck to that plastic. It was very interesting because they hired several companies and what one of the companies did was bring bird feathers and certain compounds for the oils to adhere but the most interesting was that of chemical engineering because it was really seen how it is a very good absorbent of fats.”.

File name: Video-46 - Networking.

Subject: “Windows Server”.

“Windows server está orientado a administrar grandes volúmenes de datos optimizar el uso de la agenda de red, seguro este sistema operativo como vimos la clase pasada que nos quedamos hasta la parte de las versiones de Windows server había un Windows server que estaba orientado a administrar Hardware de alto rendimiento por ejemplo pueden tener ustedes un servidor de 64 procesadores, una memoria de 400 gigas. ¿Para qué sirve el sistema operativo? Administrar el hardware, administrar el software de la máquina los periféricos de entrada y salida”.

File name: Video-46 - Networking.

Subject: “Windows Server”.

“Windows server is aimed at managing large volumes of data, optimizing the use of the network agenda, surely this operating system, as we saw in the last class that we stayed until the part of the Windows server versions, there was a Windows server that It was aimed at managing high-performance hardware, for example, you may have a server with 64 processors and 400 gigabytes of memory. What is the operating system for? Manage hardware, manage machine software, input and output peripherals.

4.2.1.2 Analysis

The Analysis and evaluation of the complexity of the texts constitute an essential phase in the research process that focuses on identifying complex words in the academic context of Spanish. This initial stage lays the foundation for the study of the texts, which is the core of interest in this research, and also paves the way for the evaluation of a potential system in the future.

Measuring Complexity

To carry out this task, the system used a set of metrics previously selected by Saggion et al., 2015a and whose description is found in detail in Chapter II in section 2.3.1.

One of the indexes most used for its ease of calculation is the Automated Readability Index, better known as the ARI (Automated Readability Index) (Kincaid et al., 1975) for its acronym in English. This index measures the difficulty of a text based on the average number of characters (letters and numbers) per word and the average number of words per sentence, the formula indicates the following:

$$4.71 \times \frac{\text{num\_characters}}{\text{num\_words}} + 0.5 \times \frac{\text{num\_words}}{\text{num\_sentences}} - 21.43 \quad (1)$$

It refers to a readability test designed to measure the comprehensibility of a text, that is, how easy it is to understand the text (Saggion et al., 2015a), (Senter and Smith, 1967). The greater the value of this index, the more difficult the text presents. The result of this formula is a number

that approaches the level of the reader's necessary degree to understand the text.

To carry out this task, first, the plain text *.txt* files containing the transcripts of the VYTEDU videos were re-encoded using the UTF-8 format. A function was created to take as an input a list of tokens and generate a list of sentences. For this, we carried out the treatment of each text through the following steps:

1. The text of the file read was passed to this function as input and the result was the generation of a list of tokens, each token corresponded to a term (word, number or punctuation mark). Abbreviations such as US, U.S.A, among others, words with intermediate hyphens, currencies and percentages were allowed, e.g.: \$ 12.40, 35%, 36.3 €, an ellipsis "...", more than two spaces, also the symbols [.,; ""' ? () : - _"] were considered isolated tokens, and once the tokens were extracted from the text they were converted to lowercase.
2. Then we grouped the obtained tokens by sentences, for this the end-of-sentence markers were specified.
3. We proceeded to the elimination of words of low semantic content, those words are also known as empty words (Rodríguez, 2017), these are words that have no meaning or as they are also called: stop words (in English). To achieve that, it was enough to go through the tokens and eliminate those terms that appear in a dictionary of empty words, NLTK incorporates one of these dictionaries for Spanish.
4. Currently, NLTK does not have a POS tagger. Commonly called POS Tagging, it is the labeling of the words of a text according to its grammatical category, but we can use the NLP group of Stanford University; the tagger was downloaded from Stanford University and placed on the same route where the notebook is located, proceeding to unzip it (Reyes, 2016), (Fernández et al., 2009).

Later, each token was converted into a tuple, the same one that was made up of the token and the POS (Part-Of-Speech tagger that specifies if the token is a noun, verb, pronoun, preposition, adverb, conjunction, participle and article). Finally, we proceeded to filter to stay alone with adjectives, verbs, and nouns. With all this, the following structures were obtained:

- List of words keeping adjectives, verbs, and nouns.

- List of words but without empty words.
- List of words but associating the POS of each word.
- List of words keeping adjectives, verbs, and nouns.

Low frequency words were calculated, and then the number of words with different content, the number of sentences, the total number of total content words, and the index of low-frequency words were calculated. The data previously obtained allowed us to calculate the Index of Lexical Distribution - LDI.

Finally, the LC - Lexical Complexity was obtained, whose values came from the calculations previously made, you can see Figure 4.8. As a result of the application of the formulas to obtain the index of the lexical complexity in each university text, the following were obtained: the number of low-frequency words, the number of different words, and the number of sentences, which in turn allowed obtaining the indexes of the lexical distribution, index of words of low frequency and the lexical complexity of the text. The basic results were obtained that allowed us to complement the analysis of the formulas proposed in the section 4.2.1.2. See Figure 4.8.

```
Trabajando en: Video-01-Licenciatura en sistema de informacion.txt .....

LEXICAL COMPLEXITY INDEX:
Number of low frequency words (N_lfw):  491

Number of distinct content words (N_dcw) =  281
Number os sentences (N_s):  46

Number of total content words (N_cw):  491

Lexical Distribution Index (LDI) =  6.108695652173913
Index Low Frequency Words (ILFW) =  0.7462006079027356
Lexical Complexity (LC) =  3.427448130038324
```

FIGURE 4.8: Execution of the application for the calculation of the metrics of the lexical complexity in the VYTEDU corpus.

Some statistical formulas were also implemented in the developed code that helped to complement the analysis of the lexical complexity of the texts, such as: *minimum value, maximum value, median and quartiles*, see Table 4.1. As can be seen, the most relevant data is shown after processing each text of the corpus, which will show us the legibility of the corpus texts, obtained according to the proposed formula, it can be observed that both the number of low-frequency words such as the index of low-frequency words are within normal values which means that the

lexicon used by two teachers facilitates learning in students.

Results from the Lexical Complexity in the archives of the VYTEDU corpus							
	LC	SSR	ASL	CS	SCI	ARI	PM
count	55.0000	55.0000	55.0000	55.0000	55.0000	55.0000	55.0000
mean	6.4006	194.7720	51.8628	0.5163	26.1845	26.3411	111.9636
std	3.1845	43.4010	32.3615	0.2602	16.2967	15.0902	55.7116
min	2.5688	128.6732	13.8603	0.0416	6.3130	8.3771	26.0000
25%	4.1312	165.2364	27.2140	0.2996	13.8049	15.4333	74.5000
50%	5.7914	184.0231	39.2888	0.50000	19.9666	20.5077	102.0000
75%	8.3438	209.2635	69.9372	0.7100	35.3450	33.4690	138.5000
max	15.2745	332.3202	132.6666	1.0000	66.7500	65.3331	344.0000

TABLE 4.1: Implementation of statistical formulas in the results from the Lexical Complexity in the archives of the VYTEDU corpus.

It should be noted that according to the analysis of the data, the LC (Lexical Complexity) in some texts is high, which means that the content becomes difficult in some cases, this is also due to the number of rare words that the teachers are using when they teach their classes. With the above data, the SSR could already be calculated. See Table 4.2.

The result obtained when classifying the SSR - Spaulding Spanish Readability in the texts of the corpus is the following: the mean is considered in the value of 184 what it means according to the table of reference values proposed (120 - and above) (Spaulding, 1956), the texts present an exceptional difficulty of being read the texts written in Spanish. 25% of the texts have a value of 165, which according to the table of values is also within the range of exceptional difficulty, while 75% and the rest are above the values described previously. The university texts used in this analysis are within the level of excessive difficulty, their content can not be easily read without having to resort to mental skills exercises and implementation of types of concentration to understand their content.

The study also included the calculations of the formulas: SCI as a measure of the complexity of sentences in a literary text, the ASL (Average Sentence Length) and the CS (Complex Sentence), obtaining as a final result, the SCI (Sentence Complexity Index).

Punctuation marks (PUNCT) refers to the average number of punctuation marks per text. This can be obtained with a POS and counting the frequency of the PUNCT label (with Freeling, for example). The punctuation marks directly influence the clarity and comprehension of the

The SSR in VYTEDU corpus				
	N_w	N_{rw}	N_s	SSR
count	55.0000	55.0000	55.0000	55.0000
mean	1369.8000	363.8363	33.8545	194.7720
std	507.5091	134.7234	15.1528	43.4010
min	541.0000	131.0000	6.0000	128.6732
25%	992.0000	272.5000	21.0000	165.2364
50%	1282.0000	358.0000	36.0000	184.0231
75%	1506.0000	431.0000	45.5000	209.2635
max	2918.0000	723.0000	64.0000	332.3202

TABLE 4.2: Calculation of the SSR-Spaulling's Spanish Readability.

content of the text contributing to its correct interpretation (De Jesús Torres, 2017). Table 4.1 presented the results after implementing the statistical measures in the results obtained from the calculation of the metrics.

For the analysis of the data obtained from the application of the lexical complexity metrics in the texts of the university videos, some texts were chosen taking into account the number of words and their respective result of the lexical complexity - LC. It is considered that not necessarily the texts that are composed of more words are those that have higher LC index, for this we have considered as an example the texts Video-17 and Video-10. See Table 4.3.

LC and SSR formulas in VYTEDU corpus											
	N_{chrs}	N_w	N_{dcw}	N_{cw}	N_{lfw}	N_s	N_{cs}	LDI	ILFW	LC	SSR
Video-01	6529	1510	281	658	491	46	14	6.10	0.74	3.42	182.70
Video-10	13034	2918	516	1218	723	35	19	14.74	0.59	7.66	238.35
Video-17	2598	541	198	246	140	16	4	12.37	0.56	6.47	162.26
Video-24	2966	611	195	306	216	44	4	4.43	0.70	2.56	161.64
Video-55	3197	768	180	295	162	6	6	30.0	0.54	15.27	297.94

TABLE 4.3: Calculation of the LC-Lexical Complexity and SSR-Spaulling's Spanish Readability formulas.

The text Video-17 has a quantity of 541 words and the text Video-10 has a total of 2,918 words corresponding to the minimum and maximum value of words respectively according to the statistical data presented in Table 4.3. The values of the measure of LC are 6.47 and 7.668 respectively, that is, with a fairly short difference; if we compare it with the number of words in the Video-55 text that we have taken as reference because according to Table 4.3 it is the one with the highest value of LC, it is composed of 768 words and an LC of 15,274 is located like the one of greater LC of all the corpus, that is, to say a text of great complexity.

It is worth mentioning that there are texts such as Video-24 which is composed of 611 words, just a little above the total number of words in the Video-17 text, but it has an LC of 2.56, which is much lower than the LC of the Video-17 text that it is 6.47, which can be concluded as a text of easy comprehensibility. Regarding the SSR measure, in the values presented in Table 4.3, if we take these texts as a sample, we would have the data presented for the SSR in the same Table 4.3.

The SSR is related to the number of words, which is why we will analyze the texts of Table 4.3. We can observe, according to the data presented in the SSR column, that the text Video-01 has a total of 1510 words and an SSR of 182,70, while the Video-55 text has a much smaller number of words and has an SSR of 297,94. Also, the text Video-24 has a total of 611 words, that is, at the moment it is the smallest of the texts of this analysis of the SSR presenting an SSR of 161.64. The text Video-10 has an SSR of 238.35 when its total of words is 2918, that is to say, a total of words much greater than those presented in the previous texts, and its SSR did not surpass the calculated of the text Video-55.

Regarding the SCI analysis of the texts of the corpus, we have taken a sample of texts, observing a very high sentence complexity index; for this, the columns of the number of words, number of sentences, number of complex sentences, and the SCI have been considered, see Tab 4.4. The text Video-24 contains 611 words and has an SCI of 6.98 while the text Video-55 has 768 words and an SCI totally superior to the previous one being this one of 56.45; text Video-46 is slightly larger in number of words than text Video-55 and its SCI is 66.7, when analyzing the text Video-07, this contrasts the previous ones since the SCI is 11.96.

SCI formula in VYTEDU corpus				
	N_w	N_s	N_{cs}	SCI
Video-24	611	44	4	6.98
Video-38	1119	10	10	56.45
Video-55	768	6	6	64.50
Video-48	796	6	5	66.75
Video-07	1502	64	30	11.96

TABLE 4.4: Calculation of the Sentence Complexity Index - SCI.

In the calculation of the ARI made in VYTEDU it can be seen that there is an average of 26.34 where the minimum value calculated is 8.37, which means that it is an easy-to-read text, while the maximum value calculated is 65.33, this being a text difficult to understand because of the

high value obtained. See Table 4.5.

The 25% of the texts of the corpus have an ARI lower than 15.43 which is easily accessible, another 25% has an ARI that is between 15.43 and 20.50 which means that it is slightly complex, another 25% is between 20.50 and 33.46, that is, 50% of the texts of the corpus present a moderately complex index of difficulty, while the last 25% have a very complex percentage of comprehension difficulty since it is above 33.46.

ARI formula in VYTEDU corpus	
	ARI
count	55.00
mean	26.34
std	15.09
min	8.37
25%	15.43
50%	20.50
75%	33.46
max	65.33

TABLE 4.5: Calculation of the ARI formula - Automated Readability Index.

The calculations were obtained for Punctuation Marks (PM). See Table 4.6 where 25% of the texts have a number of punctuation marks lower than 74, another 25% is in the range of 74.50 to 102, while 25% is between 102 to 138, and the last 25% of the texts of VYTEDU have an amount greater than 138.50 with the maximum number of 344 punctuation marks in the content of the text.

Punctuation Marks in VYTEDU corpus	
	PM
count	55.00
mean	111.96
std	55.71
min	26.00
25%	74.50
50%	102.00
75%	138.50
max	344.00

TABLE 4.6: Calculation of the PM formula - Punctuation Marks.

In analyzing the corpus of academic texts, we have identified a significant number of complex words that present a challenge for comprehension. This finding not only highlights the need for effective tools to address lexical complexity in academic texts, but also opens the door to new research in this area. Predicting and treating complex words is the focus of our research, and developing this can facilitate the accessibility and comprehension of specialized texts. We consider it essential to create more resources, such as tagged corpora, that help advance this line of research. These corpora would provide a solid database for the study of lexical complexity, and would serve as a benchmark for evaluating new methodologies and approaches in the future. Furthermore, such resources could contribute to the creation of automatic tools that support both researchers and students in navigating complex academic texts.

4.2.2 VYTEDU-CW - A Corpus of Videos and Transcriptions for Research in the Education Domain - Complex Word

4.2.2.1 Origin Of VYTEDU-CW

The VYTEDU-CW (Educational Videos and Texts - Complex Word) is a corpus in Spanish made up of university educational texts containing difficult words. VYTEDU-CW began in 2017 as the genesis of the VYTEDU corpus, thus marking the beginning of our research project. In its construction process, the work consisted in the recognition and labeling of the complex words contained in the different texts that make up the VYTEDU corpus. It should be noted that one of the difficulties in reading comprehension is complex words, which become a barrier especially for those with special cognitive abilities. The objective of this work is to determine that difficult words in educational texts become one of the main problems in the reading comprehension in the day to day of university students (Ortiz-Zambrano et al., [2019](#)).

4.2.2.2 Related Work

Numerous text simplification projects have established as a first step the identification of complex words or sentences present in the textual content (De Belder and Moens, [2010](#)). The detection of complex words plays an important role in the lexical simplification process. For this reason, in accordance with previous research, it was imperative to label the texts included in the VYTEDU corpus. The CWI (Complex Word Identification) aims to identify complex words candidates for simplification in a given text (Alarcón, Moreno, and Martínez, [2021](#)).

An exhaustive bibliographic analysis confirms that the identification of complex words constitutes a recurring component in the proposals for lexical simplification projects, which provides us with a solid basis for its implementation in our own project. Next, we will present some of the methodologies commonly used in this context:

The PSET project directed to the creation of accessible texts for aphasic readers. As part of the simplification process, difficult words were identified and then the lexical simplification process was carried out, which consisted of replacing the difficult words with others that were easier to understand, thus reducing the lexical complexity of the text (De Belder and Moens, 2010). Consisted of a simplifier of texts for English in which both lexical and syntactic simplification was implemented. In the lexical simplification phase, the detection of complex words in which they were identified in the text was carried out through the frequency of values and the use of psycholinguistic resources (Ferrés et al., 2016).

Suter, Ebling, and Volk, 2016 presented an automatic text simplification system for German with a focus based on simplification rules supported by linguistic motivation guidelines to transform standard German into simplified German. The process of simplification of text was based on the identification of difficult words, then they were replaced by their respective definition from the dictionary of simplified definitions in German called Hurraki, in this way a more readable text was obtained.

Pioneering work on Text Simplification was carried out in the Czech Republic. For the identification of difficult words, the author marked what he considered to be complex, then he commissioned three scorers to add or delete some complex words and then the three scorers checked the scores they made. All words marked as complex by at least one annotator were used in the set of evaluations. Finally, the generation of substitution was tested using these words, it was first evaluated separately for the sources based on dictionaries and for the sources based on inlays, and then jointly (Burešová, 2017).

4.2.2.3 Construction Process Of The VYTEDU-CW Corpus

About the EIL application

In the first step, we developed a software called EIL (Environment in Natural Language Processing Research) to load the texts from the VYTEDU corpus. As part of the process of labeling complex words, we carried out visits to the various faculties of the University of Guayaquil.

There, we selected a representative sample of students by major and proceeded to carry out the annotation process. See Figure 4.9.

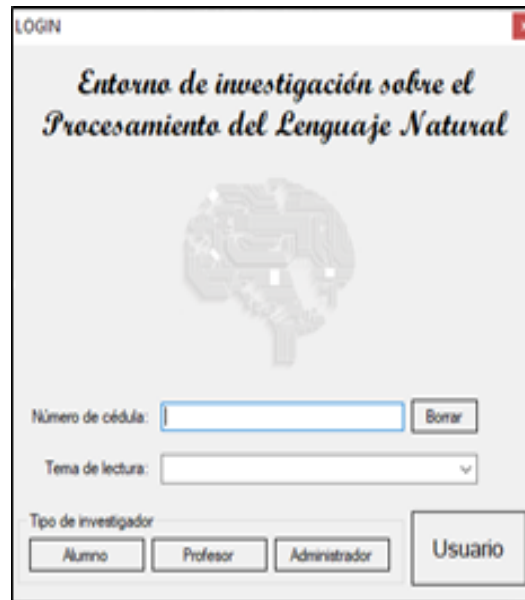


FIGURE 4.9: EIL Software.

Labeling Of Complex Words

The selection of students was carried out considering their level of studies in relation to the content of the texts in the corpus, with the aim of guaranteeing that the text was not too easy for those who were in higher grades. To enter the system, students were asked to provide their personal identification and then select the major they belonged to before pressing the “student” button. See Figure 4.10.

Showing up next the Figures 4.11 which presents an example of the complex word labeling process.

Subsequently, the application loads the texts related to the career to which the student belongs, who must begin to choose the text and start the reading. During the process of reading the material, the student proceeds to identify the word or words that are difficult to understand and then proceeds to label it as a difficult word, the same as the system will show shaded. This operation is repeated throughout the text with all words equal to the one chosen by the student. See Figures 4.12 and 4.13.

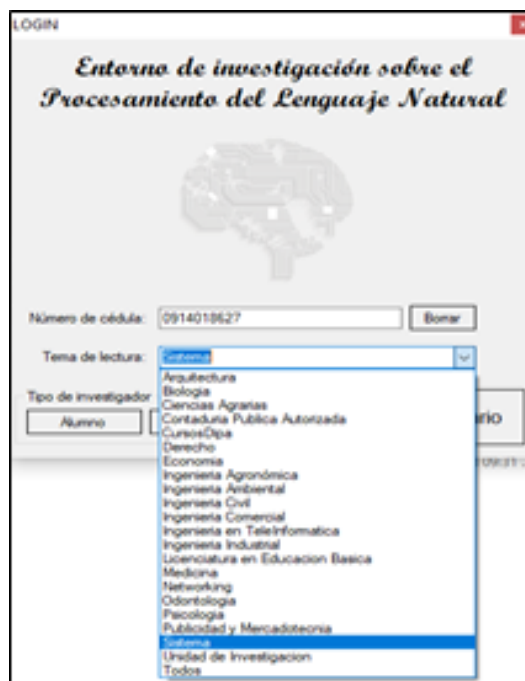


FIGURE 4.10: EIL Software: Complex word labeled system.

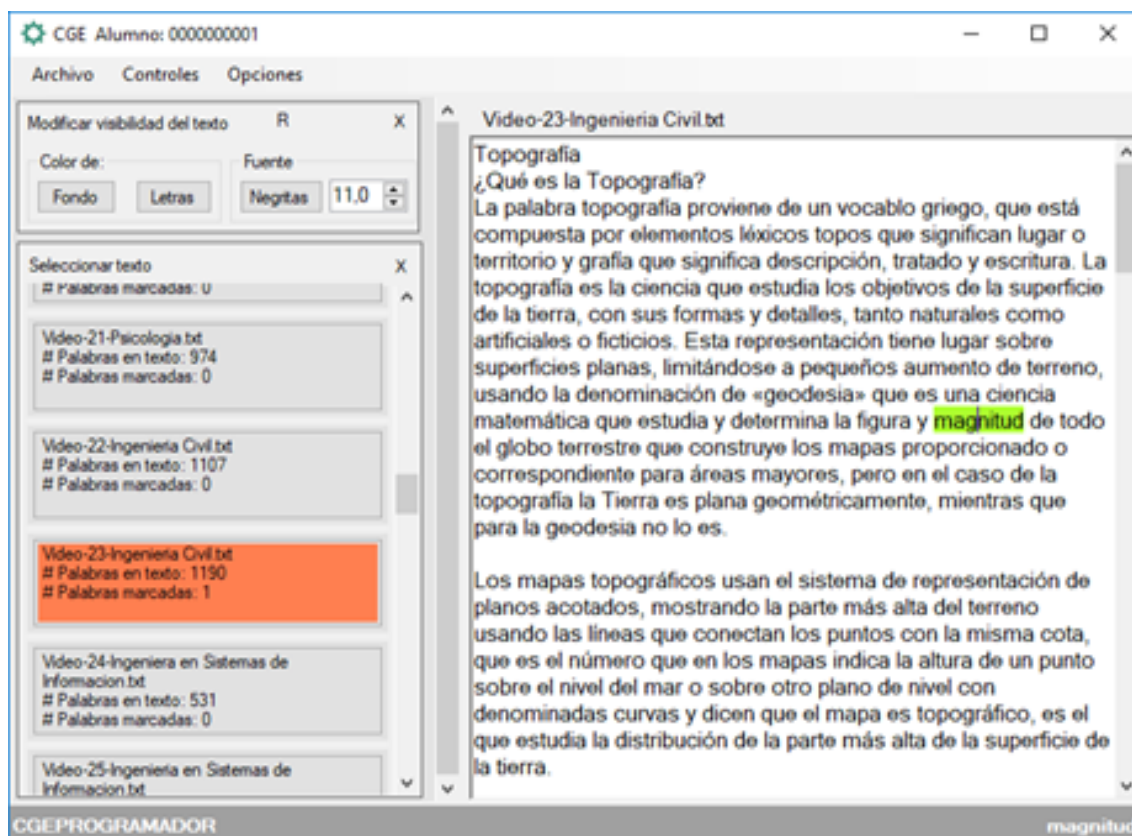


FIGURE 4.11: EIL Software - Labeling of difficult words identified in the text.

In this way, it is stored the word that was identified as difficult along with the type of user who accessed the system (since it could be a teacher or a student), the type of word that has been



FIGURE 4.12: Students labeling difficult words in texts.



FIGURE 4.13: Students labeling difficult words in texts.

marked (since it can be a difficult word or specialized word, where the specialized words are specific to the subject and as such should be learned without alternatives), the user who tagged it, the name of the file of the text that was read, the initial position of the word in the text (it is a number), the length of the word, and the date and time when the text was read. See Figure 4.14.

4.2.2.4 Corpus Statistics VYTEDU-CW

Very important data was obtained in the progress of this investigation regarding the identification of difficult words that contain educational texts. We can appreciate the following relevant data. See Table 4.7.

complex word	tipo	Student	tipo-usua	texto	pos en el	longitud	repetidor
genoma	Difícil	922313671	Alumno	Video-02-Unidad de Investigacion.txt	1783	6	3
chemistry	Difícil	922313671	Alumno	Video-02-Unidad de Investigacion.txt	7498	9	12
verborragia	Difícil	922313671	Alumno	Video-04-Unidad de Investigacion.txt	4954	11	8
maceración	Difícil	930404041	Alumno	Video-05-Unidad de Investigacion.txt	2265	10	4
taxonómico	Difícil	930404041	Alumno	Video-05-Unidad de Investigacion.txt	4707	10	10
prodequest	Difícil	930404041	Alumno	Video-06-Unidad de Investigacion.txt	3526	10	6
ISBN	Difícil	930404041	Alumno	Video-36-Unidad de Investigacion.txt	880	4	4
maceración	Difícil	928757343	Alumno	Video-05-Unidad de Investigacion.txt	2265	10	4
bioética	Difícil	928757343	Alumno	Video-05-Unidad de Investigacion.txt	3571	8	8
Intoxicate	Difícil	928757343	Alumno	Video-05-Unidad de Investigacion.txt	3945	10	9
taxonómico	Difícil	928757343	Alumno	Video-05-Unidad de Investigacion.txt	4707	10	10
mío	Difícil	928757343	Alumno	Video-06-Unidad de Investigacion.txt	2446	3	4
obtusa	Difícil	928757343	Alumno	Video-06-Unidad de Investigacion.txt	3074	6	6
ponencia	Difícil	928757343	Alumno	Video-36-Unidad de Investigacion.txt	938	8	4
Twitter	Difícil	952348837	Alumno	Video-46-Networking.txt	638	7	2
Windows	Difícil	952348837	Alumno	Video-46-Networking.txt	138	7	2
Linux	Difícil	952348837	Alumno	Video-46-Networking.txt	1028	5	4
Office	Difícil	952348837	Alumno	Video-46-Networking.txt	1599	6	4
Hardware	Difícil	952348837	Alumno	Video-46-Networking.txt	1987	8	6

FIGURE 4.14: The data library of the corpus VYTEDU-CW.

Definition	Total
All records	
# of terms	1628
# of students	422
# of corpus entries	573
Records without repetition	
# of terms	661
# of students	242
Corpus	
# of words	68301
# of words without repetition	9175
Corpus vs. marcas	
# of terms	2.88%
# of terms without repetition	7.20%

TABLE 4.7: The statistics of the corpus VYTEDU-CW.

where:

- The number of words in all the texts contained in the corpus is 68.301.
- The total number of users who have tagged words is 422 students.
- 573 students were the ones who accessed the application.
- There is a total of 661 different words that were marked throughout the corpus.
- 242 is the number of users who have entered the application once.
- 9175 is the total number of words in the whole corpus without repeating.

The value *Corpus vs marks* means that 2.38% of words were tagged from the total number of words in the corpus, while the % of terms without repetitions is the percentage of words that do not repeat with respect to the total number of words.

4.2.2.5 Analysis

We were closely following the work done by Burešová, 2017. Through the EIL application, students were able to identify and tag the complicated words present in the texts of the VYTEDU corpus. This process conclusively revealed the existence of complex words that represent an obstacle to reading comprehension.

We carried out another investigation using the Vytedu-cw corpus, called “Barriers in Reading Comprehension of University Students: Analysis of Complicated Words Annotated in the VYTEDU-CW Corpus”, in order to address the challenge that complex words represent in texts university students in Spanish corpus (Ortiz-Zambrano and Montejo-Raéz, 2020) where the analysis of the words labeled as complex confirms the conclusions reached by the research work of García Muñoz, 2014 and corroborates the recommendations of the Easy Reading guide prepared by Europe, 1998, as well as the guidelines proposed in the “EASY TO READ” guide in the automatic identification of complex words (CWI) (Alarcon et al., 2019).

These research work have attracted attention in recent years, with the advent of deep learning approaches (Gooding and Kochmar, 2019) and multilingual challenges (Finnimore et al., 2019), which contributes to the evaluation of words labeled as severe, such as specialized words, common lexicon words, slang, English words, acronyms, among others. According to the results of the analysis, the students had difficulty understanding the words. In some cases they were

unaware of its meaning or had some idea or notion of it. Among other things, this analysis helped characterize the type of complex words identified in the VYTEDU-CW corpus. We were able to reach the conclusion that the words labeled as difficult were composed of specialized words, that is, typical of the subject, words of the common lexicon, jargon, English words, acronyms.

Here are some examples:

- *Use of technicalities by teachers* when teaching their classes (technical voices used in the language of science) as is the case of the degree “Degree in Information Systems” in which words were labeled as: *break, for, while* or *variable*.

In the “Networking” career the words: *cluster, hosting, kernel*; in the career of “Odontology” the words: *deflection, homeostasis, muñón*. In the “Architecture” career, the words: *poetry, hepatization, hydrolyzing*.

- *Use of sophisticated vocabulary*. Some teachers use an improved, cultured terminology, which makes it difficult to understand the content for many students in the classroom. For example, students of the “Odontology” career labeled words such as: *contributive, alienation, tariff, optional, praxis* or *revocation*.

In the “Authorized Public Accounting” career, words such as: *inference, subsidiary* or *contingencies*. In the case of the “Research Unit” it was determined according to the analysis performed that the teacher uses a lexicon so sophisticated and special that very few understand words such as: *ultrasonic, taxonomic, maceration, genome, carto-print, bioethics* or *synthesization*.

- *Use of abbreviations* is also a cause of difficulty in reading comprehension. It was possible to determine that in the case of the students of the “Networking” career, words corresponding to their own abbreviations were labeled, such as *Mbps, IPAN, GEAR, LAN*. In the case of the “Research Unit” the acronym *ISBN*.

- *Use of words that do not belong to the level of teaching or specialization* by teachers when proposing examples. While they teach their classes, they use words that even the students

of a certain level of studies do not know, as is the case of the students of the “Degree in Information Systems” career, words such as: routers, tree structure, script.

In the case of the career in “Agricultural Science Engineering” the term: Python.

In the case of the “Research Unit”, the teacher used computer terms with the group of students who are of different specializations and are the ones who take the course to train and be able to prepare their end-of-course research work. It was observed that many do not understand because they are obviously out of the specialty; words were tagged as: virtual, pixels, logical operators or jpg.

- *The use of metaphors and figurative language* causes difficulties when it comes to understanding what the teacher wants to convey, such is the case of the students of the “Research Unit” who labeled the phrase “if you want to describe the truth, leave the elegance for the tailors”.

- *The use of long, hard-to-pronounce and unusual words* makes it difficult to recognize them immediately. To cite several examples: the case appears again:

In the “Research Unit”, where students had labeled words such as: qualitatively, quantitatively, chronologically, consent, particularly.

In the “Law career”, words such as: homogeneity, consideration, methodological, interculturality appear.

In the “Advertising and Marketing” career, words such as: imprescriptible, inalienable, decentralized.

In the “Economy” career the words: decentralized, totalitarianism.

In the “Civil Engineering” career, the words: correlating them, deontology, sclerometry, workability.

In the career of “Agricultural Engineering” the words: unanimously, carrageenans, paleobotany, carrageenans, oleoresins.

In the career of “Authorized Public Accounting” the word: granting.

In “Degree in Information Systems” the words: oversize, overestimate. In “Agricultural Science Engineering”: unanimously.

- *The use of verb nominalization* was also a cause for word tagging. Several examples are cited as those found:

In the “Networking” career in the word: emphasized.

In the career of “Law” the words: scrutinized, emanated, revalued. In the “Research Unit” the words: delimited, conditioned, synthesized, standardized.

In the career “Degree in Information Systems” the words: digitized, cohesive.

In the career “Engineering in Agricultural Sciences” the word: enchanted.

In the “Advertising and Marketing” career the word: decentralized.

In “Civil Engineering” the word: workability.

- *The use of compound words* was also a cause of difficulty in students, to name a few examples:
 In “Degree in Information Systems” the expressions: intermediary nucleus, tree structure.
 In the “Research Unit” the words: ultrasonic, taxonomic view.
 In “Economy” the term: stock market circuit.
 In “Civil Engineering” the term: flexo traction.
 In “Tele-Computing Engineering” the word: preset.
 the “Odontology” career labeled words as: consideration.
- *The use of colloquial language*, words with colloquial vocabulary are not observed, which could be explained by being transcripts of university lessons in which an academic language is presupposed.

The results of word labeling and their statistical analysis validate the achievement of one of the objectives established at the beginning of an essential phase of our project, focused on the identification of complex words in the academic field in Spanish.

4.2.3 CLexIS² - Lexical Complexity in Computer Systems and Software Engineering Careers

Our interest in this field is constantly growing and we are eager to conduct research to actively contribute to research on the identification of complex words in the academic context of Spanish.

We were interested in the premise of Pitkowski and Gamarra, 2009 where they states “The processing of extensive collections of electronic texts plays an essential role in their applicability in various fields of study in the field of linguistics, such as the teaching of a second language” (L2), lexical and syntactic simplification, text prediction, automatic translation, information

retrieval and extraction, speech synthesis and language analysis, among others.

After conducting our research and analyzing the results from the VYTEDU corpus, we've found that it contains many complex words, making comprehension difficult. We've also assessed the labeling results in creating the VYTEDU-CW corpus (Ortiz-Zambrano et al., 2019), (Ortiz-Zambrano and Montejo-Raéz, 2020). Building on the findings of Davidson et al., 2020, who noted the scarcity of Spanish corpora for NLP researchers, along with the lack of annotations in some existing corpora, we recognize an urgent need. Thus, we aim to address this gap by creating a new corpus tailored for Computer Systems and Software Engineering majors at the Faculty of Mathematical Sciences, University of Guayaquil. This corpus, named CLeXIS² - Lexical Complexity in Computer Systems and Software Engineering Careers (Ortiz-Zambrano and Montejo-Raéz, 2021), aims to provide valuable resources for NLP model development.

4.2.3.1 The Methodology

As a first step, the subjects that make up each semester of study were identified. The first four semesters correspond to the Software career and the following four semesters to the Computer Systems Engineering career, giving a total of eight study semesters, where each semester consists of five subjects.

Next, the recordings of the classes (academic videos) taught by teachers in virtual classes in the last two semesters of study were selected, which were stored in each teacher's work cloud.

Using the Dictation²⁰ application, the transcription process of each of the academic videos was carried out. The automatic transcriber did not have precision in the accentuation of the words as well as in the punctuation of the sentences, therefore, a process of grammar revision was carried out manually in each text to achieve its correct accentuation; it was also considered vitally important to separate the text into sentences for better understanding.

Finally, the number of texts per topic corresponds to an average of 100 texts, and in turn, each text contains an average of 77.29 words. In the Table 4.8 you can see other details that correspond to the content of the dataset. The table 4.9 shows the descriptive statistics of CLeXIS², with a total of 3,887 texts. The definition of the variables is detailed in Table 4.10.

²⁰Dictation - <https://dictation.io/speech>

<i>CLexIS²</i> corpus	
Detail	Total
Total number of documents per course	An average of 500 texts
Total number of documents per subject	An average of 100 texts
Global number of documents	3887 texts

TABLE 4.8: Contents of the *CLexIS²* corpus.

The Statistics of <i>CLexIS²</i>							
	N_w	N_{cw}	N_{dcw}	N_{rw}	N_{lfw}	N_s	N_{cs}
Mean	77.29	39.59	51.66	36.91	7.04	2.51	1.06
Std. Dev	20.09	9.26	11.66	9.08	4.38	1.49	0.80
Min	9.0	9.00	9.00	7.00	0.00	1.00	0.00
Max	167.00	91.00	92.00	81.00	49.00	18.00	6.00
Sum	300420.00	153885.00	200785.00	143464.00	69803.58	9756.00	4101.00

TABLE 4.9: Descriptive Statistics of different counters over documents in *CLexIS²*.

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 4.10: Definition of the columns in Table 4.9.

4.2.3.2 About Annotators

The students who participated in the labeling *CLexIS*² are Ecuadorian university students with enrollment in the Computer Systems Engineering career, and the Software career in the regular academic period 2020-CII. Five annotators were chosen for each semester of study to carry out the corpus labeling work.

The average score of the academic performance of the participants, according to a university expedient was of 8.72 over 10 points. It should be noted that no distinctions were made in the selection of students who would carry out the process of labeling the complex words of the corpus texts. The students came from different levels of secondary education (private school, national government), economic and geographical locations including vulnerable sectors such as suburban neighborhoods, rural parishes, and several housing cooperatives located on the outskirts of the city.

4.2.3.3 Labeling Process

We developed an application with free software tools Python, Firebase, and Cloud Firestore for the creation and management of the database, and the texts were loaded into the system. The taggers had to begin by reading the texts that corresponded to their study semester and then identify and write down the words that were difficult for them to understand. The annotated data was collected for later management. The dataset consisted of the following fields: the token (the difficult word), the annotator identification, the position of the token in the text, the name of the text, the length of the token, and its frequency. Table 4.11 presents several examples of the words that were tagged in the texts of the *CLexIS*² corpus (in bold).

Note: The original text is written in Spanish, so we have included the respective translation into the working language of this thesis.

Then we obtained the calculation to obtain the total number of words noted by the taggers, which we can see in Table 4.13. A large number of words labeled as complex presented a LCP complexity of 0.2, they correspond to the total number of complex words identified by one annotator, which means that there are no coincidences that these words have been annotated by the other taggers.

Course	Sentence	Target word	LCP score
Algebra Lineal	Sea A una matriz de dimensión mxm.	dimensión	0.20
Legislacion Informática	¿Qué son las ordenanzas distritales? Las ordenanzas distritales expedidas por los gobiernos de los distritos metropolitanos.	expedidas	0.20
Base de Datos S.O.D	La sentencia GOTO permite cambiar la línea secuencial de la ejecución hacia una etiqueta. Service Statelessness (No estado) Un servicio no debe guardar ningún tipo de información.	goto servicio	0.80 0.20
Algoritmos	Metodología para resolver un problema ¿Cómo se hace un programa?	metodología	0.60
Estructura de Datos	Un grafo G es desconexo, si dos o más de sus nodos no están conectados por caminos simples Grafo regular.	nodos	0.40
Economía	La economía es el estudio de la humanidad en sus quehaceres cotidianos.	quehaceres	0.20
Auditoría de Sistemas	El triunfo de una corporación depende de los controles de evaluación de la eficiencia y vigencia de sus sistemas de TI.	corporación	0.20

TABLE 4.11: Examples from the CLeXIS² dataset.

Course	Sentence	Target word	LCP score
Linear Algebra	Let A be an mxm matrix .	matrix	0.20
Computer Legislation Databases	What are district ordinances? District ordinances issued by the governments of metropolitan districts. The GOTO statement allows changing the sequential line of execution to a label.	issued goto	0.20 0.80
O.S.D	Service Statelessness (No state) A service should not retain any type of information.	service	0.20
Algorithms	Methodology for solving a problem How is a program made?	methodology	0.60
Data Structures	A graph G is disconnected if two or more of its nodes are not connected by simple paths Regular graph.	nodes	0.40
Economics	Economics is the study of humanity in its daily activities .	daily activities	0.20
Systems Audit	The success of a corporation depends on the controls for evaluating the efficiency and effectiveness of its IT systems.	corporation	0.20

TABLE 4.12: Examples from the CLeXIS² dataset (translated).

Software degree						Computer Systems degree					
Number of annotators						Number of annotators					
Sem	1	2	3	4	5	Sem	1	2	3	4	5
1st	1615	712	365	185	71	5th	2237	922	482	272	102
2nd	1964	788	358	182	44	6th	2119	809	499	246	79
3rd	1930	654	342	161	45	7th	2060	841	429	217	95
4th	1316	638	417	279	136	8th	1967	736	352	161	58

TABLE 4.13: Total number of complex words annotated by the taggers per semester in the careers of Software and Engineering in Computer Systems.

A summary of the annotation process is shown in Table 4.13. The columns in the table represent the number of scorers. Each row contains the semester of study and the total number of words rated as difficult in that semester. In addition, both in the Computer Systems Engineering degree and in the Software degree, the total number of words annotated by a single tagger is much greater than the words annotated by more than one tagger. The amount corresponding to the number of difficult words annotated decreases as the number of words annotated by more than one tagger increases, as the intersection among annotators is lower when the number of annotators increases.

4.2.3.4 Lexical Complexity Of The Corpus

The evaluation of the complexity of the *CLexIS*² corpus texts was carried out by applying the seven measures of lexical complexity for the Spanish language as in Saggion et al., 2015a and applied in our research work (Ortiz-Zambrano and Montejo-Ráez, 2017) and (Ortiz-Zambrano and Varela Tapia, 2018). These formulas were proposed by (Rebollo, 2008), with the exception of the SSR whose measurement was provided by (Spaulding, 1956). Details of each of the formulas are presented in Chapter II in section 2.3.1. The metrics are:

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

- The Percentage of Complex Sentence - CS

This evaluation was realized based on two factors:

- First: at the lexical complexity of reading texts for which the readability indices LC, LDI, ILFW, and SSR were applied.
- Second: the syntactic complexity of the texts where the measures applied were SCI, ASL, SC.

For data processing, the open source statistical software *Jasp* version 0.14.1 was used, obtaining the descriptive statistics of the subjects in Systems Engineering and Software degrees (40 subjects in total). The table 4.14 shows the values that correspond to the lexical complexity metrics detailed in the previous paragraph.

	LC	SSR	SCI	ARI	MTLD
Mean	2300.18	24759.84	2167.81	2462.55	1278.99
Std. Deviation	717.83	5673.33	1056.13	1005.39	266.26
Range	3566.30	32002.12	4512.75	4579.49	1526.32
Minimum	225.25	2089.50	260.75	271.90	88.59
Maximum	3791.55	34091.62	4773.50	4851.39	1614.91

TABLE 4.14: Lexical complexity measurements by subject.

The analysis of the results shows that the indices obtained determine that the texts corresponding to the first four semesters of the Software career and the remaining four to the Computer Systems Engineering career show an increase in terms of complexity at each semester, which represents that students who enter their university studies begin from the beginning to face the use and application of a new lexicon. As students are promoted to other semesters, the subjects to learn are new, and others correspond to the continuity of what was learned in the previous semester, which implies that students are constantly learning and using the technical vocabulary present in their studies.

In the case of the Systems Engineering career, the semesters of study correspond from the fifth semester to the eighth. The calculated index indicates a high complexity, with a value of 12,264.92, since in that semester the student learns subjects whose content involves a combination of programming languages and data that lead to the development of more complex

solutions. These involved courses are: *Database II*, *Organizational Behavior* and *Talent Human*, *Object Oriented Software Engineering*, *Artificial Intelligence*, and *Computational Organization and Architecture*.

The LC has a slight decrease in the sixth semester down to 10,129.64; in that semester the course exhibits more theoretical subjects than practical, being this *Elective III*, *Legislation in Computing*, *Financial Mathematics*, *Microprocessors*, *Simulation*.

We take as an example the texts of the seventh semester, the LC shows an increasing order and reaches the value of 11,751.32; this is because the student has subjects whose LC is between 22301.81 and 26,971.89. The subjects are: *Computer Center Administration*, *Compilers*, *Economics*, *Information Security and Distributed Operating Systems*.

In the case of the eighth semester, the LC has a quite evident decrease, decreasing to 9,975.89, it is the last semester of studies for the student, has subjects that have been learned are applied throughout their university stay, and they are subjects that are more oriented to the administrative part and its approach is directed to the development of the student's degree project, these subjects are: *Systems Auditing*, *Elective IV*, *Finance*, *MIS (Information Systems Administration)* and *Management Information Systems*.

Figure 4.15 shows the lexical complexity by subjects according to the LC metric. This allows us to demonstrate that there are subjects whose content is considerably more complex compared to others. In Figure 4.16 you can see the detail of what was specified in the previous paragraphs. The application of the lexical complexity metrics allowed evaluation of the complexity of the content of the corpus texts, determining that in a large number of subjects, the lexicon that teachers use when teaching their classes contains complex sentences, a technical language, and sophisticated causing difficulty in the understanding of students.

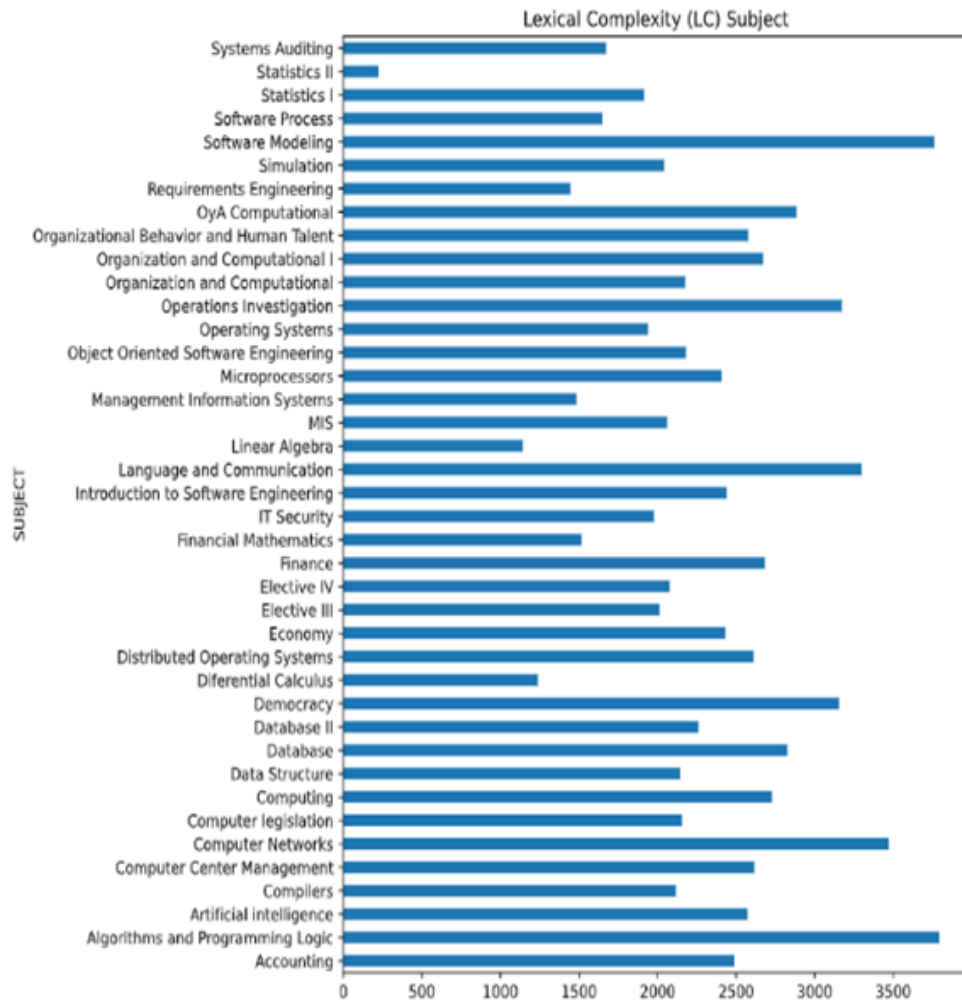


FIGURE 4.15: LC in the texts of Software Engineering and Computer Systems careers.

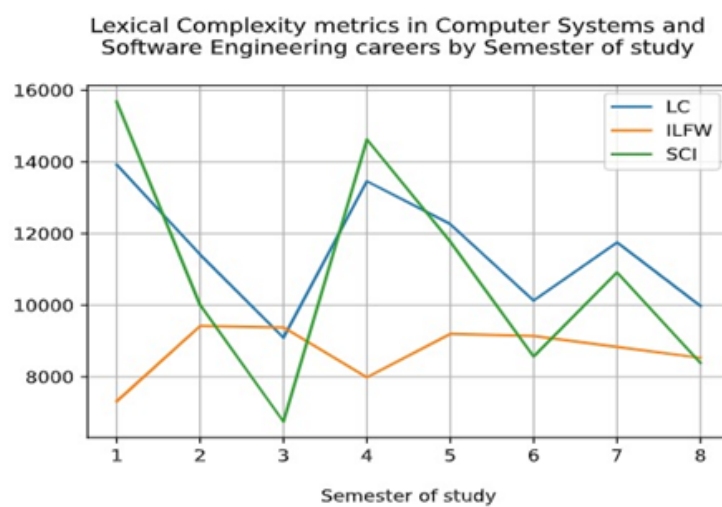


FIGURE 4.16: LC in the texts of Software Engineering and Computer Systems careers.

4.2.4 LegalEc

4.2.4.1 Motivation

With the aim of enriching the resources available to the scientific community dedicated to research on complex words in the Ecuadorian Spanish language, in the context of Lexical Simplification, we have created a new corpus called *LegalEc: A New Corpus for Complex Word Identification Research in Law Studies in Ecuadorian Spanish* (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023a). This corpus, focuses on the legal field within the academic context.

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 understand 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 legal processes and administrative in everyday life formalities without the help of an expert (Döring, 2021).

4.2.4.2 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 Guayaquil²¹. We took as a reference the works carried out on 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 75,649 students.

²¹<http://repositorio.ug.edu.ec/>

For the construction of the dataset, we followed the format used in the SemEval-2021 task 1: Lexical Complexity Prediction competition²² whose objective was to predict lexical complexity (Shardlow, Cooper, and Zampieri, 2020a). 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 enrich the data set, 23 linguistic features were included and calculated for each sentence. Importantly, we have provided a detailed description of these features in Chapter III in section 3.4.2.

Some statistics on the corpus texts are presented in Table 4.15. The Table 4.45 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.

The Statistics of <i>LegalEc</i>								
	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 4.15: Descriptive Statistics of different counters over documents in *LegalEc*.

²²SemEval-2021 - The 15th International Workshop on Semantic Evaluation. Available on <https://semeval.github.io/SemEval2021/>

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 4.16: Definition of the columns.

4.2.4.3 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 the first to third semester of studies.
- Group B (*Middle Level*): students from the fourth to sixth semester of studies.
- Group C (*Advanced Level*): students from the 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 4.17 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 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 4.17 presents several examples of the words identified and annotated as

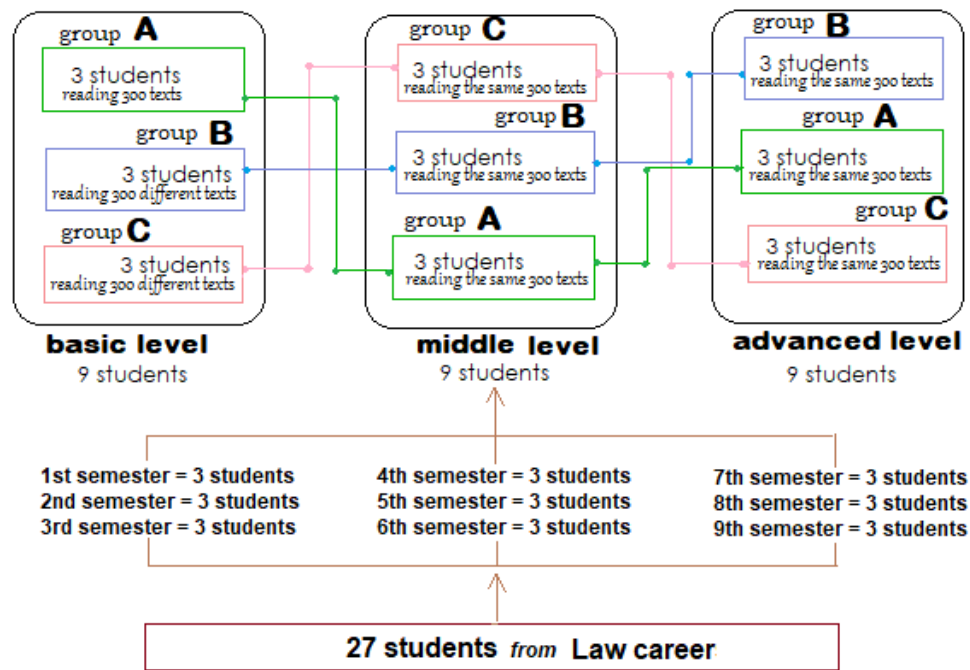


FIGURE 4.17: Applied methodology for the assignment of *LegalEc* texts.

complex in the corpus during the *LegalEc* tagging process. Next, Table ?? shows the number of words annotated by each tagger.

Note: The original text is written in Spanish, so we have included the respective translation into the working language of this thesis.

4.2.4.4 Inter-annotator Agreement

Table 4.19 shows the total number of matches obtained between taggers. 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 *Slight agreement* level according to the table of interpretations proposed by Landis and Koch, 1977. See Table 4.20.

4.2.4.5 Analysis Of Textual Complexity

The lexical complexity metrics for Spanish described in Chapter II in section 2.3.1 were applied for *LegalEc* dataset, see Table 4.22. For the analysis of the data and interpretation of the results according to the applied formulas, we consider the works by Saggion et al., 2015a and Camposa et al., 2020 and applied in our research (Ortiz-Zambrano and Montejo-Ráez, 2017),

Words tagged by the annotators in the texts of the corpus <i>LegalEc</i>		
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 4.17: Examples of words tagged by the annotators in the texts of the *LegalEc* corpus.

Words tagged by the annotators in the texts of the <i>LegalEc</i> corpus		
No. text	Sentence	Complexity
00042	<i>The issuance of advertising that induces [.] sexism, religious or political intolerance, [.]</i>	0.33
00077	<i>The State will be civilly liable for damages and [.]</i>	0.33
00078	<i>To maintain the inalienable ownership of its lands communal, which will be inalienable, [.]</i>	0.44
00152	<i>Contributions, subsidies and grants that are agreed upon in its favor by public and private institutions, [.]</i>	0.89
00230	<i>[.] and even after approval, they will have unobstructed access to justice, against any injury or harm [.]</i>	0.33
00234	<i>Article 75 of the Constitution of the Republic of Ecuador [.] and to effective, impartial, and expedited protection of its [.]</i>	0.56
00241	<i>If, through a partition act, real estate or part [.] that was previously owned in common, [.]</i>	0.89
00279	<i>Factual error vitiates consent when [.] as if one of the parties understood loan, [.]</i>	0.89

TABLE 4.18: Examples of words tagged by the annotators in the texts of the *LegalEc* corpus (translated).

(Ortiz-Zambrano and Varela Tapia, 2018), (Ortiz-Zambrano and Montejó-Ráez, 2021).

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 4.21.

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 4.19: Number words tagged as complex agreed by the different number of annotators.

Interpretation of the Fleiss Kappa statistic

range	Kappa Interpretation
1.00 – 0.81	Almost perfect agreement
0.80 – 0.61	Substantial agreement
0.60 – 0.40	Moderate agreement
0.40 – 0.21	Fair agreement
0.20 – 0.00	Slight agreement
< 0 Poor	agreement or no agreement

TABLE 4.20: Interpretation of the Fleiss Kappa statistic by Landis and Koch, 1977.

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 4.21: Interpretation of the SSR metric by Camposa et al., 2020.

For the calculation of the LC, the formula uses a list of words taken from the CREA²³ 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 the return of this formula (Camposa et al., 2020). The results of the LC are in the range of 5.86 and 53.09, which shows that the content of the texts has 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 sentence, with a result in the range of 1.79

²³<http://corpus.rae.es/lfrecuencias.html>

Lexical Complexity Metrics for Spanish in <i>LegalEc</i>								
	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 4.22: Results of the application of lexical complexity metrics for Spanish in corpus *LegalEc*.

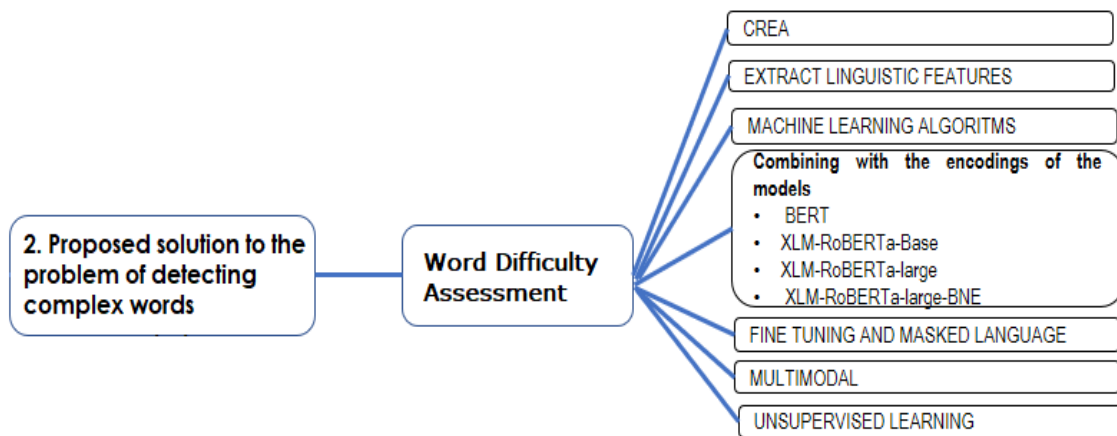


FIGURE 4.18: Representation of the explored solutions to the problem of detecting complex words.

and 75.00, demonstrating that the texts have a high complexity index at the sentence level. See Table 4.22.

4.3 Proposed Solution To The Problem Of Detecting Complex Words

After having outlined the resources generated to carry out the research, we proceed to provide a detailed description of the experiments carried out on the generated corpora, as well as on other relevant benchmarks in the context of addressing the challenge of complex word detection. Figure 4.18 presents a diagram that illustrates the previously mentioned proposal.

To address Word Difficulty Assessment, we have applied a variety of techniques, such as: using a list of frequent words in Spanish (CREA), classic machine learning algorithms, deep learning, and unsupervised learning. These techniques have been used on data sets in Spanish and English to carry out an analysis that evaluates the performance of the algorithms in different languages. Next, we will detail the techniques used in evaluating the complexity of words.

4.3.1 Complex Word Identification Based ON CREA

Based on the definitions of the experiments for the *complex word identification in Spanish* carried out by Saggion et al., 2015b and also those made by Rebollo, 2008 from the calculation of low frequency words, being those words whose frequency is less than 1,000. We built a system for our work called “*CLexIS²: A New Corpus for Complex Word Identification Research in Computing Studies*” (Ortiz-Zambrano and Montejo-Ráez, 2021) that allowed the detection of complex words contained in the corpus *CLexIS²* using CREA²⁴. Figure 4.19 shows the process for detecting complex words in Spanish in the corpus *CLexIS²* using CREA.

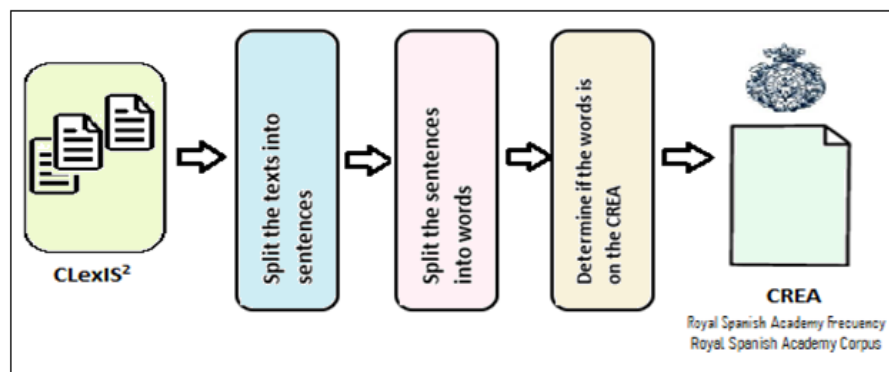


FIGURE 4.19: Process for the detection of complex words in Spanish in the corpus *CLexIS²* using CREA.

We evaluated the effectiveness of this simple approach in terms of precision, recall, F1-score and accuracy. The results show that, the proportion of predictions that the model classified correctly corresponds to an *accuracy* = 0.4394, a *precision* = 0.5165 that belong to the really correct positive identifications, the hit rate allowed us to obtain a *recall* = 0.4709, and the proportion of real negatives correctly identified whose measure is the specificity with a value = 0.3963, finally, an *F1* = 0.4927 was obtained that indicates the precision and robustness of the applied model. See Figure 4.20.

The analysis of the data determines that the Software career has a higher Lexical Complexity than the Computer Systems career. Once again, first-semester students find it difficult to move from high school to undergraduate education. The subjects that first semester students have are: Programming Algorithms and Logic, Differential Calculus, Democracy, Introduction to

²⁴CREA - Corpus of the Royal Spanish Academy. Royal Spanish Academy frequencies - Available on <http://corpus.rae.es/lfrecuencias.html>

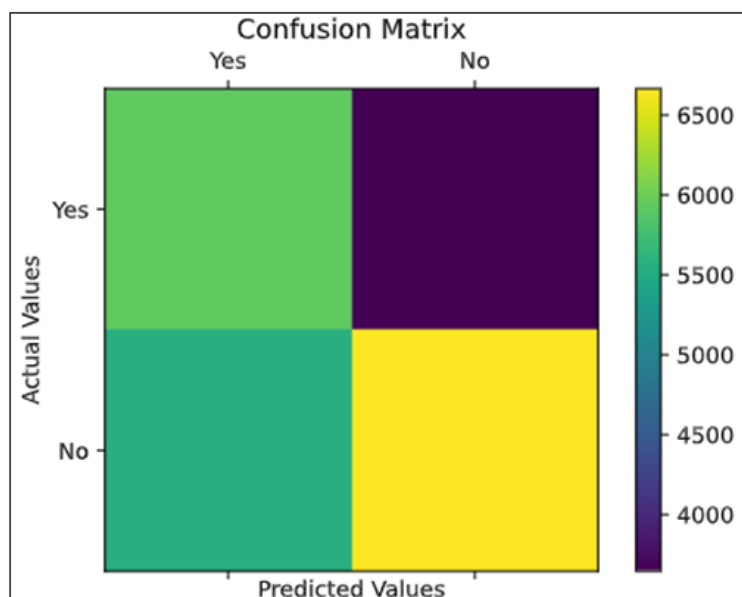


FIGURE 4.20: Relationship between the predictions of a model and the actual classes of a dataset.

Software Engineering, and Language and Communication.

In our work called “LegalEc: A New Corpus for Complex Word Identification Research in Law Studies in Ecuadorian Spanish” (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023a) for the calculation of the LC of the corpus, we also apply the solution based on the CREA as a proposal to the problem of identifying complex words for the calculation of the Lexical Complexity of the texts of the LegalEc corpus, specifically in The formula for calculating the index of low frequency words (ILFW). The results of the LC corresponded to the range of 5.86 and 53.09, which demonstrated that the content of the texts has a high level of lexical complexity since they are based on a lexicon infrequent.

4.3.2 Applying Linguistic Features With Classical Machine Learning Algorithms In English Dataset

Lexical complexity in Spanish texts represents a significant challenge for both readers and natural language processing systems. Identifying which words may be difficult to understand is essential to improve the accessibility and comprehension of texts, especially in academic and educational settings. Despite advances in this field, many of the current approaches have been developed primarily for the English language as detailed in the 1.3, which raises the need to adapt and evaluate these methodologies in the context of Spanish. To address this problem, it is proposed to apply linguistic features in combination with classical machine learning algorithms

used in English datasets, in order to determine their effectiveness in detecting complex words in Spanish texts. This methodology will allow us to evaluate whether the techniques used in other languages can be useful to solve problems specific to Spanish, and thus advance research on the topic of text simplification and accessibility improvement.

In our work called “SINAI at SemEval-2021 Task 1: Complex word identification using Word-level features” (Zambrano and Montejo-Ráez, 2021), our participation was aimed at the shared task of Lexical Complexity Prediction (LCP) hosted at SemEval 2021 (Task 1): located in the Lexical Semantics track, the task was to predict the complexity value of words in context. We prepared a system to predict word complexity with a new English dataset annotated with a Likert scale. We carried out a machine learning approach based on word frequency and a total of fifteen linguistic features calculated at the word level.

Linguistic Features

For our proposed solution, we calculated a total of 15 linguistic features at the word level, and added them to the English data set called CompLex (provided by the workshop organizing committee), taking into consideration several of the linguistic features from the work done in the research presented in the CWI in BEA 2018 (Yimam et al., 2018a) and others that correspond to other researchers. These are the features obtained on the target word (token) and are described in detail in Chapter III and section 3.4.2. The linguistic features are:

1. *Absolute frequency*
2. *Relative frequency*
3. *Word length*
4. *Number of syllables*
5. *Target word position*
6. *Number of words in the sentence*
7. *Part Of Speech (POS)*
8. *Relative frequency of the word before of the token.*

9. *Relative frequency of the word after of the token.*
10. *Length of previous word to the token.*
11. *Length of the after word to the token.*
12. *Lexical diversity (MTDL-diversity).*

Additionally, the following WordNet features Fellbaum, 2010 were also considered for each target word:

13. *Number of synonyms (number-synonyms).*
14. *Number of hyponyms (number-hyponyms).*
15. *Number of hyperonyms (number-hyperonyms).*

Classic Machine Learning Algorithm

The numeric input variables were scaled to a standard range, as many machine learning algorithms have been found to perform better when the data set is normalized. Normalization of data is a preprocessing approach that entails scaling or altering the data to ensure that every attribute contributes equally. The performance of ML algorithms depends on the data quality used. Recent research has highlighted the importance of data normalization for improving data quality and, subsequently, the performance of ML algorithms (Ali et al., 2023). A polynomial transformation on the linguistic features was then applied with a degree value of 2, so new features were created.

A forest of trees was built with the training set (X, y) , where we assigned to the independent variable (X) an array that contains all the word-level features that were obtained from the token; and the value of the dependent variable (y) corresponds to the level of complexity's word.

To build the Random Forest Regression Model, we split the dataset into the training set and test set, so 10% of the data set was used as test set, and the remaining 90% was used as the training set. Figure 4.21 shows the training process applying the random forest regression algorithm.

To calculate the prediction value of the word complexity on the data of the evaluation corpus, the (X, y) , where we assigned to the independent variable (X) which we called X_{Test} , was

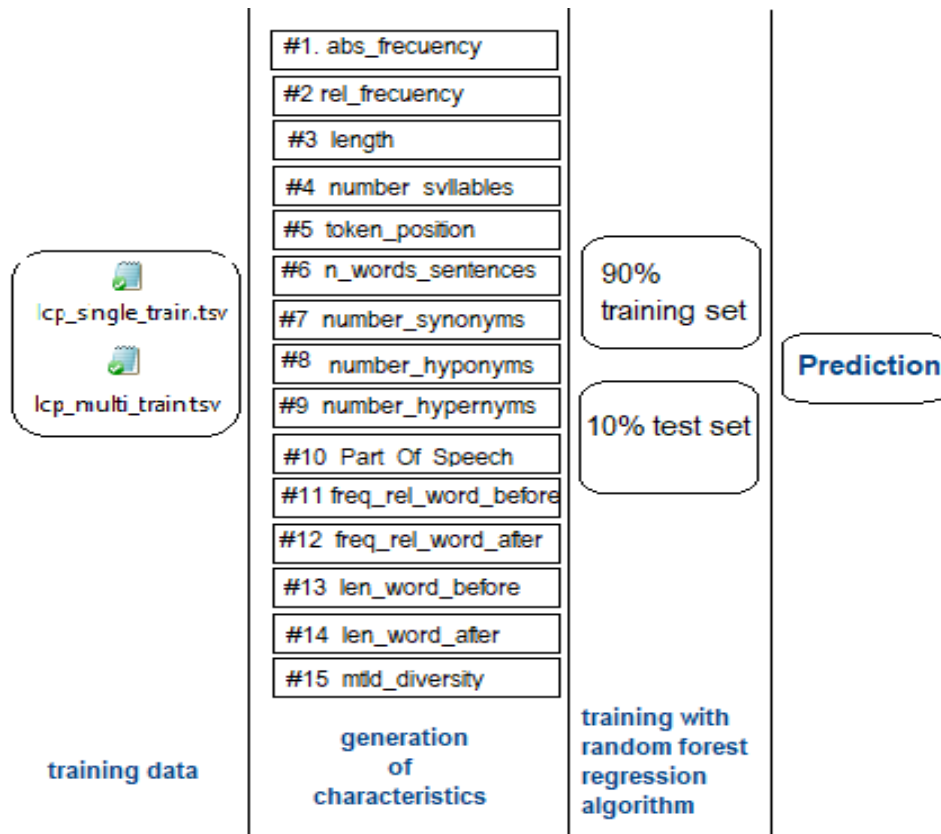


FIGURE 4.21: Prediction of lexical complexity by applying the Random Forest Regressor with linguistic features algorithm by Zambrano and Montejo-Ráez, 2021.

built, was an array that contained all the word-level features that were obtained from the token. Finally, we train the algorithm with the evaluation data and predict the results of the test set with the model trained on the testing set values using the regressor predict function.

Several runs were made with different values to observe the performance of the algorithm and fine-tune the hyperparameters of the model. Our best configuration was with 150 nodes and 7 features, selected by their F ANOVA between label / feature. The selected features were: `abs_frecuency`, `rel_frecuency`, `length`, `number_syllables`, `token_position`, `number_synonyms`, `Part_of_speech`, see Table 4.23. Finally, the prediction value of the words for the test data set was obtained, obtaining the best result: $MAE = 0.07347$, $MSE = 0.00938$, and $RMSE = 0.096871$. See Table 4.24 where we present the results obtained from our system versus the results compiled by the organizers.

The final results published in Table 4.24, present the winners of the first three positions.

#Trees	K	MAE	MSE	RMSE
150	7	0.07347	0.00938	0.09687
130	7	0.07354	0.00940	0.09700
150	8	0.07356	0.00942	0.09710

TABLE 4.23: Results obtained in the execution of the algorithm in the training set of the simple words.

#	Team Name	MAE	MSE	R^2
1	JUST_Blue	0.0609	0.0062	0.6172
2	DeepBlueAI	0.0610	0.0061	0.6210
3	Alejandro M.	0.0619	0.0064	0.6062
50	SINAI	0.0875	0.0131	0.1930

TABLE 4.24: Final results of the Lexical Complexity Prediction task on the single words dataset by Shardlow et al., 2021a.

Taking into account the number of competitors (quite large) and the result obtained by the first place winner (MAE = 0.0609) versus our result (MAE = 0.0875), we see that there is a small difference, which allows us to trust our simple approach.

4.3.3 Applying Linguistic Features With Classical Machine Learning Algorithms In Spanish Dataset

In order to address the challenge of lexical complexity in Spanish texts, we propose to apply a combination of linguistic features and classic machine learning algorithms that have been successfully used on English datasets. The goal is to evaluate the effectiveness of these techniques in identifying complex words within Spanish texts. This strategy will not only allow us to determine whether the approaches applied in other languages are transferable to Spanish, but will also offer new perspectives to advance research into text simplification and improve accessibility in various linguistic contexts.

For our solution proposal presented in *RanNLP 2021* (Ortiz-Zambrano and Montejo-Ráez, 2021), we carried out the prediction of complex words using a machine learning approach over different lexical features in a Spanish dataset.

Data from the annotated corpus *CLexIS²* was used, which is extensively detailed in section 4.2.3. A total of 15 features were generated to feed the learning algorithm, following a process similar to that used in the work carried out with *CLexIS²* corpus for the detection of complex

words process is detailed in the section 4.3.2. A supervised learning approach was applied using the Random Forest algorithm. Multiple runs with different configurations were carried out in order to evaluate the performance of the algorithm.

Numerical features were scaled to a standard range. The Random Forest algorithm was selected as the learning approach. For the construction of the Random Forest regression model, the data set was divided into: 10% used as a test set and the remaining 90% as a training set. Multiple runs with different value settings were carried out to evaluate the performance of the algorithm and tune the model hyperparameters. The test set results lead the configuration with the best scores, which consisted of 241 nodes and the 15 considered features. The results obtained from the test set highlighted the most successful configuration, which consisted of 241 nodes and 15 linguistic features. The best results achieved were: MAE = 0.174282, MSE = 0.044231, RMSE = 0.210312.

The Random Forest algorithm was run to predict lexical complexity, obtaining highly accurate results. The values generated by the algorithm show a remarkable proximity to the real values. To visualize this performance, see Figure 4.22.

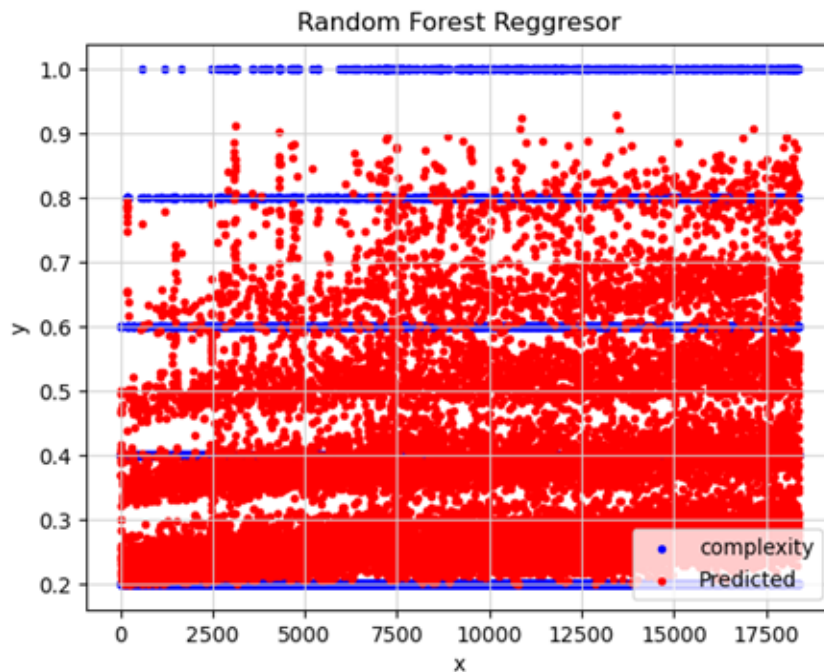


FIGURE 4.22: Prediction of word complexity applying Random Forest algorithm in Spanish dataset.

4.3.4 Applying Linguistic Features With Transformers Embeddings (pre-trained) Executed On Various Classic Machine Learning Algorithms In Spanish Dataset

In order to address the challenges associated with lexical complexity in the Spanish language, it is proposed to apply linguistic features combined with pre-trained transformer embeddings, executed on classic machine learning algorithms. This methodology will allow leveraging the representation power of transformer models together with more traditional supervised learning techniques, in order to improve the identification of complex words in Spanish texts. By implementing this strategy on a Spanish dataset, it is expected not only to increase the accuracy of the models, but also to provide more advanced and effective solutions in the field of textual simplification and accessibility.

In our proposal presented at SEPLN 2022 in our work called “Transformers for Lexical Complexity Prediction in Spanish Language” (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2022b), we highlight the relevance of evaluating the performance of classical supervised learning algorithms on multilingual data sets that are characterized by their diversity of features. We intend to demonstrate through the executions carried out to demonstrate how the combination of different types of features contributes to better performance of the models (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2022b) based on research carried out by Zaharia, Cercel, and Dascalu, 2021; Paetzold, 2021.

We will refer to the linguistic features as Hand-Crafted Features (HCF) versus the Transformers encodings coming from the models: BERT in Spanish, XLM-RoBERTa multilingual, and RoBERTa-large-BNE executed on various classic Machine Learning algorithms. This approach demonstrates the need to understand how traditional algorithms behave in diverse linguistic environments, which is essential for a complete understanding of their effectiveness and applicability in multilingual contexts.

Dataset

We carried out our research taking as a reference point the Spanish corpus known as CLexIS², the details of which are exhaustively set out in the 4.2.3 section.

Features

To obtain the morphological aspects of the text, we perform several experiments applying a total of 23 linguistic features following a similar approach to that used in an English corpus detailed in Chapter III in section 3.4.2 and combine them with the word and sentence embeddings of previously trained deep learning models. All these features were normalized with a z-score transformation before passing them to the learning algorithm.

Transformer-Based Language Models

The models were taken from the *Transformers*²⁵ library.

- The pre-trained **BERT** model that we chose was the one that the Spanish community uses mostly in research work to date, which is *bert-base-uncased* (BETO) (Canete et al., 2020).

The **BERT-base** model has the number of layers $L=12$, the hidden size $H=768$, the number of self-attention heads $A=12$, and total parameters equal to 110M.

BERT-large model has the number of layers $L=24$, the hidden size $H=1024$, the number of self-attention heads $A=16$, and total parameters equal to 335M (Conneau et al., 2020).

- The **XLNet-RoBERTa** model applied was *xlnet-roberta-base* (Conneau et al., 2020).

The **XLNet-RoBERTa-base** model has the number of layers $L=12$, the hidden size $H=768$, the number of self-attention heads $A=12$, and the total parameters equal to 270M.

XLNet-RoBERTa-large model has the number of layers $L=24$, the hidden size $H=1024$, the number of self-attention heads $A=16$, and total parameters equal to 550M (Conneau et al., 2020).

²⁵<https://huggingface.co/>

XLM-RoBERTa-large is a transformer-based masked language model for the Spanish language. It is based on the RoBERTa large model²⁶.

- The **RoBERTa-large-BNE** model used was *PlanTL-GOB-ES/roberta-large-bne* being the largest Spanish-specific model to date (Gutiérrez-Fandiño et al., 2021).

The **RoBERTa-large-BNE** model has the number of layers $L=24$, the hidden size $H=1024$, the number of self-attention heads $A=16$, and total parameters equal to 335M (Conneau et al., 2020).

Transformers Embeddings (pre-trained)

The *linguistic features* were combined with the *Transformers encoders* of the models: BERT in Spanish, XLM-RoBERTa multilingual and RoBERTa-large-BNE.

- **BERT vector:** The *bert-base-uncased* model from the Hugging Face transformer library (Wolf et al., 2020) was applied. We take all the 768-dimensional numerical representation produced by the pre-trained BERT model (Devlin et al., 2018) in the different combinations of sentence and target word encodings, for both the pre-trained model and the model fine-tuned, reaching a total of 1,559 linguistic features of different nature.
- **XLM-RoBERTa vector:** As in the case of the BERT model, we take all the 768-dimensional numerical representation produced by the pre-trained RoBERTa model (Conneau et al., 2020) in the different combinations of sentence and target word encodings, for both the pre-trained model and the model fine-tuned, reaching a total of 1,559 linguistic features of different nature.
- **RoBERTa-large-BNE vector:** Regarding this model, we take all the 1024-dimensional numerical representations produced by the pre-trained RoBERTa-large-model model (Gutiérrez-Fandiño et al., 2021), in the same way that they were applied in the previous models, the data sets were made up of the different combinations of sentence and target word encodings, for both the pre-trained model and the fine-tuned model, reaching a total

²⁶<https://huggingface.co/PlanTL-GOB-ES/roberta-large-bne>

of 2,071 linguistic features of different nature.

Next, we applied a fitted model on top of the previously trained model to demonstrate how the fitted model run on the previously trained model contributed to more accurate LCP results. Finally, the execution of the different supervised learning algorithms was applied to the training data set to evaluate which of them achieved the best prediction score. A three-fold cross-validation was performed to ensure that the partitions contained independent data for training and testing. See Figure 4.23.

Classic Machine Learning Classifiers

Similar to the work done in section 4.3.5 in the case of the algorithms, the training and evaluation of the different combinations of the sets was carried out with the same eight supervised algorithms for regression.

Several experiments were carried out for each of the data sets, where different configurations were explored for each of the algorithms. We apply the default values for the algorithms in their configurations, except for the case for tree-based algorithms where we manage to determine the best hyper-parameters with the following number of nodes:

- AdaBoost with 100 nodes.
- Random forest with 241 nodes.
- Gradient Boosting algorithm with 350 nodes.

Figure 4.23 comprehensively illustrates the process flow used to calculate lexical complexity by combining features of various natures applying classical machine learning algorithms.

Features Sets

We build several datasets made up of the combination of the features described above to run them on the pre-trained BERT model. Table 4.25 contains the description of the abbreviations.

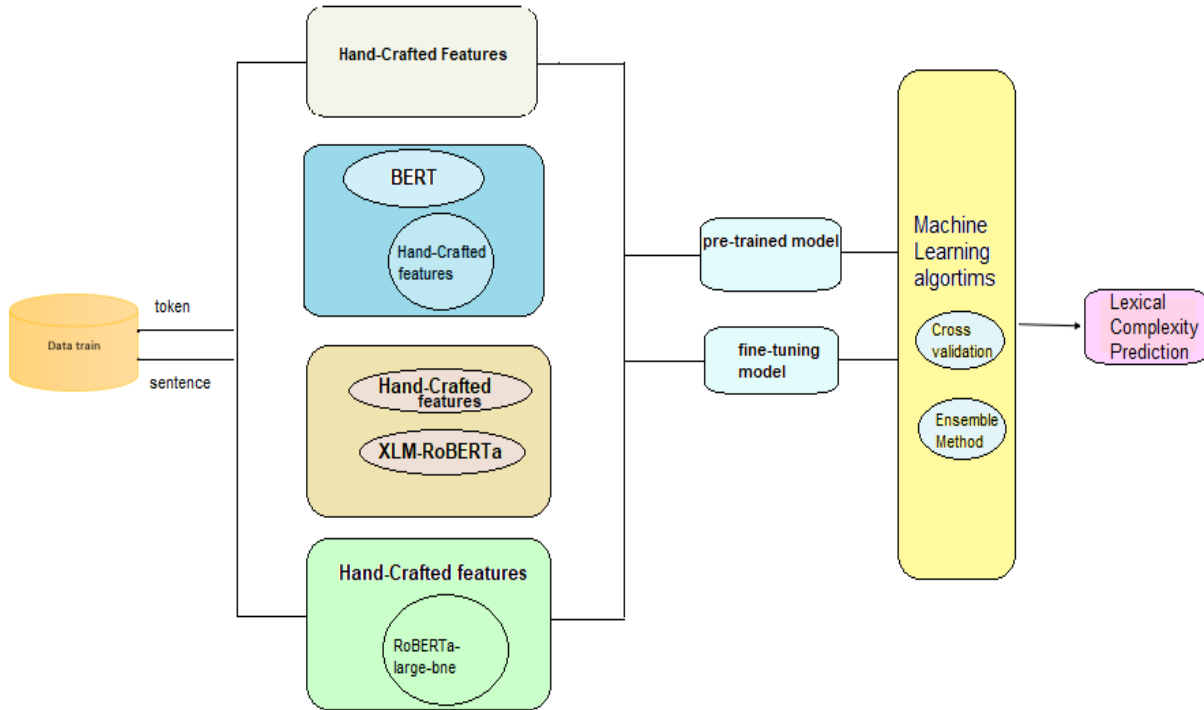


FIGURE 4.23: Representation of the workflow of the process for LCP.

Features identifier	Description
HCF	Hand-crafted features
$BERT_{sent}$	Sentence encodings from pre-trained BERT models
$BERT_{word}$	Token encodings from pre-trained BERT models
$XLMR_{sent}$	Sentence encodings from pre-trained XLM-RoBERTa model
$XLMR_{word}$	Token encodings from pre-trained RoBERTa model
$RBNE_{sent}$	Sentence encodings from pre-trained RoBERTa-large-BNE model
$RBNE_{word}$	Token encodings from pre-trained RoBERTa-large-BNE model

TABLE 4.25: Feature sets.

The details of the features is the next:

- The *Hand-Crafted Features* (linguistic features) with the features coming from the 768-dimensional vector of the initial [CLS] token as sentence embeddings ($BERT_{sent}$).
- The *Hand-Crafted Features* with the 768-dimensional vector corresponding to the target token as word embeddings ($BERT_{word}$).
- The *Hand-Crafted Features* with encodings of the [CLS] token and encodings of the target token.
- The encodings of the [CLS] token.

- The encodings of the target token.
- The encodings of the [CLS] token with the encodings of the target token.

For the evaluation of the trained and fine-tuned models, those that were widely applied to LCP for the shared LCP task hosted in SemEval 2021: Mean Absolute Error (MAE), Mean Square Error (MSE), Mean Square Error (RMSE) and Pearson correlation (Shardlow et al., 2021a).

BERT Model *Pre-trained*

Table 4.26 presents the ten best performances corresponding to different combinations of features executed with BERT pre-trained. As can be seen, the three best results in predicting lexical complexity were achieved by the ABR (AdaBoost Regressor) and SVR (Super Vector Regressor) algorithms. The best performance was achieved by the ABR algorithm presenting the best prediction for the Spanish language with an MAE = 0.1632 and a Pearson = 0.999 in the execution with the data set made up of the combination of the features generated at the sentence level and at the word level - BERT_{sent} $\oplus$ BERT_{word}.

BERT model pre-trained with CLexIS ²					
Features	Alg	MAE	MSE	RMSE	Pearson
BERT _{sent} $\oplus$ BERT _{word}	ABR	0.1632	0.0502	0.2343	0.9999
BERT _{word}	SVR	0.1634	0.0432	0.2074	0.9023
BERT _{word}	ABR	0.1643	0.0512	0.2332	0.9947
BERT _{word}	GBR	0.1653	0.0494	0.0447	0.6977
BERT _{word} $\oplus$ HFC	GBR	0.1655	0.0454	0.2088	0.7040
BERT _{sent} $\oplus$ BERT _{word} $\oplus$ HFC	GBR	0.1659	0.0418	0.2074	0.7147
BERT _{sent} $\oplus$ BERT _{word}	GBR	0.1694	0.0444	0.2039	0.7167
BERT _{sent} $\oplus$ BERT _{word} $\oplus$ HFC	ABR	0.1699	0.0554	0.2334	0.9939
BERT _{sent} $\oplus$ BERT _{word}	SVR	0.1699	0.0488	0.2037	0.8799
BERT _{word} $\oplus$ HFC	SVR	0.1704	0.0503	0.2204	0.0483

TABLE 4.26: Results of the model BERT pre-trained with features of different nature.

BERT Model *Fine-tuned*

We applied the fine-tuned BERT model on top of the pre-trained BERT model to improve the accuracy of the model to our data set. You can see in Table 4.27 the three best results resulting from the performance of the RFR (Random Forest Regressor) algorithm and the GBR (Gradient Boosting Regressor) algorithm.

The best performance was obtained with the dataset composed of the combination of the features with target word encodings together with sentence encodings from BERT fine-tuned. The same combination of features achieved the best performance in the pre-trained model but with lower results.

It should be noted that the RFR (Random Forest Regressor) algorithm does not appear within the top ten places in the execution of the pre-trained model, but it achieves its best result when the model is refined, placing first and third within the three best executions tuned. RFR presented the best prediction for the Spanish language with an MAE = 0.1592 and a Pearson = 0.988 combining $BERT_{sent} \oplus BERT_{word}$.

<i>BERT model fine-tuned with CLexIS²</i>					
Features	Alg	MAE	MSE	RMSE	Pearson
$BERT_{sent} \oplus BERT_{word}$	RFR	0.1592	0.0379	0.1982	0.9883
$BERT_{sent} \oplus BERT_{word} \oplus HFC$	GBR	0.1600	0.0367	0.1979	0.8202
$BERT_{sent} \oplus BERT_{word} \oplus HFC$	RFR	0.1610	0.0401	0.1988	0.9987
$BERT_{sent} \oplus BERT_{word} \oplus HFC$	ABR	0.1610	0.0506	0.2242	0.9998
$BERT_{sent} \oplus BERT_{word} \oplus HFC$	ABR	0.1621	0.0487	0.2300	0.9999
$BERT_{word} \oplus HFC$	GBR	0.1622	0.0430	0.1984	0.8983
$BERT_{word}$	SVR	0.1622	0.0429	0.2018	0.9183
$BERT_{word}$	GBR	0.1632	0.0472	0.0429	0.7083
$BERT_{word}$	ABR	0.1632	0.0501	0.2284	0.9967
$BERT_{word} \oplus HFC$	RFR	0.1640	0.0407	0.2100	0.9898

TABLE 4.27: Results of the model BERT tuned with features of different nature.

XLM-RoBERTa Model Pre-trained

Similar to the BERT model, the top ten sites were taken from all the runs that were done on the different data sets. The results of the three best places for the pre-trained XLM-RoBERTa model were achieved by the ABR (AdaBoots) algorithm with an MAE = 0.1623 and a Pearson 0.9973 result of the combination of the features with target word encodings together with sentence encodings from BERT fine-tuned and the HCF - $XLMR_{sent} \oplus XLMR_{word} \oplus HFC$, as can be seen in the table 4.28. It can be clearly shown that the pre-trained XLM-RoBERTa model has a better performance compared to the pre-trained BERT model, achieving a better prediction of Lexical Complexity.

XLM-RoBERTA model pre-trained with CLexIS ²					
Features	Alg	MAE	MSE	RMSE	Pearson
$\text{XLMR}_{\text{sent}} \oplus \text{XLMR}_{\text{word}} \oplus \text{HFC}$	ABR	0.1623	0.0527	0.2270	0.9973
$\text{XLMR}_{\text{word}} \oplus \text{HFC}$	ABR	0.1623	0.0513	0.2273	0.9973
$\text{XLMR}_{\text{sent}} \oplus \text{HFC}$	ABR	0.1630	0.0524	0.2293	0.9973
$\text{XLMR}_{\text{word}} \oplus \text{HFC}$	GBR	0.1653	0.0433	0.2073	0.4848
$\text{XLMR}_{\text{sent}} \oplus \text{XLMR}_{\text{word}} \oplus \text{HFC}$	GBR	0.1658	0.0434	0.2074	0.4874
$\text{XLMR}_{\text{sent}} \oplus \text{HFC}$	GBR	0.1663	0.0433	0.2082	0.4807
$\text{XLMR}_{\text{word}} \oplus \text{HFC}$	SVR	0.1680	0.0483	0.2194	0.3095
$\text{XLMR}_{\text{word}} \oplus \text{HFC}$	RFR	0.1690	0.0445	0.2093	0.9803
$\text{XLMR}_{\text{sent}} \oplus \text{HFC}$	SVR	0.1692	0.0486	0.2204	0.2713
$\text{XLMR}_{\text{sent}} \oplus \text{XLMR}_{\text{word}} \oplus \text{HFC}$	RFR	0.1692	0.0443	0.2093	0.9803

TABLE 4.28: Results of the model XLM-RoBERTA pre-trained with features of different nature.

XLM-RoBERTa Model Fine-tuned

We also highlight that in the execution of the XLM-RoBERTa fine-tuned model, it achieved a significant improvement compared to the results of the pre-trained model, reaching an MAE = 0.1601 and a Pearson = 0.998 result of the combination of the features with target word encodings together HCF - $\text{XLMR}_{\text{word}} \oplus \text{HFC}$ from XLMR fine-tuned. See table 4.29. Comparing the results of the BERT and XLM-RoBERTa both fine-tuned models, BERT tuned is so far the one that has an important performance achieved by an MAE = 0.1592 with the execution of the RFR algorithm combining $\text{XLMR}_{\text{sent}} \oplus \text{XLMR}_{\text{word}}$.

XLM-RoBERTA model fine-tuned with CLexIS ²					
Features	Alg	MAE	MSE	RMSE	Pearson
$\text{XLMR}_{\text{word}} \oplus \text{HFC}$	ABR	0.1601	0.0501	0.2251	0.9987
$\text{XLMR}_{\text{sent}} \oplus \text{XLMR}_{\text{word}} \oplus \text{HFC}$	ABR	0.1620	0.0526	0.2268	0.9987
$\text{XLMR}_{\text{sent}} \oplus \text{HFC}$	ABR	0.1620	0.0519	0.2287	0.9979
$\text{XLMR}_{\text{sent}} \oplus \text{HFC}$	GBR	0.1630	0.0420	0.2062	0.4790
$\text{XLMR}_{\text{word}} \oplus \text{HFC}$	GBR	0.1638	0.0429	0.2034	0.4800
$\text{XLMR}_{\text{sent}} \oplus \text{XLMR}_{\text{word}} \oplus \text{HFC}$	GBR	0.1652	0.0430	0.2069	0.4930
$\text{XLMR}_{\text{word}} \oplus \text{HFC}$	SVR	0.1660	0.0482	0.2172	0.3083
$\text{XLMR}_{\text{word}} \oplus \text{HFC}$	RFR	0.1669	0.0427	0.2013	0.9849
$\text{XLMR}_{\text{sent}} \oplus \text{HFC}$	SVR	0.1682	0.0487	0.2202	0.2729
$\text{XLMR}_{\text{sent}} \oplus \text{XLMR}_{\text{word}} \oplus \text{HFC}$	SVR	0.1688	0.0482	0.2191	0.2932

TABLE 4.29: Results of the model XLM-RoBERTA fine-tuned with features of different nature.

RoBERTa-large-BNE Model Pre-trained

Runs were also performed with the pre-trained and fine-tuned RoBERTa-large-BNE model. The ten best results are displayed in the table 4.31. The three best positions in predicting lexical

complexity were achieved by the ABR (AdaBoost) and GBR (Gradient Boost) Regressor algorithms, with ABR reaching an MAE = 0.1609 and a Pearson = 0.6754 combining the sentence encodings together with the target word encodings and the HCF - RBNE_sent $\oplus$ RBNE_word $\oplus$ HFC.

It should be noted that the pre-trained model RoBERTa-large-BNE is the one that achieves a better prediction for lexical complexity in the Spanish language compared to the pre-trained models BERT and XLM-RoBERTa, reaching BERT an MAE = 0.1632, XLMR an MAE = 0.1623 and RBNE an MAE = 0.1609. See Table 4.30.

The Spanish Language Models Pre-Trained Best Result				
Model	Features	Alg	MAE	
RBNE	RBNE _{sent} $\oplus$ RLBNE _{word} $\oplus$ HFC	ABR	0.1609	
XLMR	XLMR _{sent} $\oplus$ XLMR _{word} $\oplus$ HFC	ABR	0.1623	
BERT	BERT _{sent} $\oplus$ BERT _{word}	ABR	0.1632	

TABLE 4.30: Best results with pre-trained models.

RoBERTa-large-BNE model pre-trained with CLexIS ²						
Features	Alg	MAE	MSE	RMSE	Pearson	
RBNE _{sent} $\oplus$ RBNE _{word} $\oplus$ HFC	ABR	0.1609	0.0421	0.2047	0.6754	
RBNE _{sent} $\oplus$ RBNE _{word}	ABR	0.1675	0.0556	0.2347	0.9952	
RBNE _{sent} $\oplus$ RBNE _{word} $\oplus$ HFC	GBR	0.1691	0.0434	0.2073	0.6607	
RBNE _{word}	ABR	0.1693	0.0563	0.2360	0.9948	
RBNE _{word}	GBR	0.1696	0.0447	0.2101	0.6400	
RBNE _{sent} $\oplus$ RBNE _{word}	GBR	0.1698	0.0447	0.2102	0.6450	
RBNE _{sent} $\oplus$ RBNE _{word} $\oplus$ HFC	SVR	0.1708	0.0507	0.2224	0.2363	
RBNE _{sent} $\oplus$ HFC	SVR	0.1708	0.0507	0.2224	0.0857	
RBNE _{word}	SVR	0.1708	0.0507	0.2224	0.0390	
RBNE _{sent}	SVR	0.1708	0.0507	0.2224	0.0052	

TABLE 4.31: Results of the RoBERTa-large-BNE pre-trained model with features of different nature.

RoBERTa-large-BNE Fine-tuned Model

Executing the RoBERTa-large-BNE tuned model, the results are encouraging, there is an improvement compared to the results of the pre-trained model. The first places were reached by the GBR-Gradient Boosting Regressor and SVR-Super Vector Regressor algorithms. It presents an acceptable improvement, achieving in its performance an MAE = 0.1609 and a Pearson = 0.6754 combining the sentence encodings together with the target word encodings and the HCF - RLBNE_sent $\oplus$ RLBNE_word $\oplus$ HFC, and the second and third places prove it in comparison

with the pre-trained model. See Table 4.32.

RoBERTa-large-BNE model fine-tuned with CLexIS ²					
Features	Alg	MAE	MSE	RMSE	Pearson
$\text{RBNE}_{\text{sent}} \oplus \text{RBNE}_{\text{word}} \oplus \text{HFC}$	GBR	0.1609	0.0421	0.2047	0.6754
$\text{RBNE}_{\text{sent}} \oplus \text{RBNE}_{\text{word}} \oplus \text{HFC}$	SVR	0.1630	0.0435	0.2070	0.4883
$\text{RBNE}_{\text{word}} \oplus \text{HFC}$	SVR	0.1666	0.0466	0.2136	0.4220
$\text{RBNE}_{\text{word}}$	ABR	0.1677	0.0551	0.2336	0.9952
$\text{RBNE}_{\text{sent}} \oplus \text{RBNE}_{\text{word}}$	SVR	0.1684	0.0472	0.2152	0.4425
$\text{RBNE}_{\text{sent}} \oplus \text{RBNE}_{\text{word}} \oplus \text{HFC}$	GBR	0.1686	0.0432	0.2067	0.6854
$\text{RBNE}_{\text{word}}$	SVR	0.1686	0.0468	0.2146	0.5021
$\text{RBNE}_{\text{sent}} \oplus \text{RBNE}_{\text{word}} \oplus \text{HFC}$	ABR	0.1689	0.0558	0.2351	0.9951
$\text{RBNE}_{\text{word}}$	GBR	0.1690	0.0442	0.2090	0.6585
$\text{RBNE}_{\text{sent}} \oplus \text{RBNE}_{\text{word}}$	ABR	0.1699	0.0598	0.2391	0.9891

TABLE 4.32: Results of the RoBERTa-large-BNE model fine-tuned with features of different nature.

It should be noted that the pre-trained model RoBERTa-large-BNE is the one that achieves a better prediction for lexical complexity in the Spanish language compared to the tuned models BERT and XLM-RoBERTa, reaching BERT an MAE = 0.1592, XLM-R an MAE = 0.1601 and RBNE an MAE = 0.1609. See Table 4.33.

The Spanish Language Models Fine-Tuned Best Result			
Model	Features	Alg	MAE
BERT	$\text{BERT}_{\text{sent}} \oplus \text{BERT}_{\text{word}}$	RFR	0.1592
XLMR	$\text{XLMR}_{\text{word}} \oplus \text{HFC}$	ABR	0.1601
RBNE	$\text{RBNE}_{\text{sent}} \oplus \text{RBNE}_{\text{word}} \oplus \text{HFC}$	GBR	0.1609

TABLE 4.33: Best results with fine-tuned models.

It can be seen that the models based on adjusted Transformers make an important contribution to the Prediction of Lexical Complexity in the Spanish language. Table 4.34 presents the best five best results of all the experiments carried out with the models both pre-trained and adjusted, where the adjusted models are those that achieved the best performance.

The BERT tuned model with the combination of the features executed on the RFR-Random Forest Regressor algorithm reached first place with an MAE = 0.1592 and a Person = 0.9883, the feature is $\text{RBNE}_{\text{sent}} \oplus \text{RBNE}_{\text{word}}$.

Summary of best results on the CLexIS ² corpus						
Model	Features	Alg	MAE	MSE	RMSE	Pearson
BERT _{<i>fine-tuned</i>}	BERT _{<i>sent</i>} $\oplus$ BERT _{<i>word</i>}	RFR	0.1592	0.0379	0.1982	0.9883
BERT _{<i>fine-tuned</i>}	BERT _{<i>sent</i>} $\oplus$ BERT _{<i>word</i>} $\oplus$ HFC	GBR	0.1600	0.0367	0.1979	0.8202
XLMR _{<i>fine-tuned</i>}	XLMR _{<i>word</i>} $\oplus$ HFC	ABR	0.1601	0.0501	0.2251	0.9987
RBNE _{<i>fine-tuned</i>}	RBNE _{<i>sent</i>} $\oplus$ RBNE _{<i>word</i>}	GBR	0.1609	0.0421	0.2047	0.6754
BERT _{<i>fine-tuned</i>}	BERT _{<i>sent</i>} $\oplus$ BERT _{<i>word</i>} $\oplus$ HFC	RFR	0.1610	0.0401	0.1988	0.9987

TABLE 4.34: Summary of best results on the CLexIS² dataset.

4.3.5 Applying Linguistic Features With Transformers Embeddings (pre-trained) Executed On Various Classic Machine Learning algorithms In English dataset

In particular, English texts present lexical challenges that can make them difficult to interpret for non-specialist readers or those with language limitations (Anggia and Habók, 2023). To address this issue, we propose to apply a combination of linguistic features and embeddings from pre-trained transformers, executed on various classical machine learning algorithms. This technique allows to take advantage of the semantic representation power of transformers, while optimizing the results using traditional machine learning methods. By applying this approach to an English dataset, we aim to improve the accuracy in the detection of complex words, thus providing an effective tool to simplify texts.

This research focuses on the evaluation of the prediction of complex words in the English language through the application of technological solutions with emphasis on Deep Learning models. Our challenge is to improve the prediction results of the lexical complexity task presented in the SemEval 2021 competition with the CompLex 2.0 corpus. Our approach leverages the combination of several classical machine learning algorithms and their integration into Transformer-based deep learning models: BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020).

Data

The data set is made up of features of different natures: embeddings, linguistic, syntactic, statistical, and semantic. We implement a fine-tuned model executed on a previously trained model for which we execute several of the recommendations applied by the different winning teams of the competition (Shardlow et al., 2021a), especially, the approach applied by (Zaharia, Cercel, and Dascalu, 2021).

The results obtained in our solution proposal, in our research called *“Combining Transformer Embeddings with linguistic features for Complex Word Identification”* (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2022a), highlight the usefulness of linguistic features in Natural Language Processing (NLP) tasks, reaffirming our hypothesis by demonstrating that these features can effectively improve the performance of deep learning systems. We have demonstrated through the study carried out, the combination of tuned and pre-trained transformer networks with different types of traditional systems linguistic features. We apply these features on classical machine learning classifiers.

System Description

Our approach is based on the combination of encodings from pre-trained language models with different linguistic features to produce a final vector that is fed into some classic machine learning algorithms, as is graphically described in Figure 4.24. As can be seen, from a given text, two feature vectors are generated: the embeddings from the transformer model and the linguistic features computed from the text (normalized with a Z-score transformation). These vectors are combined and serve as input to the classification algorithm.

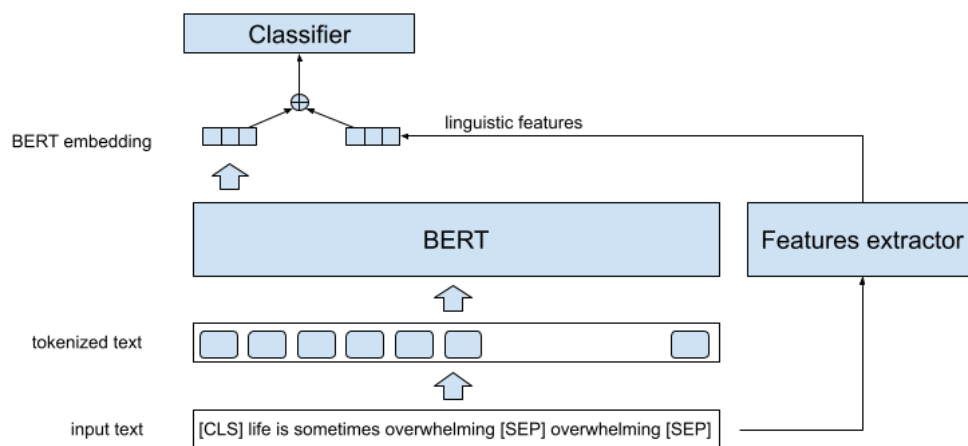


FIGURE 4.24: Architecture of the system.

As research in recent times has shown that Transformers achieve significant performance for most NLP tasks (Wolf et al., 2020), we take the initiative applied by the top-ranking teams in the SemEval task 2021, and select these two well-known large language models:

- BERT represents the baseline of Transformer-based models. This model has been extensively pre-trained in English (Devlin et al., 2019).

- RoBERTa usually performs better on downstream tasks, as it improves upon BERT by modifying key hyperparameters and trains with mini-batches and higher learning rates (Zhuang et al., 2021).

The BERT and RoBERTa models are fine-tuned with final dense layers as the classifier network using the architecture used by default in the sequence classification version of these models, as it is the Hugging Face implementation of these architectures. No hyperparameter exploration is done, so, again, the default parametrization from the taken implementation is applied. Regarding classic machine learning methods, they are trained with the default parameters as set by Scikit-learn implementations. Various machine learning algorithms have been evaluated to train the final classifier. The features considered are detailed below, as well as the classic machine learning models used in this analysis.

Features

We conducted various experiments using a total of 23 linguistic features detailed extensively in Chapter III in section 3.4.2, and combined these features with word and sentence embeddings from pre-trained deep learning models. This approach is similar to our research presented in in section 4.3.4 only with the difference that now we uses an English corpus. Before entering them into the learning algorithm, all of these features were normalized using a z-score transformation. Linguistic features are computed as classical features in computational linguistics (Liebeskind, Elkayam, and Liebeskind, 2021; Mosquera, 2021a), and “semantic features” are those that result from the word- and sentence-level embeddings described above. Linguistic features have been shown to be a major contributor to the prediction of lexical complexity; therefore, we rely on the applied features (Zambrano and Montejo-Ráez, 2021).

Subsequently, we carried out a series of runs where linguistic features were combined with word and sentence embeddings generated by previously pre-trained or fine-tuned deep learning models, following the methodology proposed by Liebeskind, Elkayam, and Liebeskind, 2021. In this context, we obtained sentence-level embeddings of the token source as well as word-level embeddings. This is based on the idea that contextual token information is essential to improve the suitability of word-level embeddings that may be suitable for LCP (Zaharia, Cercel, and Dascalu, 2021).

Classic Machine Learning Classifiers

The algorithms that were used with the various datasets are detailed below, which made it possible to evaluate their performance and determine which of them exhibits better performance based on the LCP results:

1. AdaBoost — AB (Paetzold, 2021).
2. Decision Tree — DT (Shardlow, Evans, and Zampieri, 2022).
3. Gradient Boost — GB (Vettigli and Sorgente, 2021).
4. Stochastic Gradient — SG (Bottou, 2010).
5. K-Nearest Neighbor — KNN (Liebeskind, Elkayam, and Liebeskind, 2021).
6. Support Vector Machines — SVM (Liebeskind, Elkayam, and Liebeskind, 2021).
7. Passive Aggressive — PA (Crammer et al., 2006).
8. Random Forest — RF (Zaharia, Cercel, and Dascalu, 2021), (Desai et al., 2021a).

We explored the BERT and RoBERTa models, which were pre-trained and fine-tuned, encoding both the token and the sentence. We implement various training strategies, considering previously trained or fine-tuned models, with or without the inclusion of linguistic features. Additionally, approaches involving encoding the target word or the entire sentence, among other variants, were explored. The results of these different executions are presented in Table 4.36. On the other hand, in Table 4.35, we provide a detailed description of the abbreviations used to improve the understanding of the features used in our experiments.

Feature Identifier	Description
LF	Linguistic Features.
BERT _{sent}	Sentence encodings from BERT model.
BERT _{word}	Token/word encodings from BERT model.
XLMR _{sent}	Sentence encodings from XLM-RoBERTa model.
XLMR _{word}	Token/word encodings from RoBERTa model.

TABLE 4.35: Description of the explored feature sets.

Results for the Deep Learning Approaches with CompLex							
Configuration	Model	Alg	MAE	MSE	RMSE	Pearson	R2
BERT-W $\oplus$ BERT-S $\oplus$ LF	<i>fine-tuned</i>	SVR	0.068898	0.009908	0.094294	0.891137	0.80
BERT-W $\oplus$ BERT-S $\oplus$ XLMR-W $\oplus$ XLMR-S $\oplus$ LF	<i>fine-tuned</i>	SVR	0.068899	0.009908	0.094296	0.8911367	0.80
BERT-W $\oplus$ BERT-S $\oplus$ XLMR-W $\oplus$ XLMR-S $\oplus$ LF	<i>pre-trained</i>	SVR	0.068899	0.009908	0.094296	0.891136	0.79
BERT-W $\oplus$ BERT-S $\oplus$ LF	<i>fine-tuned</i>	GBR	0.068972	0.011898	0.099297	0.927208	0.87
BERT-W $\oplus$ BERT-S $\oplus$ XLMR-W $\oplus$ XLMR-S $\oplus$ LF	<i>fine-tuned</i>	GBR	0.068974	0.011898	0.099297	0.927218	0.86
BERT-W $\oplus$ BERT-S $\oplus$ XLMR-W $\oplus$ XLMR-S $\oplus$ LF	<i>pre-trained</i>	GBR	0.068974	0.011898	0.099297	0.927218	0.87
BERT-W $\oplus$ LF	<i>fine-tuned</i>	SVR	0.069900	0.009913	0.094392	0.890911	0.79
BERT-W $\oplus$ LF	<i>fine-tuned</i>	GBR	0.070124	0.012018	0.099372	0.926726	0.86
BERT-W $\oplus$ BERT-S	<i>fine-tuned</i>	SVR	0.071623	0.009204	0.095901	0.874019	0.77
BERT-W $\oplus$ LF	<i>pre-trained</i>	SVR	0.074342	0.009909	0.095558	0.864029	0.75
BERT-W $\oplus$ BERT-S	<i>pre-trained</i>	SVR	0.074394	0.009323	0.096034	0.873943	0.75
BERT-W $\oplus$ LF	<i>pre-trained</i>	GBR	0.075430	0.012348	0.100003	0.900043	0.80
BERT-W	<i>pre-trained</i>	RLM	0.075552	0.009734	0.098232	0.789901	0.63
BERT-W	<i>fine-tuned</i>	SVR	0.075897	0.009520	0.097119	0.864816	0.76
BERT-W	<i>pre-trained</i>	SVR	0.075938	0.009559	0.097123	0.864000	0.76

TABLE 4.36: Results for different feature combinations, models, and classifiers.

We evaluated the results of our system by applying the regression metrics in the execution of the supervised learning algorithms, specifically MAE, MSE, RMSE, Pearson, and R2. We emphasize that we apply the methodologies of the winning teams, which are based on the application of language models based on the pre-trained and adjusted Transformers BERT and RoBERTa (Song et al., 2021; Yaseen et al., 2021; Taya et al., 2021), together with the linguistic, syntactic and statistical features, and the embedding results at the word and sentence level.

The Table 4.36 presents the 15 best results. Only the GBR (Gradient Booting Regressor) and SVR (Super Vector Regressor) algorithms appear in the first places. The SVR algorithm is the one with the best performance for predicting the complexity of simple words, the results are: MAE = 0.068898, MSE = 0.009908, RMSE = 0.094294. In the publication “Combining transformer embeddings with linguistic features for the identification of complex words” (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2022a), we demonstrate through the study, the combination of tuned and pre-trained transformer networks with different types of traditional systems linguistic features. We apply these features on classical machine learning classifiers. Our best results are obtained by applying Super Vector Regressor on an English corpus in an LCP task solved as a regression problem obtaining a Mean Absolute Error (MAE) = 0.068898 and Pearson = 0.8911.

4.3.6 Applying Linguistic Features With Transformers Embeddings (tuned) In Spanish And English Corpora (End-to-End Approach)

On this occasion we have designed a solution proposal based on the End-To-End approach using as a first instance the Spanish data set called LegalEc (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023a), from which we extracted 23 linguistic features together with the encodings generated by various Transformer-based models designed for Spanish, such as XLM-RoBERTa-large, XLM-RoBERTa-base and RoBERTa-large-BNE (MarIA), both of which have been pre-trained and are widely recognized for their outstanding contributions. These models have been successfully used in the creation of cutting-edge solutions for various tasks (Paetzold, 2021). LegalEc has been described in detail in the corresponding section 4.2.4 of this Chapter.

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 4.25.

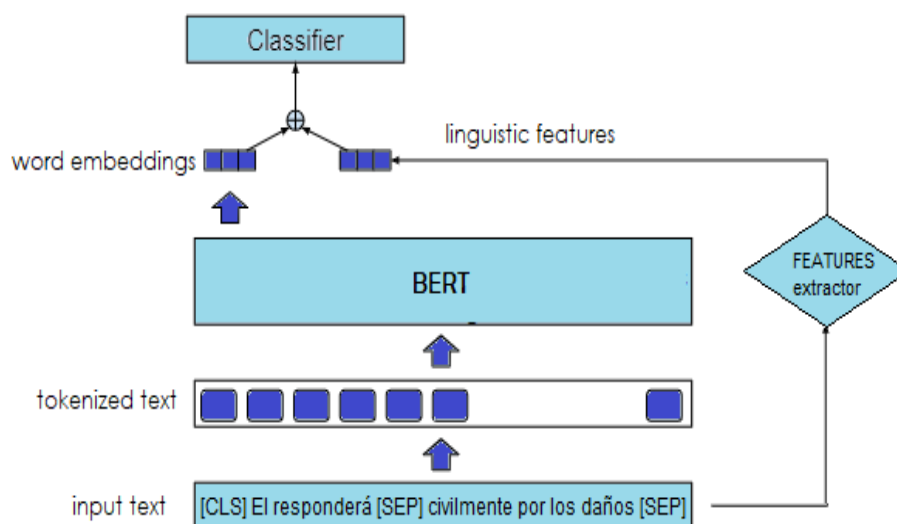


FIGURE 4.25: Process flow methodology integrating linguistic features with embeddings from the Transformer model.

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-large, or RoBERTa-large-BNE), the sentence embedding is concatenated with a min-max scaler vector of the linguistic features.
3. The resulting vector enters 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.

Executions

We have applied the metrics used to evaluate the different configurations which are: Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), and Pearson's correlation coefficient.

In the experiments, we applied 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 training was 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 4.37. The best result was achieved by the RoBERTa-large-BNE model with an MAE = 0.1560.

<i>Spanish Language Model pre-trained with LegalEc</i>						
Model	with 10 epochs					
	MAE	MSE	RMSE	R2	Poisson	Pearson
XLM-RoBERTa-Base	0.2286	0.0710	0.2665	-0.0010	0.3970	0.0323
XLM-RoBERTa-Base $\oplus$ 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 $\oplus$ 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 $\oplus$ LF	0.1770	0.0682	0.2612	-0.0821	0.4388	-0.0400

TABLE 4.37: Results of the pre-trained models applying 10 epochs.

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 an MAE of 0.1341; RoBERTa-large-BNE + LF an MAE of 0.1363, and XLM-RoBERTa-large + LF an MAE of 0.1360. See Table 4.38 for detailed results.

<i>Spanish Language Model pre-trained with LegalEc</i>						
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 $\oplus$ 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 $\oplus$ 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 $\oplus$ LF	0.1360	0.0256	0.1600	0.6614	0.1632	0.9959

TABLE 4.38: Results of the pre-trained models applying 30 epochs.

In view of the fact that the improvement of the results was evident when increasing the epochs, we executed one more training over 50 epochs. The prediction error reached by RoBERTa-large-BNE + LF is an MAE = 0.1349, and XLM-RoBERTa-large + LF with an MAE = 0.1338, achieving an improvement over previous runs. See Table 4.49.

<i>Spanish Language Model pre-trained with LegalEc</i>						
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 $\oplus$ 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 $\oplus$ 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 $\oplus$ LF	0.1338	0.0252	0.1587	0.6611	0.1594	0.9950

TABLE 4.39: Results of the pre-trained models applying 50 epochs.

The Best Results of the execution of the pre-trained models through the MAE metrics are presented below:

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 4.26. We hypothesize that the wealth of knowledge present in transformer-based models can help in extracting complementary clues of contextual complexity (Paetzold, 2021).

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 $\oplus$ LF	0.1642	0.1341	0.1382
RoBERTa-large-BNE	0.1560	0.2284	0.2085
RoBERTa-large-BNE $\oplus$ LF	0.1580	0.1363	0.1349
XLM-RoBERTa-large	0.2100	0.2360	0.2375
XLM-RoBERTa-large $\oplus$ LF	0.1770	0.1360	0.1338

TABLE 4.40: Best Results of the execution of the pre-trained models through the MAE metric.

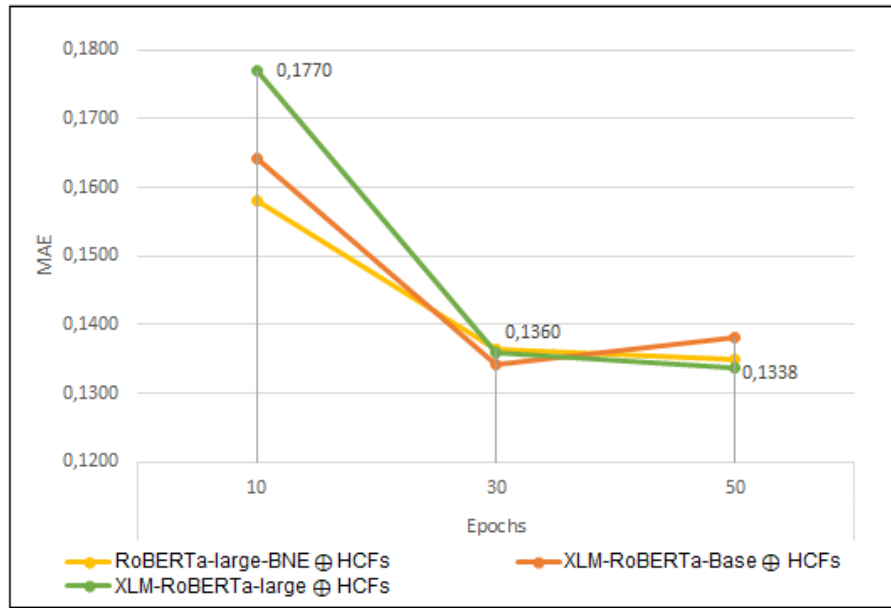


FIGURE 4.26: Integration of HCF and increase in times.

When evaluating different combinations of features, it was observed that in the Spanish corpus CLexIS² the best performance was achieved with a MAE of 0.1323 using the fine-tuned version of the Spanish BERT model. Table 4.41, which summarize the performance of various models based on the Transformers architecture after being tuned.

Transformers Embeddings fine-tuned End-To-End approach CLexIS ² corpus				
Feature	MAE	MSE	RMSE	Pearson
BERT	0.1323	0.0368	0.1918	0.5305
BERT + HFCs	0.1361	0.0387	0.1967	0.5257
XLM-RoBERTa + HCFs	0.1790	0.0478	0.2187	0.0228
XLM-RoBERTa	0.1801	0.0489	0.2211	-0.0097

TABLE 4.41: Summary of the best results on the CLexIS² corpus.

On the other hand, in the English CompLex corpus, where a MAE of 0.0683 was obtained, BERT fine-tuning was again highlighted as the optimal option, as long as additional linguistic

features were incorporated into the model. Detailed results of these evaluations are presented in Tables 4.42, which summarize the performance of various models based on the Transformers architecture after being tuned.

Transformers Embeddings fine-tuned End-To-End approach				
CompLex corpus				
Feature	MAE	MSE	RMSE	Pearson
BERT + HCFs	0.0683	0.0077	0.0876	0.7452
BERT	0.0834	0.0109	0.1046	0.6539
XLM-RoBERTa + HFCs	0.0951	0.0157	0.1254	0.2029
XLM-RoBERTa	0.1000	0.0165	0.1284	-0.0830

TABLE 4.42: Summary of the best results on the CompLex corpus.

The most notable results of the execution of Transformers Embeddings (tuned), combining linguistic features in the spanish and english datasets, were evaluated by calculating the Mean Absolute Error (MAE), as shown in Table 4.43.

The Transformers Embeddings fine-tuned End-To-End approach					
Final Results					
Model	Corpus	language	MAE	MSE	RMSE
BERT	CLexIS ²	Spanish	0.1323	0.0368	0.1918
XLM-RoBERTa-large $\oplus$ LF	LegalEc	Spanish	0.1338	0.0252	0.1587
BERT + HFCs	CompLex	English	0.0683	0.0077	0.0876

TABLE 4.43: Results of the fine-tuned end-to-end approach.

Note: The data presented in the Table 4.43 correspond to those that will be published in the article “Deep Encodings vs. Linguistic Features in Lexical Complexity Prediction”, except the data that correspond to the LegalEc corpus (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023a).

We created a Spanish corpus that we have named GovAIEc, a new annotated complex lexical corpus created with institutional texts in Ecuadorian Spanish. We submitted our contribution to the 2024 Conference on Empirical Methods in Natural Language Processing LREC-2024 and is under review. Our goal is to provide 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 literacy levels, in order to ensure a comprehensive evaluation. We used Lexical Complexity metrics as units of analysis, and

applied 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 to the reading comprehension of users who interact with bureaucratic procedures of various entities in Ecuador.

For many people, the way a text is written can become an obstacle to understanding its content (Saggion et al., 2015a). Public institutions are not exempt from this reality, many individuals encounter significant obstacles in understanding texts related to public administration (Yuan et al., 2023). 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, Trussel, and Davenport, 2023.

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 Municipality of Guayaquil* (GMO)²⁷. 2) *The Internal Revenue Service* (IRS)²⁸. 3) *The National Telecommunications Corporation* (NTC)²⁹. 4) *The National Electoral Council* (NEC)³⁰. 5) *The Municipal Transit Authority* (MTA)³¹. 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 for 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.

²⁷Illustrious 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/>

²⁸Internal 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>

²⁹National 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/>

³⁰National 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/>

³¹Municipal 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/>

The contributions of this research can be summarized as follows:

- A new corpus consists of 1,500 texts in Spanish which we have called GovAIEc. 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.

Measures Of Lexical Complexity For Spanish

These metrics have been used in previous research on text simplification in Spanish, and have also been applied in our own work as a result of our research. They are detailed in detail in Chapter II in section 2.3.1. Among the most notable studies are those by Saggion et al. (2015a), Ortiz-Zambrano and Montejo-Ráez (2017), Ortiz-Zambrano and Montejo-Ráez (2021), Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez (2023a), among others.

The GovAIEc Corpus

GovAIEc 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. GovAIEc has a total of 7,813 complex words identified and a total of 12,095 annotations. The objective of GovAIEc 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 users who participate in administrative processes. Secondly, it provides a valuable resource for the scientific community, allowing progress in research within the field of Lexical Simplification in the Spanish language.

Each sample of the GovAIEc 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:** The presence of linguistic features has been shown to improve the accuracy of large linguistic models by carefully selecting and extracting features relevant to specific tasks, since without these essential features, the algorithm does not work properly and the result also decreases. It is about feeding the model and making it more fruitful.

Feature engineering through linguistic feature extraction has allowed combining linguistic features with encodings in the data sets to be trained with more robust models and achieving important results in the prediction of word complexity thanks to the models and the presence of linguistic features, achieving an improvement in their performance and the effectiveness of natural language processing systems, as well as their relevance in the continuous advancement of research in the field of Lexical Simplification, specifically in the study of the identification and prediction of complex words that cause barriers in reading comprehension in our case in the Spanish language. Research has shown that the incorporation of linguistic features and the increase in the number of training epochs can significantly improve the performance of models (Ortiz-Zambrano and Montejo-Ráez, 2021).

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, Anke, Saggion, et al., 2016, Shardlow, Cooper, and Zampieri, 2020a, Paetzold, 2021, Mosquera (2021b), Desai et al., 2021b, Shiroyama, 2022, among others. We use the following linguistic features: the absolute frequency, the number of characters of the token, the relative frequency of

the word before the token, the relative frequency of the word after the token, the relative frequency of the token, the number of syllables, the position of the target word in the sentence, the number of words in sentence, the number of characters in the word before the token, the number of characters in the word after the token, the Lexical diversity, the number of synonyms, the number of hyponyms, the number of hyperonyms, the Part Of Speech category, the number of nouns, singular or massive, the number of auxiliaries verbs, the number of adverbs, the number of symbols, the number of numeric expressions, the number of verbs, the number of nouns, and the number of pronouns.

Annotation Process

A graphical user interface (GUI) was created using the Tkinter library. 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*. By “neutral” is a moderately difficult category, referring to words that are neither difficult nor extremely difficult, allowing us to make a more granular assessment of difficulty and thus improving the accuracy of labeling complex words.

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 ?? 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.

Results

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

Firstly: We applied the Lexical Complexity metrics for Spanish.

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

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

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 4.44, while the definition of the variables is shown in table 4.45. It is notable that the number of rare words (N_{rw}) is considerably greater than that of less frequent words (N_{lfw}). Table 4.46 presents several examples of the words identified and annotated as complex in the corpus during the GovAIEc tagging process.

The Statistics of GovAIEc								
	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 4.44: Descriptive Statistics of different counters over documents in GovAIEc.

Inter-annotator Agreement

A total of 7,813 complex words identified and a total of 12,095 annotations in the 1,500 texts of the GovAIEc corpus. We applied the Fleiss-Kappa coefficient and we obtained a

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 4.45: Definition of the columns in Table 4.44.

Words tagged by the annotators in the texts of the corpus GovAIEc		
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
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

TABLE 4.46: Examples of words tagged by the annotators in the texts of the GovAIEc corpus.

value of 0.165, which indicates a slight level of agreement. 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. The table 4.47 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, while annotators of the high level have recorded a smaller number of complex words compared to the annotators of the previous groups.

Application Of Complexity Measures For Spanish

In the second phase of results, we evaluate the GovAIEc corpus using seven complexity measures for Spanish detailed in detail in the section 4.3.6. See Table 4.48.

The application of the LC metric has been fundamental to evaluate the quality of the texts by analyzing their level of lexical complexity. We have obtained an average value of 28.87, which indicates considerable complexity in the evaluated texts, which corresponds to various aspects,

Agreement between labelers						
	Academic Level					
	Low		Middle		University	
PI	Taggers					
	women	men	women	men	women	men
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 4.47: Analysis of the degree of agreement between the annotators in relation to the academic level.

Lexical Complexity Metrics for Spanish in GovAIEc									
	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
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 4.48: Results of the application of lexical complexity metrics for Spanish in corpus GovAIEc.

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.

Evaluation Applying LLMs

In this study, to meet our ultimate goal of evaluating the GovAIEc 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. These models have been widely used to create state-of-the-art solutions for numerous tasks (Paetzold, 2021). Executions were carried out with each of the models, applying 30, 50, 70 and 100 epochs. Our strategy focuses on integrating the 23 linguistic features of the corpus together the encodings generated by previously trained models; the objective is to evaluate whether this

combination provides satisfactory answers to our research hypothesis. 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 4.27.

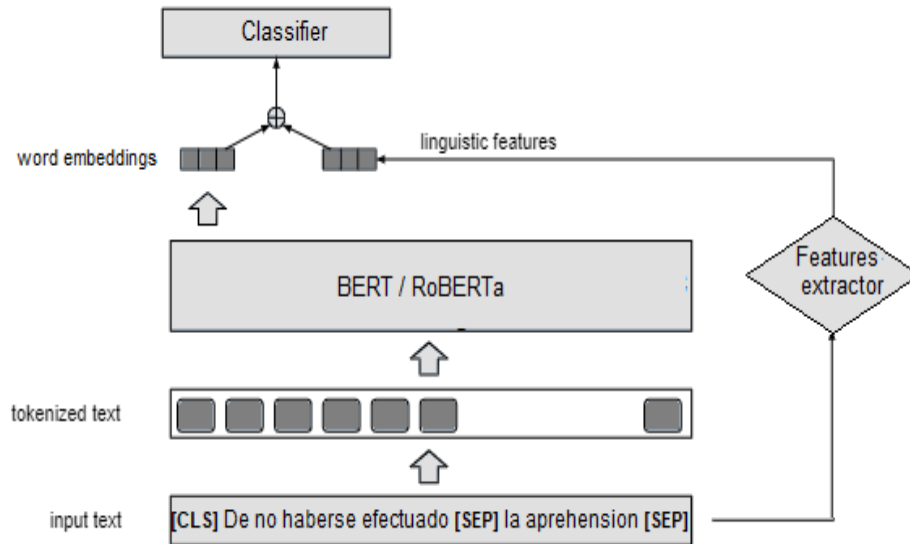


FIGURE 4.27: Process flow methodology integrating linguistic features.

<i>Spanish Language Model pre-trained with GovAIEc</i>				
Model	MAE	MSE	RMSE	R2
XLM-RoBERTa-Large	0.20618	0.05533	0.23521	-0.00692
XML-RoBERTa-Large $\oplus$ LF	0.19824	0.06221	0.24943	-0.11003
RoBERTa-Large-BNE	0.21077	0.05623	0.23712	-0.00321
RoBERTa-Large-BNE $\oplus$ LF	0.20399	0.05112	0.22610	0.02651
XML-RoBERTa-BASE	0.16807	0.06074	0.24645	-0.09903
XML-RoBERTa-BASE $\oplus$ LF	0.19492	0.04825	0.21966	0.09482
BERT	0.14641	0.05374	0.23181	0.01266
BERT $\oplus$ LF	0.14777	0.05151	0.22697	0.01875

TABLE 4.49: Results of the pre-trained models applying 50 epochs.

Description Of The Models

Below is a brief detail of the models used:

- **The pre-trained BERT** model that we chose was the one that the Spanish community uses mostly in research work to date, which is bert-base-uncased (BETO). **BERT-base** model has the number of layers $L=12$, the hidden size $H=768$, the number of self-attention heads $A=12$, and Total Parameters=110M.

- **BERT-large** model has the number of layers=24, the hidden size=1024, the number of self-attention heads=16, and Total Parameters=335M.
- **The XLM-RoBERTa-large** model has the number of layers L=24, the hidden size H=1024, the number of self-attention heads A=16, and Total Parameters=550M.
- **RoBERTa-large-BNE** model used was PlanTL-GOB-ES/roberta-large-bne being the largest Spanish-specific model to date.
- **XLM-RoBERTa-base** model has the number of layers L=12, the hidden size H=768, the number of self-attention heads A=12, and Total Parameters=270M.

The data was split into two sets: 80% was used for model training and the remaining 20% was used for evaluation. This split is indicated by the variable `train_portion = 0.8`. The batch size during model training was 32 samples per device, `per_device_train_batch_size=32`. It should be noted that during the evaluation, the model does not update its parameters, but the batch size is still considered important for the efficiency of the process, `per_device_eval_batch_size=32`. The epochs ensured that after the completion of each epoch, an evaluation of the model's performance is performed on the validation or test set, `evaluation_strategy='epoch'`. As for the training process logs, such as performance metrics and loss, which will be generated at the end of each epoch, we consider it very important and useful to monitor the training progress and track how the model improves over time, `logging_strategy='epoch'`.

We evaluated the results using the common error metrics MAE, MSE, RMSE and R2. The MAE (Mean Absolute Error) measures the average magnitude of the errors in a set of predictions without considering their direction; thus calculating the average of the absolute values of the differences between the predictions and the real values. The analysis of the MAE (Mean Absolute Error) for the different Spanish language models pre-trained with GovAIEc and trained with 50 epochs told us, on average, how far the model's predictions are from the real values.

The BERT model according to the results presents the lowest MAE with a value of 0.14641, which indicates that, on average, its predictions are closer to the real values compared to the rest of the models. In the case of the BERT model in combination with Linguistic Features, it also presents a low MAE = 0.14777, very close to the pure BERT, which shows that the addition of Linguistic Features does not significantly improve the MAE in this case. In the case of the RoBERTa-Large-BNE model, it presents a higher MAE = 0.21077, indicating that, on average, its

predictions are further away from the real values compared to the other models. Overall, BERT stands out as the best performing model for both English and Spanish because it is highly effective at understanding and generating text, which explains its high performance in many NLP tasks compared to traditional models. According to the results the addition of LF (Linguistic Features) improves the MAE for some models, but not for all. In particular, it does not improve for XML-RoBERTa-BASE.

BERT and RoBERTa have been widely adopted by the NLP development and scientific community. Both BERT and RoBERTa models have proven to be suitable for complex word identification and prediction tasks, each with their strengths. BERT showed slightly better performance in the presented results, being more efficient in terms of MAE. Its manageable size and wide adoption in the scientific community make it a solid choice, because of this, the BERT model is known for its ability to capture the bi-directional context of words, which is very important for complex word identification and prediction tasks, as it allows for a better understanding of the relationships between words and the context in which they are used.

The experiments revealed a significant improvement in the performance of the models when integrating the linguistic features obtained from the texts. The results of the evaluation indicated that the combination of these features contributes to improving the prediction of lexical complexity, which confirms our hypothesis. The lexical complexity metrics for Spanish showed that the terminology used in texts addressed to users becomes a difficulty in understanding their content.

4.3.7 Applying Prompting (unsupervised): Zero-Shot, Few-Shot, and Chain-Of-Thought

Identifying complex words in Spanish academic texts represents a significant challenge, as this type of language often incorporates technical and specialized terminology that can make them difficult to understand, even for advanced readers (Degraeuwe and Goethals, 2024). In the educational field, this problem directly impacts access to knowledge and the effectiveness of learning (Lipková, n.d.). Traditional techniques for simplifying and detecting complex words still face limitations in terms of accuracy and adaptability to specific academic contexts. To address this challenge, the application of Prompting in an unsupervised approach is proposed, using Zero-Shot, Few-Shot, and Chain-of-Thought techniques. This approach allows for improved identification of complex terms without the need for large amounts of labeled data,

leveraging the power of advanced language models to offer more adaptive and accurate solutions.

This section presents an analysis of the implementation and evaluation of the Prompting-based experimentation strategy, which has been applied in different unsupervised learning scenarios. In particular, the Zero-Shot, Few-Shot, and Chain-of-Thought approaches have been tested, allowing their effectiveness in specific tasks to be observed. The contexts in which these techniques have been tested are detailed below, as well as the results obtained in each of them.

The unsupervised learning approach we have explored was based on the use of large language models and their “prodding”. This is illustrated by our latest experiments with GPT-3 and its successors presented at the CLEF 2023 and CLEF 2024 international workshop in task 2 of SimpleText@CLEF-2023³², SimpleText@CLEF-2024³³ and LREC-COLLING 2024 - Determit!³⁴ focused on identifying and explaining difficult concepts.

The results obtained in our research, titled “SINAI Participation in SimpleText Task 2 at CLEF 2023: GPT-3 in Lexical Complexity Prediction for General Audience” (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023b), highlight the relevance of low and zero probability learning strategies in the GPT-3 autoregressive model. Several tests were carried out to implement these strategies and our results stood out as some of the best in our participation in Tasks 2.1 and 2.2 of the SimpleText track of CLEF 2023. These results demonstrated strong performance on the task of predicting complexity lexical.

The SimpleText track at CLEF faces the challenges of text simplification methods with the purpose of promoting access to scientific information. It provided a corpus of scientific literature abstracts and popular science requests. We have only participated in task 2, complexity spotting. In the Second task proposed, the goal of complexity detection (what is unclear?) is to identify what is unclear in a passage and a query. It seeks to classify the terms or concepts that require additional explanation to understand the passage, including definitions, context, and applications (Ermakova et al., 2022b).

³²SimpleText@CLEF-2023. Available at <http://simpletext-project.com/2023/clef/>

³³SimpleText@CLEF-2024. Available at <https://simpletext-project.com/2024/en/>

³⁴LREC-COLLING 2024 - Determit!. Available at <https://determit2024.dei.unipd.it/>

Description of Task 2: What Is Unclear?

Complexity Spotting: Identifying and explaining difficult concepts for a general audience. The goal of this task is:

1. To decide which terms (up to 5) require explanation and contextualization to help a reader understand a complex scientific text – for example, with regard to a query, terms that need to be contextualized (with a definition, example, and/or use-case)
2. To provide short (one/two sentence) explanations/definitions for the detected difficult terms. For the abbreviations, the definition would be the extended abbreviation.

Datasets for Task 2.2

For Task 2.2, the training set includes the same set of 203 difficult terms. However, in Task 2.2, the training set is extended by the addition of corresponding definitions for each of these difficult terms. To evaluate the performance of the runs submitted for this task, a test set comprising approximately 800 terms with ground-truth definitions is used. This test set serves as a benchmark against which the performance of the participants' systems is measured. The runs were evaluated by comparing the systems' results to the ground-truth definitions. The utilization of a substantial number of terms in the test set ensures a comprehensive evaluation of the systems' performance in interpreting and providing accurate definitions for the given terms. From this set, approximately 300 terms are selected for annotation using a clustering mechanism, ensuring that the test set contains a sufficient number of annotated samples for the majority of the runs.

The test set comprises a total of 15,056 sentences containing at least one of these terms. These sentences are used to evaluate the performance of the submitted runs. For the evaluation of abbreviation expansion, a set of approximately 1,000 abbreviations is manually annotated. In addition, an additional 4,374 abbreviations are extracted using the Schwartz and Hearst algorithm [11] from the sentences in Task 1, resulting in a total of approximately 5,000 abbreviations. The final test set consists of 38,416 sentences containing at least one of these abbreviations, and this set of sentences is used for the final evaluation of this subtask (Ermakova et al., 2023).

Proposed System

As stated before, our approach consists of applying zero-shot or few-shot prompting to the GPT-3 model extracting the predictions from the generated sequence. The details of the model are given in Table 4.50. The model is queried through its Python API and OpenAI playground is a convenient web interface that can be used to generate the needed code in a fast and easy way.

Parameter	Value
Model	text-davinci-003
Temperature	0.7
Max_tokens	250
Top_p	1
Frecuency Penalty	0
Presence Penalty	0
Logprobs	5
Stop	“###”

TABLE 4.50: GPT-3 model details.

Experiments With Few-Shot Learning

We apply three executions with few-shot learning with different prompts. The values of the settings also varied, it was observed that in the determination of the level of complexity (difficult, very difficult, or neutral), the model tended to identify terminology beyond its verbatim expression in the text. After the experiments, we show that this approach is important because the model can be adapted to a specific task. Next is the detail of the different prompts tested:

Prompt PRM_FS_TASK2_1_V1:

```
To retrieve up to 5 difficult terms in a given passage from a
scientific abstract. Rank the list of difficult terms with
corresponding scores on the scale 1-3 where 3 is the most
difficult term. You have a maximum of 250 tokens you can t go over
that. (here is an example)
context: quantum computing
Text: This paper describes a tool that converts Unreal levels to web-ready
environments in VRML and X3D.
##
Term 1: X3D
Difficulty:2
Term 2: vrml
Difficulty:2
Now it is your turn
```

```
##
Context: drones
Text: In an attempt to achieve the above mentioned tasks, we
propose an imitation learning based, data-driven solution to UAV
autonomy for navigating through city streets by learning to fly
by imitating an expert pilot.
###
```

Prompt PRM_FS_TASK2_1_V2:

```
To decide which terms (up to 5) require explanation and
contextualization to help a reader to understand a complex
scientific text. Rank the list of difficult terms with
corresponding scores on the scale 1-3 where 3 is the most
difficult term. You have a maximum of 250 tokens you can t go
over that. (here is an example)
context: quantum computing
Text: This paper describes a tool that converts Unreal levels to web-ready
environments in VRML and X3D.
##

Term 1:X3D
Difficulty:2
Term 2: vrml
Difficulty:2
Now it is your turn
##
Context: misinformation
Text: We find that Italy is the most polarized country, followed
by France, Germany and lastly Spain.
###
```

Prompt PRM_FS_TASK2_2_V1:

```
Select a number from 1 to 5 terms within the sentence named
text, rank them from 1 to 3 is the most difficult and finally
give the meaning of each difficult term and add an example and
an use-case. Use next form (term, difficulty, 'definition,
example, use-case'). You have a maximum of 250 tokens you can t go
over that. Here is an example:
context: quantum computing
Text: This paper describes a tool that converts Unreal levels to
web-ready environments in VRML and X3D.
##

Term 1:3D
Difficulty:2
```

Definition: X3D is an ISO-ratified, file format and run-time architecture to represent and communicate 3D scenes and objects. X3D fully represents 3-dimensional data.

Term 2:vrml

Difficulty:2

Definition: Virtual Reality Modeling Language (VRML) is a standard file format for representing 3-dimensional interactive vector graphics, designed particularly with the World Wide Web in mind. It has been superseded by X3D.

Now it is your turn

##

Context: drones

Text: This paper discusses preliminary work regarding robot companionship applications by using a miniature humanoid capable of fetching different toys based on voice command.

###

Table 4.51 presents the results generated by applying few-shot learning, where the model has been able to learn from a very small number of training examples.

Experiments with Zero-Shot Learning

In the case of zero-shot learning, no examples are provided in the prompt. Below are the specifications of our prompt.

Prompt PRM_ZS_TASK2_1_V1:

To retrieve up to 5 difficult terms in a given passage from a scientific abstract. Rank the list of difficult terms with corresponding scores on the scale 1-3 where 3 is the most difficult term. you have a maximum of 250 tokens you can t go over that.

##

Context: drones

Text: With the ever increasing number of unmanned aerial vehicles getting involved in activities in the civilian and commercial domain, there is an increased need for autonomy in these systems too.

###

# Execution	Prompt	Term_rank_snt	Term	Difficulty
1 - Task 2.1	PRM_FS_TASK2_1_V1	1	UAV	2
		2	Autonomy	3
		3	Imitation Learning	3
		4	Data-Driven	2
		5	Expert Pilot	2
2 - Task 2.1	PRM_FS_TASK2_1_V2	1	Polarized	3
		2	Italy	1
		3	France	1
		4	Germany	1
3 - Task 2.2	PRM_FS_TASK2_2_V1	1	Voice command <i>Definition:</i> Voice command is a type of technology that enables a user to control devices or services using voice commands. <i>Example:</i> Voice command technology is used in Amazon's Alexa to control smart home devices. <i>Use Case:</i> Voice command technology can be used to control home appliances, such as lights, thermostats, and security systems.	2
		2	Robot companionship <i>Definition:</i> Robot companionship is a type of human-robot interaction in which robots are used to provide companionship to humans. <i>Example:</i> A robotic pet can be used to provide companionship to elderly people who live alone. <i>Use Case:</i> Robot companionship can be used to provide comfort and companionship to people who are isolated or have difficulty forming relationships with other people.	3

TABLE 4.51: Predictions generated by applying GPT-3 with *few-shot learning*.**Prompt PRM_ZS_TASK2_1_V2:**

To decide which difficult terms (up to 5) are in the complex scientific text. Rank the list of difficult terms with corresponding scores on the scale 1-3 where 3 is the most difficult term. you have a maximum of 250 tokens you can t go over that.

##

Context: drones

Text: In an attempt to achieve the above mentioned tasks, we propose an imitation learning based, data-driven solution to UAV autonomy for navigating through city streets by learning to fly by imitating an expert pilot.

###

Prompt PRM_ZS_TASK2_2_V1

```

Select a number from 1 to 5 terms within the sentence named
text, rank them from 1 to 3 is the most difficult, and finally
give the meaning of each difficult term and add an example and
a use-case. Use the next form (term, difficulty, 'definition,
example, use-case'). you have a maximum of 250 tokens you can t go over that.

##
Context: drones
Text: Derived from the classic image classification algorithms,
our classifier has been constructed in the form of a fast 39-
layered Inception model, which evaluates the presence of roads
using the tomographic reconstructions of the input frames.
###

```

Table 4.52 presents the results generated by applying zero-shot learning, where the model has been able to generate its own results without the need for it to learn from a small number of examples.

In the example of one of the texts presented by Table 4.52 for task 2.1 on the first run - PRM_ZS_TASK2_1_V1, the model has identified a total of 5 words (corresponds to the column *Term_rank_snt*) achieving to generate its response according to the request made. The model has also managed to identify the complex terms of the text (corresponds to the column *Term*), The model has also been able to provide a complexity value for each term (corresponds to the column *Difficulty*).

In the second execution for the generation of results of task 2.1 - PRM_ZS_TASK2_1_V2, the model has not only been able to identify a total of 5 difficult terms from the content of the text, but it has also been able to generate a brief description of the difficult terms; we must emphasize that it had not been specified in the request. Finally, taking as a sample one of the examples resulting from the third execution PRM_ZS_TASK2_2_V1, and as a requirement specified in task 2.2, the model has generated a definition, an example, and a use case to better illustrate the reader about the term identified as difficult. In the run example, the model has identified 4 complex words in the text. According to the examples of the analyzed sample, GPT-3 has shown an outstanding ability to recognize the definitions of acronyms, generating coherent definitions, examples, and use cases, the issue of identification and recognition of complex words has been

# Execution	Prompt	Term_rank_snt	Term	Difficulty
1 - Task 2.1	PRM_ZS_TASK2_1_V1	1	Autonomy	3
		2	Unmanned Aerial Vehicles (UAVs)	2
		3	Civilian	1
		4	Commercial	1
		5	Activities	1
2 - Task 2.1	PRM_ZS_TASK2_1_V2	1	UAV autonomy:	3
		2	Imitation learning: A type of machine learning algorithm that learns from observing and mimicking an expert.	2
		3	Data-driven Using data to inform a decision-making process.	2
		4	City streets Urban roadways.	1
		5	Expert pilot A highly skilled and experienced pilot.	1
3 - Task 2.2	PRM_ZS_TASK2_2_V1	1	Inception model <i>Definition:</i> A type of convolutional neural network, often used for image classification and object ecognition. <i>Example:</i> GoogleNet, a popular <i>Inception model</i> , was used to classify images in the ImageNet challenge <i>Use Case:</i> Drones can use <i>Inception models</i> to identify roads in tomographic reconstructions of input frames.	3
		2	Image classification <i>Definition:</i> The process of recognizing the content of an image. <i>Example:</i> A photo of a dog may be classified as an image of an animal. <i>Use Case:</i> Drones can use <i>image classification</i> algorithms to identify features in tomographic reconstructions of input frames.	2
		3	Tomographic reconstructions <i>Definition:</i> The process of reconstructing an image from a series of slices taken from different angles. <i>Example:</i> A computed tomography (CT) scan is a type of <i>tomographic reconstruction</i> . <i>Use Case:</i> Drones can use <i>tomographic reconstructions</i> to analyze the presence of roads in input frames.	3
		4	ImageNet challenge <i>Definition:</i> A competition to classify images into different categories. <i>Example:</i> The <i>ImageNet challenge</i> involves classifying images into categories like animals, plants, and vehicles. <i>Use Case:</i> Drones can use the results of the <i>ImageNet challenge</i> to enhance their capabilities in various ways.	1

TABLE 4.52: The Experiments applying GPT-3 with Zero-shot learning.

Run	total	BLEU BLEU	ROUGE precision	ROUGE recall	ROUGE fmeasure	semantic match
Abbreviation extension results:						
SINAI_task_2.1_PRM_FS_TASK2_2_V1	228	1.35	0.05	0.81	0.1	0.53
SINAI_task_2.1_PRM_ZS_TASK2_2_V1	228	1.61	0.07	0.76	0.13	0.55
Definition generation results:						
SINAI_task_2.1_PRM_FS_TASK2_2_V1	75	4.17	0.19	0.37	0.24	0.78
SINAI_task_2.1_PRM_ZS_TASK2_2_V1	89	3.2	0.17	0.29	0.21	0.68
Results on the extended test set:						
SINAI_task_2.1_PRM_FS_TASK2_2_V1	257	4.64	0.17	0.39	0.22	0.77
SINAI_task_2.1_PRM_ZS_TASK2_2_V1	649	03.08	0.16	0.34	0.19	0.69

TABLE 4.54: SINAI results for task 2.2.

addressed with success.

In Table 4.53 and Table 4.54 we show the results obtained in our participation in tasks 2.1 and 2.2 respectively, according to the scores officially released by the organizers. The runs were named against our team ID (SINAI) together with the prompt strategy used from the ones described in previous sections. We have omitted our team ID for simplification. Our results in task 2.1 show a significant performance of the proposed approach close to top performing systems. It is interesting to note that, in general, version 1 for the prompt is better and that zero-shot learning prompts are better at detecting a difficult word, while few-shot learning prompts are better at determining a difficulty score. A mixture of both approaches could be worth exploring. In task 2.2 we obtain high values at semantic match. For providing definitions or explanations over difficult terms, few-shot learning prompting is consistently better than zero-shot learning prompts.

Run	total	evaluated	TERM LIMITS OK	Correct difficulty scores	Correct LIMITS & Correct difficulty scores
PRM_ZS_TASK2_1_V1	11081	1044	970 (5 th)	450 (3 rd)	428 (3 rd)
PRM_FS_TASK2_1_V1	10768	1025	942 (6 th)	498 (2 ^d)	471 (2 ^d)
PRM_ZS_TASK2_1_V2	10952	861	794 (7 th)	426 (4 th)	402 (4 th)
PRM_FS_TASK2_1_V2	8836	800	739 (8 th)	397 (5 th)	378 (5 th)
UAms_Task_2_RareIDF	675090	1200	1068 (1 st)	503(1 st)	476 (1 st)

TABLE 4.53: SINAI results for task 2.1. The ranking reached over the 34 runs submitted by different teams is shown. The results obtained by the best submitted run are in the last row.

In the submissions for Task 2.1 were a total of 12 teams participated and submitted a combined total of 39 runs for this task. The outcomes of these runs are presented in Table 4.55, which contains the following metrics: the total number of evaluated terms, the number of terms with

correct term limits, the number of correctly attributed scores (regardless of term limits), and the number of correctly limited terms with correctly attributed scores (+Limits). Among all the runs for Task 2.1, our run called SINAI_task_2.1_PRm_ZS_TASK2_1_V1 (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023b) achieved the highest number of correctly detected terms and scores. The performance of different runs on both train and test sets (only on correctly limited terms) is presented in Table 4.56. Generative models, like auto-regressive language models, can accomplish the task of CWI. Most approaches achieve comparable performance on both train and test sets.

The estimation of difficulty scores for terms proved to be a challenging task, as the majority of the submitted runs struggled to provide accurate scores for more than half of the detected difficult terms. This indicates the difficulty and subjectivity inherent in determining the level of complexity or difficulty associated with specific terms in scientific texts. The variability in assessing the difficulty of terms highlights the need for further research and improvement in developing robust methods for accurately assigning difficulty scores to terms.

Participants employed a range of approaches, including Large Language Models (LLMs) and unsupervised methods, to tackle the task. However, several runs were limited or incomplete due to token constraints imposed by LLMs or the time required for their execution. It was observed that the results of the same methods varied significantly depending on the specific implementation, fine-tuning techniques, and prompts utilized during the process. In terms of difficult term detection, LLMs demonstrated comparable performance to other methods as RareIDF, TextRank, and YAKE!. However, it is worth noting that the term difficulty scores assigned by the models differed considerably from the lay annotations such (Ermakova et al., 2023).

In the results obtained in our research, titled “Enhancing Lexical Complexity Prediction through Few-shot Learning with Gpt-3” in LREC 2024 - DETERMIT! Conference, we evaluated the ability of the GPT-3 language model to classify terms regarding their lexical complexity. This was achieved through the creation and evaluation of different versions of the model: text-Davinci-002 y text-Davinci-003 and prompts for few-shot learning to determine the complexity of the words. The results obtained on the CompLex dataset achieve a minimum average error of 0.0856. Although this is not better than the state of the art (which is 0.0609), it is a performing and promising approach to lexical complexity prediction without the need for model fine-tuning (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024b).

	Total	Evaluated		Score	
		+ Limits		+ Limits	
SINAI_task_2.1_PRM_ZS_TASK2_1_V1	11081	1322	1185	556	507
UAms_Task_2_RareIDF	675090	1293	1145	309	241
SINAI_task_2.1_PRM_FS_TASK2_1_V1	10768	1235	1122	440	405
Smroltra_task_2.1_keyBERT_FKgrade	11099	1215	1061	379	341
Smroltra_task_2.1_keyBERT_F	11099	1215	1061	223	171
UOL-SRIS_2.1_KeyBERT	23757	1215	1061	0	0
MiCroGerk_task_2.1_TextRank	21516	1275	1002	482	391
Smroltra_task_2.1_TextRank_FKgrade	10056	1275	1002	456	363
SINAI_task_2.1_PRM_ZS_TASK2_1_V2	10952	1075	965	366	330
SINAI_task_2.1_PRM_FS_TASK2_1_V2	8836	1004	915	346	316
Smroltra_task_2.1_YAKE_D	11112	1576	905	627	422
MiCroGerk_task_2.1_YAKE	23790	1576	905	582	362
Smroltra_task_2.1_YAKE_Fscore	11112	1576	905	409	209
MiCroGerk_task_2.1_GPT-3	15892	968	889	487	459
UBO_task_2.1_FirstPhrases	14088	1032	831	210	161
UBO_task_2.1_PositionRank	13881	1071	825	237	181
UBO_task_2.1_SingleRank	14088	981	748	200	151
UBO_task_2.1_TfIdf	14340	1206	740	263	187
UBO_task_2.1_TextRank	14088	960	722	189	139
Smroltra_task_2.1_RAKE_AUI	10660	1016	713	378	288
Smroltra_task_2.1_RAKE_F	10660	1016	713	255	170
UBO_task_2.1_TopicRank	13912	824	663	174	144
UBO_task_2.1_YAKE	14337	1118	576	265	116
MiCroGerk_task_2.1_BLOOM	9600	608	535	235	218
Aiirlab_task_2.2_KBIR	4797	498	429	158	135
TeamCAU_task_2.1_ST5	2234	484	418	222	201
Smroltra_task_2.1_SimpleT5	2234	460	406	259	239
Smroltra_task_2.1_SimpleT5_COLEMAN_LIEAU	2234	460	406	168	152
TheLangVerse_task_2.2_openai-curie-finetuned	2234	445	391	0	0
ThePunDetectives_task_2.1_SimpleT5	152072	428	371	110	91
Aiirlab_task_2.2_YAKEIDF	4790	465	241	154	75
Aiirlab_task_2.2_YAKE	4790	486	234	169	78
TeamCAU_task_2.1_AI21	100	10	6	3	2
Smroltra_task_2.1_Bloom	100	4	2	1	1
TeamCAU_task_2.1_BLOOM	100	1	1	0	0

TABLE 4.55: SimpleText Task 2.1: Results for the official runs Ermakova et al., 2023.

Dataset

We used the *CompLex* corpus proposed by Shardlow, Cooper, and Zampieri, 2020b. *CompLex*³⁵ is the first English multiple domain dataset, where words are scored in a context concerning their complexity using a five-point Likert scale to label complex words in texts of three sources/domains: the Bible, Europarl and Biomedical texts. The corpus is split into single-word and multiple-word annotations. The corpus contains a total of 9,476 sentences, each annotated by approximately 7 annotators, see Table 4.57.

Each entry contains the name of the source corpus, the sentence representing the context of the word, the targeted word to classify, and a score (in training data) that is an average of the scores given by different annotators, taking into account that each class has a predefined

³⁵CompLex: Is available at <https://github.com/MMU-TDMLab/CompLex>

	Total	Test		Train	
		Evaluated	Score	Evaluated	Score
SINAI_task_2.1_PRM_ZS_TASK2_1_V1	11081	1185	507	94	56
UAms_Task_2_RareIDF	675090	1145	241	40	21
SINAI_task_2.1_PRM_FS_TASK2_1_V1	10768	1122	405	81	40
Smroltra_task_2.1_keyBERT_FKgrade	11099	1061	341	41	4
Smroltra_task_2.1_keyBERT_F	11099	1061	171	41	7
UOL-SRIS_2.1_KeyBERT	23757	1061	0	42	0
MiCroGerk_task_2.1_TextRank	21516	1002	391	87	61
Smroltra_task_2.1_TextRank_FKgrade	10056	1002	363	87	29
SINAI_task_2.1_PRM_ZS_TASK2_1_V2	10952	965	330	94	53
SINAI_task_2.1_PRM_FS_TASK2_1_V2	8836	915	316	76	41
Smroltra_task_2.1_YAKE_D	11112	905	422	71	21
MiCroGerk_task_2.1_YAKE	23790	905	362	71	51
Smroltra_task_2.1_YAKE_Fscore	11112	905	209	71	32
MiCroGerk_task_2.1_GPT-3	15892	889	459	79	43
UBO_task_2.1_FirstPhrases	14088	831	161	49	19
UBO_task_2.1_PositionRank	13881	825	181	71	29
UBO_task_2.1_SingleRank	14088	748	151	67	19
UBO_task_2.1_Tfidf	14340	740	187	50	13
UBO_task_2.1_TextRank	14088	722	139	67	16
Smroltra_task_2.1_RAKE_AUI	10660	713	288	48	25
Smroltra_task_2.1_RAKE_F	10660	713	170	48	21
UBO_task_2.1_TopicRank	13912	663	144	61	21
UBO_task_2.1_YAKE	14337	576	116	44	11
MiCroGerk_task_2.1_BLOOM	9600	535	218	64	34
Aiirlab_task_2.2_KBIR	797	429	135	38	11
TeamCAU_task_2.1_ST5	2234	418	201	90	79
Smroltra_task_2.1_SimpleT5	2234	406	239	82	74
Smroltra_task_2.1_SimpleT5_COLEMAN_LIEAU	2234	406	152	82	21
TheLangVerse_task_2.2_openai-curie-finetuned	2234	391	0	165	0
ThePunDetectives_task_2.1_SimpleT5	152072	371	91	118	49
Aiirlab_task_2.2_YAKEIDF	4790	241	75	17	4
Aiirlab_task_2.2_YAKE	4790	234	78	19	3
TeamCAU_task_2.1_AI21	100	6	2	0	0
Smroltra_task_2.1_Bloom	100	2	1	0	0
TeamCAU_task_2.1_BLOOM	100	1	0	0	0

TABLE 4.56: SimpleText Task 2.1: Results for the official runs on train and test sets. The evaluation is done based on the terms with thier limits correctly detected Ermakova et al., 2023.

	All	Single words	Multiple words
Europarl	3,496	2,896	600
Biomed	2,960	2,480	480
Bible	3,020	2,600	420
All	9,476	7,974	1,500

TABLE 4.57: Volumetric information for the single and multiple words in each subcollection of the CompLex dataset.

weight between 0 (very easy) and 1 (very difficult) in a linear distribution over possible classes, see Table 4.58. Sample entries are shown in Table 2.12. This resource was used as a benchmark collection for SemEval 2021 Task 1: Prediction of Lexical Complexity of the 15th International Workshop on Semantic Evaluation³⁶ (Shardlow et al., 2021a).

³⁶Semeval 2021 - LCP SHARED TASK 2021 - <https://sites.google.com/view/lcpsharedtask2021>

Scale	Description	Complexity
<i>Very Easy</i>	Words which were very familiar to an annotator.	0
<i>Easy</i>	Words with which an annotator was aware of the meaning.	0.01 - 0.25
<i>Neutral</i>	A word which was neither difficult nor easy.	0.26 - 0.50
<i>Difficult</i>	Words in which an annotator was unclear of the meaning, but may have been able to infer the meaning from the sentence.	0.51 - 0.75
<i>Very Difficult</i>	Words that an annotator had never seen before, or were very unclear.	0.76 - 1.00

TABLE 4.58: Categories on the Likert scale proposed by (Shardlow, Cooper, and Zampieri, 2020b).

Methodology

We based our work on the one presented by Mostert et al. (2022), applying GPT-3 in this text classification task, specifically, a task of identification of complex words in texts and the respective categorization of them. The experiments were carried out applying a *few-shot classification* where the purpose was to carry out an analysis of the content of the texts coming from diverse sources such as the Bible, Biomedical, and Europarl to determine that GPT-3 was able to predict the complexity degree of the word. Various experiments were carried out applying different contents in the prompt, this is because GPT-3 has no concrete way of verifying the truth, logic, or meaning of any of the millions of lines of text it generates daily. The setting for the model parameters for GPT-3 is given in Table 4.60.

The steps that were carried out during the development of the experimentation are the following:

- The respective model configuration was specified.
- Different prompts were built and executed with the data set in its training and evaluation phases, to ask the model to return a “soft” response in the Likert scale.
- The probabilities were obtained to know what is the priority with which the model determines its result.
- The respective evaluation metrics were calculated: MAE, MSE, and RMSE, to determine the accuracy of the results.
- The respective comparison of the results generated by the model versus the data set was performed.

Parameter	Values
model	text-davinci-003
prompt	orden
temperature	0
maximum tokens	5
top_p	1
presence_penalty	0
logprobs	5

TABLE 4.60: GPT-3 Model Configuration.

Few Shot Prompting For Lexical Simplification

We applied the *few shot prompting* strategy by providing the model with a few samples of what we wanted it to do. In Table 4.61 the prompt provided to the model is shown. The prompt contains several examples that are complemented with the word to be evaluated from the text, and it also indicates which resource corresponds the text. After prompt specification, the values are replaced by the CompLex corpus dataset with the sentence and word to be evaluated.

Construction Of The Different Prompt

To facilitate comparison, analysis, and interpretation of the results of the model runs against the test data set assign a name for each execution as is in the Prompt Variants column, as can be seen in See Table 4.64. The differentiation in the construction of the various prompts intended for the GPT-3 model was based on their size, determined by the number of examples integrated for construction and training during their learning phase. This distinction was indicated by the first letter of the respective name: S to indicate small applications (Small) with an average of 2 to 4 examples, M for medium-sized ones (Medium) with an average of 5 to 6 examples, and L for those of large magnitude (Large) with an average of examples between 9 and 12 examples included. This process aimed to generate multiple prompts that enable the model to offer more precise results during its evaluation. This process aimed to generate multiple prompts that enable the model to offer more precise results during its evaluation.

Next, SO comes from Source, that is, whether or not the source from which the text to be evaluated came was included in the prompt specification. We also include the *nor* operator (the

I'm reading fragments from some sources such as the Bible, Biomed, and Europarl, and some words are not easy to understand. I'm classifying these words into "very easy", "easy", "neutral", "difficult" and "very difficult". The sentence is "neutral" when it is neither "very easy", nor "easy", nor "difficult", nor "very difficult". Several examples are: "However, no defects in axon pathfinding along the monosynaptic reflex arc or in muscle spindle differentiation have been noted in PV KO mice, which develop normally and show no apparent changes in their behavior or physical activity (Schwaller et al. 1999).". I find that word "spindle" is neutral.

###

The following fragment comes from the "bible" and after reading the fragment "I will sprinkle clean water on you, and you shall be clean: from all your filthiness, and from all your idols, will I cleanse you. ". I find that the word "filthiness" is easy.

###

The following fragment comes from the "biomed" and after reading the fragment "Moreover, acute dosing does not recapitulate the marked learning deficits produced in rodents [15,16] by chronic exposure to dopamine D2R antagonists [6,7].". I find that the word "antagonists" is difficult.

###

The following fragment comes from the "biomed" and after reading the fragment "Thrombus formation on fissured atherosclerotic plaques is the precipitating event in the transition from a stable or subclinical atherosclerotic disease and leads to acute myocardial infarction, ischemic stroke or peripheral arterial occlusion.". I find that word "Thrombus" is very difficult.

###

The following fragment comes from the "bible" and after reading the fragment "Mount Sinai, all it, smoked, because Yahweh descended on it in fire; and its smoke ascended like the smoke of a furnace, and the whole mountain quaked greatly.". I find that the word "fire" is very easy.

###

The following fragment comes from the @recurso and after reading the fragment @oracion I find that word @aEvaluar is

TABLE 4.61: Prompt example

result of the negation of the OR operator) and *neither* to indicate the denial of the alternatives presented, translating to "NOR" and "neither" which would mean "none" or "none", and we have used them to express that the application does not consider any of the two previous options mentioned. See Table 4.65 in the *Prompt Variants* column.

Methods Used To Calculate The Complexity Level Of Words

To calculate the level of complexity of complex words generated by GPT-3 as a value within the range [0, 1], we explored three ways based on the categories of the complex words, as detailed below. In this way, the linguistic responses of the model are transformed into numerical

values. In Table 4.58 the range of values that correspond to the complexity of each category was presented.

1. Method #1 - Middle of the Range

Half between the lower limit and the upper limit of each range of complexity values. Scores are fixed on a per category basis. For example:

$$\text{Neutral} = (0.26 + 0.50) / 2$$

$$\text{Neutral} = 0.375$$

The calculated values for each category are:

$$\text{Very Easy} = 0$$

$$\text{Easy} = 0.125$$

$$\text{Neutral} = 0.375$$

$$\text{Difficult} = 0.625$$

$$\text{Very Difficult} = 0.875$$

Table 4.62 presents in the column *Method #1* the results of an execution carried out with a total of 30 records where it can be seen that there is a large number of coincidences with the categories that correspond to the complex word of the CompLex corpus. The highlighted values correspond to matches. After the execution with the test data, it was obtained a MAE=0.1293, MSE=0.0258, RMSE=0.1608.

2. Method #2 - Average by Category

The average of complexity values of a category is used as the complexity degree for that category. Therefore, the scores are fixed on a per category basis. Again, scores are fixed on a per category basis. The table 4.62 presents in the *Method #2* column the results of the execution carried out with the test records, a total of 30, where you can see the value calculated for the level of complexity of the categories generated by the model for the complex words of the texts. After the execution with the test data, it was obtained a MAE=0.086, MSE=0.016, RMSE=0.125.

The corresponding calculated values for each category would be the following:

$$\text{Very Easy} = 0$$

$$\text{Easy} = 0.189$$

$$\text{Neutral} = 0.351$$

$$\text{Difficult} = 0.588$$

$$\text{Very Difficult} = 0.811$$

3. Method #3 - The Confidence of GPT-3

The Confidence Level of the model corresponds to the high percentage of precision and coherence with which the model has made use of its attention mechanism and the context to select the category to which the complex word in the text corresponds.

We consider for the assignment of the category generated by the model the one whose confidence level is the highest. For example: If the confidence level is 90% for the Easy category, the complexity closest to the left limit of the category range is taken. If the confidence level is 80% for the Difficult category, the complexity level that is furthest to the right of the category range is assigned. In the case of Very Difficult, the same procedure as the previous ones is considered. The table 4.62 presents in the Method #3 column the results of the execution carried out with the test records, a total of 30, where it can be seen, the value calculated for the level of complexity of the categories generated by the model for the words complexities of the texts according to the level of confidence of the model. After the execution with the test data, a MAE=0.191, MSE=0.047, RMSE=0.216 was obtained.

It is important to note that in the table 4.62, in the “Category” column of the GPT-3 section, the complexity category generated by the GPT-3 model was selected based on the highest level. high confidence in percentage terms of the probabilities associated with the predictions generated by the model that corresponds to the category of the complex word. This is observed in the execution of the model on the corpus, using a test data set composed of 30 records, as detailed in the table 4.63.

Comparing the results applying GPT-3 with Few-Shot learning in corpus CompLex							
#	Corpus CompLex			2*Category	GPT-3		
	2*Complex word	2*Category	2*Range of Values		Complexity Level		
					Method #1	Method #2	Method #3
10	voice	neutral	0.01 - 0.25	easy	0.125	0.189	87.07% - 0.032
11	darkness	easy	0.01 - 0.25	easy	0.125	0.189	76.91% - 0.058
12	behold	easy	0.26 - 0.50	neutral	0.375	0.351	29.40% - 0.381
13	camp	easy	0.01 - 0.25	easy	0.125	0.189	81.11% - 0.045
14	bonds	easy	0.01 - 0.25	easy	0.125	0.189	54.29% - 0.115
15	statutes	neutral	0.01 - 0.25	easy	0.125	0.189	51.43% - 0.127
16	snares	easy	0.01 - 0.25	easy	0.125	0.189	54.95% - 0.112
17	exhortation	difficult	0.51 - 0.75	difficult	0.189	0.588	61.30% - 0.665
18	River	easy	0.01 - 0.25	easy	0.125	0.189	86.88% - 0.033
19	generation	easy	0.01 - 0.25	easy	0.125	0.189	85.27% - 0.037
20	dainties	difficult	0.51 - 0.75	difficult	0.189	0.588	58.36% - 0.657

TABLE 4.62: Methods applied to calculate the level of lexical complexity.

Probabilities associated with the complexity category predictions generated by the GPT-3 model The results applying the <i>davinci-002</i> model and <i>few shots learning</i> approach						
#	Token	GPT-3 complexity	CompLex complexity	GPT-3 confidence level		
				Option #1	Option #2	Option #3
20	dainties	difficult	difficult	difficult 59.39%	neutral: 25.25%	easy: 10.23%
21	subjection	difficult	neutral	difficult: 92.44%	neutral: 5.0%	easy: 1.16%
22	perverseness	difficult	neutral	difficult: 92.35%	neutral: 5.26%	very: 1.3%
23	grasshoppers	easy	easy	easy: 72.01%	neutral: 13.88%	very: 9.34%
24	signet	difficult	neutral	difficult: 74.86%	neutral: 20.39%	very: 2.74%
25	snare	easy	neutral	easy: 65.44%	neutral: 18.25%	difficult: 13.4%
26	Asher	easy	neutral	easy: 76.68%	very: 9.64%	difficult: 7.44%
27	demons	difficult	easy	difficult: 59.93%	easy: 30.62%	neutral: 7.4%
28	prophet	easy	easy	easy: 88.42%	neutral: 5.55%	difficult: 3.21%
29	lion	easy	neutral	easy: 90.23%	very: 4.46%	neutral: 4.32%
30	Lion	easy	easy	easy: 84.1%	very: 12.29%	difficult: 2.07%

TABLE 4.63: Probabilities associated with the predictions generated by the GPT-3 model that correspond to the category of the complex word.

Results

Our goal is to advance research on the use of the GPT-3 model to predict word complexity in the English language by adopting a few-shot examples learning approach. We have carried out multiple iterations with the objective that GPT-3 generates more precise and coherent answers with quality and relevance, we have formulated 19 several prompts pretending to optimize the performance of the model. Through this approach, we aspire to achieve greater precision in our predictions, approaching the results obtained by the winners of the lexical complexity prediction task proposed in the framework of SemEval 2021³⁷.

³⁷<https://sites.google.com/view/lcpsharedtask2021>

We experimented with different prompts issued to OpenAI's largest available model: text-davinci-002 and davinci-003 as evidenced by Table 4.65. Our first approach uses a singular prompt template in a few-shot setting to obtain the category of word complexity: *easy - very easy - neutral - difficult - very difficult*; we further improve upon these results by combining predictions from different prompt templates as can be seen in the Table 4.64, the application of different runs performed with the evaluation data set. The results derived from our approach toward single word prediction yielded the following values: MAE = 0.0875, MSE = 0.0131, and R2 = 0.1930.

Standard applied for the construction of the Prompt Variants				
Size	Source	Connector Logical	# exp	Emphasis
L-M-S	SO	NOR	05	Em
M	SO		05	
M	SO	NOR	05	Em
M	SO		06	
S		NOR	04	
S		NOR	04	Em
S	SO	NOR	05	
L	SO	NOR	09	

TABLE 4.64: Standard applied for the construction of the prompt variants.

Final results generated with GPT-3							
The results applying the davinci-002 and davinci-003 models and few shots learning approach							
#	Prompt Variants	Score Type	Model Version	Metrics			
				MAE	MSE	RMSE	Pearson
1	M-SO-05	Means	davinci-003	0.0882	0.0136	0.1165	0.5776
2	M-SONOR-05-Em	Means	davinci-003	0.0956	0.0153	0.1238	0.5103
3	M-SONOR-05-Em	Means	davinci-002	0.1011	0.0170	0.1305	0.4661
4	S-SONOR-05	Means	davinci-003	0.1057	0.0190	0.1378	0.5016
5	M-SO-06	Means	davinci-003	0.1074	0.0199	0.1412	0.4924
6	S-SONOR-05	Means	davinci-002	0.1098	0.0208	0.1442	0.5086
7	S-NOR-04	Means	davinci-002	0.1143	0.0229	0.1512	0.4919
8	S-NOR-04	Means	davinci-003	0.1725	0.0440	0.2099	0.3826
9	S-NOR-04-Em	Means	davinci-002	0.1793	0.0512	0.2262	0.3524
10	S-NOR-04-Em	Means	davinci-003	0.1875	0.0503	0.2242	0.4477
11	M-SO-05	Half of the range	davinci-003	0.1212	0.0219	0.1480	0.5730
12	M-SONOR-05-Em	Half of the range	davinci-003	0.1292	0.0239	0.1546	0.5099
13	S-SONOR-05	Half of the range	davinci-003	0.1475	0.0310	0.1761	0.5136
14	M-SO-06	Half of the range	davinci-003	0.1555	0.0345	0.1859	0.4944
15	S-NOR-04-Em	Half of the range	davinci-003	0.2164	0.0603	0.2456	0.4650
16	S-NOR-04	Half of the range	davinci-003	0.2106	0.0580	0.2409	0.3806
17	S-SONOR-05	GPT-3 Confidence level	davinci-003	0.2333	0.0655	0.2559	0.5600
18	M-SO-06	GPT-3 Confidence level	davinci-003	0.2658	0.0816	0.2857	0.5247
19	L-SONOR-09	GPT-3 Confidence level	davinci-003	0.2431	0.0708	0.2662	0.5241

TABLE 4.65: The results applying the davinci-002 and davinci-003 models and few shots learning approach.

A test was carried out by taking a sample of 30 records to train the model applying the

few-shot learning technique. The data in the column *GPT-3 Confidence Level* represents the level of lexical complexity generated by the GPT-3 model for each token in the corpus. In the table 4.65, we can see that the “Score Type” column in the last three rows shows runs where this strategy was applied to assign complexity levels to complex words in the corpus. This is complemented by the results presented in the table 4.66 in the “GPT-3 Complexity” column, which refers to the complexity generated by the model. It is from these values that the strategy for calculating lexical complexity was derived, called *GPT-3 Confidence Level*. Additionally, the table 4.66 shows matches where the model’s complexity prediction for a token matches the complexity assigned to the token in the CompLex corpus.

Probabilities associated with the level of complexity predictions generated by the GPT-3 model The results applying the davinci-002 model and few shots learning approach							
#	Token	GPT3 category	GPT3 range	GPT3 complexity	CompLex complexity	CompLex range	Match
20	dainties	difficult	0.51 - 0.75	0.5880	0.5625	difficult	Yes
21	subjection	difficult	0.51 - 0.75	0.5880	0.4375	neutral	No
22	perverseness	difficult	0.51 - 0.75	0.5880	0.4166	neutral	No
23	grasshoppers	easy	0.01 - 0.25	0.1896	0.25	easy	Yes
24	signet	difficult	0.51 - 0.75	0.5880	0.4687	neutral	No
25	snare	easy	0.01 - 0.25	0.1896	0.3194	neutral	No
26	Asher	easy	0.01 - 0.25	0.1896	0.4285	neutral	No
27	demons	difficult	0.51 - 0.75	0.5880	0.125	easy	No
28	prophet	easy	0.01 - 0.25	0.1896	0.2222	easy	Yes
29	lion	easy	0.01 - 0.25	0.1896	0.2812	neutral	No
30	Lion	easy	0.01 - 0.25	0.1896	0.1710	easy	yes

TABLE 4.66: The results of the probabilities associated with the level of complexity predictions generated by the GPT-3 model applying the davinci-002 and few shots learning approach.

In our participation in Tasks 2.1 and 2.2 of the SimpleText track of CLEF 2024, our work focused on the implementation of zero-shot learning using the GPT-4-Turbo autoregressive model. To this end, we develop and evaluate various cues to optimize the model’s ability to predict lexical complexity. The results of our experiments indicated that GPT-4-Turbo can perform this task with remarkably robust performance, demonstrating its potential to assess language complexity effectively without the need for additional training.

The SimpleText laboratory (Ermakova et al., 2024) is part of the CLEF 2024 (Faggioli et al., 2024b) initiative, which promotes the systematic evaluation of information access systems through experimentation with shared tasks. SimpleText addresses the challenges of text simplification, especially in the context of making scientific information accessible. In this sense, it provides relevant data and reference points, as the general public tends to avoid reliable sources

such as scientific literature due to its complex language and lack of prior knowledge. Instead, they rely on superficial and derivative sources on the web and social media, which are often published for commercial or political rather than informational purposes.

Can simplifying the text help overcome these barriers to access? That is the central question that this task addresses. Its objective is to generate a simplified summary of several scientific documents, based on a popular science query. This summary will provide the user with an accessible overview of the specific topic, with the hope of making scientific information more accessible and understandable to a broader audience.

The main objective of this research is to demonstrate the capacity of the language model based on Transformers like GPT-4-Turbo for lexical simplification. To achieve this, several variants of prompts without samples have been created and evaluated. This approach allows us to determine which concepts in scientific summaries require additional explanation and contextualization to help the reader understand the scientific text more effectively.

GPT-4-Turbo For Solving NLP Tasks

State Of The Art deep learning models, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), GPT-3 (Brown et al., 2020), and GPT-4-Turbo (Shahandashti et al., 2024), significantly outperform existing traditional approaches. GPT-4-Turbo, the transformer-based large language model generation developed by OpenAI³⁸, is a prominent example of these advances. These models have demonstrated superior capabilities in various natural language processing tasks, raising the bar for performance and accuracy in the field.

GPT-3 Contains 175 billion parameters. Due to its extensive language knowledge, enormous processing power, and potential to learn from large amounts of online text data, GPT-3 can perform a wide variety of general natural language-based tasks with unprecedented ease, such as the generation and classification of text (Kublik and Saboo, 2022). The enormous scale of the model allows generating results with quality, precision and diversity of the generated content. This development has sparked considerable interest and concern in the field of NLP, the machine learning industry in general, the media, the AI ethics communities and civil society (Chan, 2023), The latest GPT-4-Turbo model of the OpenAI has been recognized for its ability to process text

³⁸<https://openai.com/>

and images (Hirano et al., 2024). Although GTP-3 is a generative model, several approaches may be necessary to classify text, mainly ranging from zero-shot classification to single-shot and few-shot classification (Kublik and Saboo, 2022). In zero-shot learning no samples of data are needed (Bhattacharjee et al., 2024).

SimpleText@CLEF-2024 Tasks

Task 2: “What is unclear?” Difficult concept identification and explanation

The goal of this task is to identify key concepts that need to be contextualized with a definition, example, and/or use-case and provide useful and understandable explanations for them. Thus, there are three subtasks:

1. To predict what are the terms in a passage of a document and their difficulty (easy/medium/difficult).
2. To generate a definition and an explanation only for the difficult terms.
3. To retrieve the provided definitions of the difficult terms in “correct” order: manual (2), generated positive 1 (1), generated positive 2 (1), generated negative 1 (0), generated negative 2 (0).

In Task 2.1, for each passage of a document, participants should provide a list of terms with corresponding difficulty scores (easy/medium/difficult). Passages (sentences) are considered to be independent, so term repetition is allowed (the same term can be detected in different sentences, even in the same document).

In Task 2.2, for each difficult term (terms that have been evaluated with the highest level of difficulty), participants should provide the definition and explanation which will be evaluated both from a qualitative point of view (manual review by terminologists) and from a quantitative point of view (applying overlapping text measures like BLUE, ROUGE, etc.).

In Task 2.3, participants should rank the set of definitions provided for the difficult terms in a way that the “best” definitions are ranked higher in the list of definitions. In particular, for each term there will be one manual definition (considered the best one) and two automatically generated good definitions that should be placed at the top of the list of retrieved definitions. Quantitative metrics (for example, P@1, P@3, rank correlation measures) will be used to evaluate

participants' results. Task 2.1 and Task 2.3 can be performed independently. Participants that want to run experiments on Task 2.2 need to accomplish Task 2.1 first.

Data

The corpus for Task 2 was based on sentences from the high-ranking summaries of Task 1 applications and collected in 2023. A total of 175 documents and 1077 sentences were used to generate the training and test data. In particular, 115 documents and 576 sentences were used to build the training set and 60 documents and 501 sentences were used to build the test set. A dataset was provided to train the systems before the evaluation phase and a test set for the evaluation phase. In particular, the dataset comprises the following files:

- The documents and their sentences.
- Terms manually extracted from each sentence and their relative difficulty.
- Definitions and explanations provided by experts for difficult terms.
- Definitions automatically generated by a large language model.

For the training set, 21 experts were hired to manually annotate each document, identifying the terms in each sentence, assessing their difficulty, and providing definitions and explanations for each difficult term. This effort resulted in the generation of 1,609 terms and 899 definitions and explanations. To further analyze consistency across experts, in some cases we deliberately assigned the same documents to multiple experts. Additionally, for each term accompanied by a definition, we created two "good" definitions and two "bad" definitions. This was done to develop a set of definitions for classification in Task 2.3, resulting in a total of 2,356 sentences, evenly split between good and bad definitions.

In addition to this initial training set, we introduced an additional set of files produced by an external expert who reviewed the annotations of all 21 experts. This secondary set, called the validation set, included the expert's additions of missing terms, definitions, or both. This revision added 677 terms, 960 definitions, and 3,732 generated definitions (evenly split between good and bad) to the training data. For the test set, we asked the external expert to annotate the remaining 60 documents. A total of 1,440 terms were extracted and 424 definitions were written from the 501 sentences in the test set. An additional 3,816 definitions were also added (evenly

split between good and bad) (Di Nunzio et al., 2024).

Proposed System

As mentioned above, our approach is to apply learning cues without prior examples to the GPT-4-Turbo model, extracting predictions from the generated sequence. Details of the model configuration are presented in Table 4.67. We use the OpenAI API in Python to interact with the model, allowing us to easily integrate GPT-4-Turbo into our workflow. In addition, the OpenAI Playground offers an intuitive web interface that facilitates experimentation and rapid generation of code necessary for our tests and developments.

Parameter	Value
Model	gpt-4-turbo
Temperature	0
Max_tokens	500
Top_p	1
Frecuency Penalty	0
Presence Penalty	0
Logprobs	5
Stop	###

TABLE 4.67: GPT-4-Turbo model details.

Experiments With Zero-Shot Learning

We execute the model applying zero-shot learning where no examples are specified for learning and a sample of a generated output.

Prompt PRM_ZS_TASK2_V1

```
Performing the following actions according to a given text:
Select a number from 1 to 5 terms within the sentence named text
Assign the difficulty of the words: easy, medium and difficult, giving
the initial letter of each difficulty as an answer ([e]fácil/[m]edium/[d]ifficul)
Give the definition of each difficult term, add an explanation and use
case according to context of the sentences.
Provide the response in following JSON format:
{
  "word": <word>,
  "difficulty": <word_difficult>,
```

```

    "definition": "<definition>",
    "explanation": <explanation>
}
Context: {{context}}
Text : {{text}}
Response:
##
Context: Digital assistant
Text : Abstract    Novel technological advances in mobile devices and
applications can be exploited in wildfire confrontation, enabling end-
users to easily conduct several everyday tasks, such as access to data
and information, sharing of intelligence and coordination of personnel
and vehicles.
Response:

json
[
  {
    "word": "exploited",
    "difficulty": "m",
    "definition": "make full use of and derive benefit from (a
resource)",
    "explanation": "In the context, 'exploited' refers to utilizing
the new technological capabilities of mobile devices and
applications to their full advantage, specifically in managing
and combating wildfires.",
    "use_case": "The novel technologies can be exploited to improve
the efficiency and effectiveness of wildfire management."
  },
  {
    "word": "confrontation",
    "difficulty": "m",
    "definition": "a hostile or argumentative meeting or situation
between opposing parties",
    "explanation": "Here, 'confrontation' is used metaphorically to
describe the act of dealing with wildfires, suggesting a direct
and active engagement with the fires.",
    "use_case": "Mobile technologies are being used in the
confrontation of wildfires, aiding in direct engagement and
management."
  },
  {
    "word": "coordination",
    "difficulty": "m",
    "definition": "the organization of the different elements of a
complex body or activity so as to enable them to work together
effectively",
    "explanation": "In this context, 'coordination' refers to the

```

```

        organized management of personnel and vehicles to tackle
        wildfires efficiently, facilitated by mobile technology.",
        "use_case": "Effective coordination of personnel and vehicles
        is crucial for the successful management of wildfire incidents."
    }
]

###

```

Prompt PRM_ZS_TASK2_V2

Depending on the text, select up to 5 terms considered complex within the sentence, then assign the complexity of the terms to easy, medium and difficult using the first letter of each word, for example if the complexity is easy add the letter "e", If the complexity of the term is medium, add the letter "m", and if the complexity of the term is difficult, you should add the letter "d". Then, from each complex term generate its definition, an explanation and a use case according to the context of the sentence. Finally, provide the response in the following JSON format:

```

{
  "word": <word>,
  "difficulty" <difficult_word>,
  "definition": "<definition>",
  "explanation" : <explanations>
}

```

##

Context: Drug discovery

Sentence: A pharmacophore analysis approach was used to investigate and compare different classes of compounds relevant to the drug discovery process (specifically, drug molecules, compounds in high throughput screening libraries, combinatorial chemistry building blocks and nondrug molecules).

Response:

```

json
[
  {
    "word": "pharmacophore",
    "difficulty": "d",
    "definition": "A set of structural features in a molecule that is recognized by a biological macromolecule and is responsible for its biological activity.",
    "explanation": "In drug discovery, a pharmacophore analysis is used to identify these structural features which can interact with a biological target, such as a protein, to induce a therapeutic effect. This helps in designing or identifying new

```

```
drugs that can mimic or block the biological activity.",
"use_case": "In the sentence, a pharmacophore analysis approach
helps in comparing different classes of compounds to find
potential drug candidates."
},
{
  "word": "combinatorial chemistry",
  "difficulty": "d",
  "definition": "A method in chemistry where different
combinations of building blocks are systematically mixed to
generate a large number of different compounds.",
  "explanation": "Combinatorial chemistry is used in drug
discovery to rapidly synthesize and screen large libraries of
compounds for potential drug activity. It allows researchers to
explore a wide variety of chemical structures.",
  "use_case": "In the sentence, combinatorial chemistry building
blocks are mentioned as part of the classes of compounds
relevant to drug discovery."
},
{
  "word": "high throughput screening",
  "difficulty": "d",
  "definition": "A method used in drug discovery to quickly
conduct millions of chemical, genetic, or pharmacological
tests.",
  "explanation": "High throughput screening (HTS) is crucial in
the early stages of drug discovery. It allows researchers to
quickly identify active compounds, antibodies, or genes that
modulate a particular biomolecular pathway.",
  "use_case": "The sentence refers to compounds in high
throughput screening libraries, indicating these are tested in
large-scale assays to find promising drug candidates."
},
{
  "word": "nondrug",
  "difficulty": "m",
  "definition": "Substances or compounds that are not considered
drugs and do not have therapeutic effects.",
  "explanation": "In the context of drug discovery, nondrug
molecules are those that are used as controls or are part of
the compound libraries but are not expected to lead to a
therapeutic drug.",
  "use_case": "The sentence compares nondrug molecules with drug
molecules and other compounds to understand their relevance or
irrelevance in the drug discovery process."
},
{
  "word": "compounds",
```

```

        "difficulty": "e",
        "definition": "Chemical substances composed of two or more
elements chemically bonded together.",
        "explanation": "In the context of the sentence, 'compounds'
refers to various chemical entities that are being investigated
in the drug discovery process to find those with potential as
drug molecules.",
        "use_case": "The sentence discusses investigating different
classes of compounds in relation to their relevance to drug
discovery."
    }
]

###

```

Prompt PRM_ZS_TASK2_V3

Identify up to a total of 5 complex words in the sentence and determine the difficulty of each complex word by classifying it into the category of "easy", "medium" or "difficult". For the "easy" difficulty place the letter "e", for the "medium" difficulty place the letter "m", and for the "hard" difficulty place the letter "d". Furthermore, in the case of words identified as complex taking into account the context of the sentence, three things must be generated: a definition, an explanation, and a use case.

Finally, provide the response in the following JSON format:

```

{
  "word": <word>,
  "difficulty" :<hard_word>,
  "definition": "<definition>",
  "explanation" : <explanations>
}

```

##

Context: Humanoid robots

Sentence: In many situations, users walk on typical paths between specific destinations at which the service of a mobile robot is needed.

Response:

json

```

[
  {
    "word": "humanoid",
    "difficulty": "m",
    "definition": "having an appearance or character resembling
that of a human.",
    "explanation": "In the context of robots, 'humanoid' refers to
robots designed to look and sometimes behave like humans, often
used to interact with environments built for human use.",
  }
]

```

```

    "use_case": "The humanoid robot was programmed to assist
customers in the store."
  },
  {
    "word": "destinations",
    "difficulty": "m",
    "definition": "the place to which someone or something is going
or being sent.",
    "explanation": "In the given sentence, 'destinations' refers to
the specific places where the mobile robot is required to
perform services.",
    "use_case": "The tourists set off early in the morning to reach
their destinations before sunset."
  },
  {
    "word": "mobile",
    "difficulty": "e",
    "definition": "able to move or be moved freely or easily.",
    "explanation": "Here, 'mobile' describes a robot that is
capable of moving around rather than being fixed in one
location.",
    "use_case": "The company developed a mobile app to increase
accessibility for users on the go."
  }
]
###

```

Results

The Table 4.68 presents various results in response to the approach of sub-task 2.1. To automatically support this task, we have proposed an innovative approach that explores the potential of GPT-4-Turbo, an advanced large language model (LLM) developed by OpenAI as in the work done by Shahandashti et al., 2024. On this occasion, we seek to evaluate the improvements and additional capabilities of GPT-4-Turbo in comparison with the proposal made for the same task in the SimpleText 2023 workshop where the results obtained were outstanding, demonstrating the effectiveness of the GPT-3 model, text-davinci-003 version (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023b).

The Table 4.69 presents several results of sub-task 2.2 applying GPT-4-Turbo with zero-shot learning according to the PRM_ZS_TASK2_V2 prompt. As we can see, the table illustrates an example that corresponds to Snt_id G01.1_1000902583_1. In the solution proposed for sub-task 2.2, the model generates a definition, an explanation, and also generates a use case illustrating

GPT-4-Turbo with Zero-Shot learning

Manual	Snt_id	Term	Difficulty
0	G01.1_1000902583_1	exploited	d
0	G01.1_1000902583_1	confrontation	m
0	G01.1_1000902583_1	coordination	m
0	G05.1_2914002216_4	CRISPR/Cas9	d
0	G05.1_2914002216_4	reagents	m
0	G05.1_2914002216_4	high-content screen	d
0	G07.2_2773680786_5	conspiracy	m
0	G07.2_2773680786_5	emergent	m
0	G07.2_2773680786_5	pervasive	d
0	G11.1_2946157960_5	consideration	m
0	G11.1_2946157960_5	applications	e
0	G11.1_2946157960_5	deployment	m
0	G11.1_2946157960_5	networked	d
0	G11.1_2946157960_5	emission	m

TABLE 4.68: Predictions generated by applying GPT-4-Turbo with zero-shot learning in Sub-task 2.1-Prompt PRM_ZS_TASK2_V1.

in a broad way the complexity of the terms for different groups of users, making suggestions for definitions, explanations and use cases based on the context of the sentence (Abstract).

As we can see, the table illustrates an example that corresponds to Snt_id G01.1_1000902583_1. In the solution proposed for sub-task 2.2, the model generates a definition, an explanation, and also generates a use case illustrating in a broad way the complexity of the terms for different groups of users, making suggestions for definitions, explanations and use cases based on the context of the sentence (Abstract).

Table 4.70, we present the official results published by the organizers (Ermakova et al., 2024), including the scores obtained in our participation in subtasks 2.1 and 2.2. The meaning of each column is the following:

- **recall overall:** the proportion of terms (independently from the difficulty) that were found.
- **recall average:** the average of the recall of terms computed per sentence.
- **recall difficult terms:** the proportion of difficult terms that were found.
- **precision difficult:** the precision of terms that were labeled as difficult.
- **bleu_nx:** the BLEU score computed with ngrams $n = 1, 2, 3, 4$.

Snt_id: G01.1_1000902583_1

Abstract:

Novel technological advances in mobile devices and applications can be exploited in wildfire confrontation, enabling end-users to easily conduct several everyday tasks, such as access to data and information, sharing of intelligence and coordination of personnel and vehicles.

Term	Definition/Explication
<i>exploited</i>	Definition: Make full use of and derive benefit from (a resource). Explication: In the context of the sentence, exploited refers to utilizing the new technological capabilities of mobile devices and applications to their fullest potential in managing and combating wildfires. Use case: The novel technologies can be exploited to improve the efficiency and effectiveness of wildfire management.
<i>confrontation</i>	Definition: A hostile or argumentative meeting or situation between opposing parties. Explication: Here, confrontation is used metaphorically to describe the act of dealing with or managing wildfires, which are seen as the opposing party to human safety and property. Use case: Technological advances aid in the confrontation of wildfires by providing better tools for managing the crisis.
<i>coordination</i>	Definition: The organization of the different elements of a complex body or activity so as to enable them to work together effectively. Explication: In this context, coordination refers to the organized management of personnel and vehicles, ensuring they operate in a synchronized and efficient manner during wildfire emergencies. Use case: Effective coordination of personnel and vehicles is crucial for rapid and efficient wildfire suppression.

TABLE 4.69: Predictions generated by applying GPT-4-Turbo with Zero-Shot learning in Sub-task 2.2 - Prompt PRM_ZS_TASK2_V1.

Official Results of SimpleText CLEF-2024 Task 2

runid	recall_ overall	recall_ average	recall_ difficult _terms	precision difficult _terms	bleu_ n1_ average	bleu_ n2_ average	bleu_ n3_ average	bleu_ n4_ average
AIIRLab_Task2.2_LLaMA	0.2792	0.3011	0.2642	0.6667	0.2883	0.1519	0.0497	0.0191
AIIRLab_Task2.2_LLaMAFT	0.0069	0.0056	0.0047	1.0000	0.2405	0.1171	0.0000	0.0000
AIIRLab_Task2.2_Mistral	0.4118	0.4415	0.1863	0.4907	0.2610	0.1338	0.0395	0.0128
Dajana&Kathy_SimpleText _Task2.2_LLAMA2_13B_CHAT	0.0118	0.0114	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
FRANE_AND_ANDREA _SimpleText_Task2.2_LLAMA2 _13B_CHAT	0.0076	0.0066	0.0094	0.3636	0.0000	0.0000	0.0000	0.0000
ruby	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Sharingans_Task2.2_GPT	0.4722	0.5302	0.5448	0.5954	0.2257	0.1039	0.0300	0.0160
SINAI_task_2_PRM_ZS_ TASK2_V1	0.0868	0.0872	0.1014	0.5244	0.2545	0.1579	0.0821	0.0578
SINAI_task_2_PRM_ZS_ TASK2_V2	0.1556	0.1636	0.1297	0.7746	0.2774	0.1574	0.0630	0.0443
SINAI_task_2_PRM_ZS_ TASK2_V3	0.0951	0.1045	0.0472	0.8333	0.2144	0.1113	0.0377	0.0229
team1_Petra_and_Regina_ Task2_ST	0.0042	0.0042	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Tomislav&Rowan_Task2.2_ LLAMA2_13B_CHAT	0.0069	0.0040	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Tomislav&Rowan_Task2.2_ LLAMA2_13B_CHAT_1	0.0083	0.0084	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
UAms_Task2-1_RareIDF	0.0854	0.0942	0.0259	0.0894	0.0001	0.0000	0.0000	0.0000
UboNLP_Task2.1_phi3-oneshot	0.5424	0.5620	0.3160	0.3743	0.0011	0.0000	0.0000	0.0000
unipd_t21t22_chatgpt	0.1340	0.1400	0.0825	0.6250	0.3045	0.1851	0.0905	0.0507
unipd_t21t22_chatgpt_mod1	0.2194	0.2371	0.1981	0.5957	0.3060	0.1783	0.0802	0.0430
unipd_t21t22_chatgpt_mod2	0.3146	0.3155	0.3420	0.6905	0.0302	0.0069	0.0031	0.0000

TABLE 4.70: Results for CLEF 2024 SimpleText Task 2 for the quality of definitions.

According to the results presented in table 4.70, the UboNLP_Task2.1_phi3-oneshot team showed the best overall performance in terms of recall_overall (0.5424) and recall_average (0.5620), which indicates a high ability to retrieve relevant terms. However, its precision in difficult terms is relatively low (0.3743). Sharingans_Task2.2_GPT also performs well with a recall_overall of 0.4722 and a recall_average of 0.5302, in addition, it has a high recall_terms (0.5448), demonstrating a good ability to handle difficult terms.

According to the performance analysis in the difficult terms, the AIIRLab_Task2.2_LLaMAFT and SINAI_task_2_PRM_ZS_TASK2_V3 teams have a precision_difficult_terms of 1.0000 and 0.8333 respectively, which indicates that these models are very precise with difficult terms, although their recall_overall and recall_average are not as high in comparison with other equipment.

In the performance in BLEU Scores, the unipd_t21t22_chatgpt teams and their variants (mod1 and mod2) have the best bleu_n1_average, bleu_n2_average, bleu_n3_average and bleu_n4_average demonstrating good performance in terms of precision of n-grams in text generation. In the case of AIIRLab_Task2.2_Mistral, this team has a good balance with a recall_overall of 0.4118 and precision_difficult_terms of 0.4907, suggesting that their model is quite robust in general terms and in difficult terms. The SINAI team variants show variations in their results, with SINAI_task_2_PRM_ZS_TASK2_V2 having the best performance among its variants, with a recall_overall of 0.1556 and precision_difficult_terms of 0.7746. There are teams with values of 0 in almost all metrics, which indicates very low performance or problems in the implementation of their models.

The model has proven to be able to generate robust responses based on the text or instruction (message) provided. We observe that GPT-4-Turbo analyzes the text, uses its extensive knowledge to identify words that can be considered complex, and categorizes them according to their complexity. Our preliminary analysis evaluates the model's ability to understand and generate arguments in specific contexts. The results show that GPT-4-Turbo is highly competent in natural language processing tasks, such as predicting lexical complexity, demonstrating its effectiveness in identifying and categorizing complex terms accurately and consistently.

We have applied the GPT-4-Turbo model in the construction of several solutions for sub-tasks 2.1 and 2.2, where the model has demonstrated robust performance in the execution of natural

language processing tasks, specifically in lexical simplification and identification of complex words. Not only has it successfully tackled these tasks, but it has also been able to generate detailed illustrations by creating definitions, explanations and use cases based on the context of the sentence (text). By identifying complex words and phrases and offering simpler versions, it facilitates access to scientific sources in an understandable way, helping readers to understand the complexity of terms in different user groups, which makes it very useful for a general audience. and especially valuable for readers with reading difficulties or cognitive disabilities. We confirmed our hypothesis that generative models, such as autoregressive language models, can perform the task of complex word identification (CWI) thanks to their ability to generate text sequences and model word complexity based on context.

The results obtained show that these models effectively identify terms that are more difficult for users, generating accessible summaries and providing texts that facilitate understanding. This reinforces the feasibility of applying these models to the CWI task. The results showed a variety of performances among the participating teams. Some teams, such as UboNLP_Task2.1_phi3-oneshot and Sharingans_Task2.2_GPT, stand out for their overall performance and ability to handle difficult terms, while other teams show specific strengths such as accuracy on difficult terms or good BLEU scores. However, there are also teams with significantly low performances, which could indicate the need for improvements in their approaches or implementations.

Chapter V

Conclusions

5.1 Main Conclusions

This chapter presents a detailed analysis of the methods and results applied in research on the identification of complex words in the academic domain in Spanish. Accurate identification of complex terms is essential in the academic context, as it directly affects students' comprehension and learning, as well as teachers' ability to communicate effectively.

The objective of this chapter is to perform this analysis of results according to the different approaches, domains and languages (Spanish and English) considered during this research. Both traditional methods based on linguistic features and advanced machine learning models will be evaluated, including those using neural network architectures such as autoregressive language models. Emphasis will be placed on comparing the effectiveness of various techniques, highlighting their strengths and limitations. In addition, the impact of integrating domain-specific linguistic features and applying zero-shot and few-shot learning approaches on model accuracy will be analyzed.

There are three main types of systems according to the state of the art of modern NLP: feature-based systems, deep learning systems and hybrid systems that combine both approaches (Shardlow et al., 2021a). In our research, we have applied these three approaches in order to validate our hypothesis: H1. *“The implementation of large linguistic models complemented with*

linguistic features emerges as a beneficial solution to identify and predict highly complex words in academic texts in Spanish.” To strengthen our hypothesis, we have explored the results using the aforementioned approaches, focusing on the study of linguistic characteristics and their impact on our findings. Through this exhaustive analysis and discussion, we seek in our field of study: to enrich a debate with the findings of other research, and to contribute to the advancement of knowledge.

We confirm our hypothesis H2. “*Generative models, like auto-regressive language models can accomplish the task of CWI*” thanks to their ability to generate text sequences and model word complexity based on context. The results obtained show that these models effectively identify terms that are more difficult for users, generating accessible summaries and providing texts that facilitate understanding. This reinforces the feasibility of applying these models to the CWI task.

Below we present the results and discussion according to each approach.

- Applying Linguistic Features with Classical Machine Learning algorithms: the results provided us with a better understanding of how different variables interact and contribute to our overall results, as has been detailed in previous sections [4.3.2](#) and [4.3.3](#).
- Applying Linguistic Features with Transformers Embeddings (pre-trained) executed on various Classic Machine Learning algorithms in Spanish and English dataset: we have examined in detail the use of deep learning model and its capabilities for generating numerical embeddings, for further combination with linguistic features. The sections [4.3.4](#) and [4.3.5](#) gave us a broader view of the influence of the model on a data set with features of different nature in our study, which allowed us to analyze how the choice of the model impacts our results and to what extent they influence its interpretation.
- Applying End-to-End Architecture (fine-tuning) with linguistic feature integration in Spanish corpus: to this end, we presented in section [4.3.6](#) the discussion of the *exploration of the model results by applying fine tuning*, highlighting its importance in the performance of the model by obtaining more precise results, which contributes to the quality of our findings.

- Applying the Use of the Unsupervised Learning Algorithm (generative models): in section 4.3.7 we presented our analysis based on the results of GPT-3 and its later versions GPT-3.5 and GPT-4-turbo, thus enriching our study and complementing our research.

We started our research based on feature based systems use a wide variety of attributes that are considered fundamental for the assessment of lexical complexity, among the most relevant of which are word length and lexical frequency (Shardlow et al., 2021a). “Linguistic features” are computed as classical features in computational linguistics, and “semantic features” are those that result from the word and sentence level embeddings described above (Liebeskind, Elkayam, and Liebeskind, 2021; Mosquera, 2021a).

The results obtained in our participation in Task 1: Lexical Complexity Prediction of SemEval 2021, we run the English data set called CompLex using the Random Forest Regressor algorithm. Furthermore, we enriched the data set by incorporating 15 additional linguistic features, where we obtained a MAE = 0.0875, while the winner of the competition obtained a MAE = 0.0609.

Subsequently, we developed a solution for our participation in the international competition RaNLP 2021. In this context, with the aim of analyzing the performance of the algorithm with a different data set, that is, this time in Spanish language, we created a corpus in Spanish called CLexIS², and we apply a process similar to that used in SemEval 2021, we use the Random Forest Regressor algorithm and apply the same 15 linguistic features, where we obtained a MAE = 0.1742.

Finally, we execute the BERT and RoBERTa models, from which we have obtained the encodings. We integrate these encodings into the data set, combining them with the linguistic features. A fairly important distinction in this research process is the increase of our set of linguistic features to 23. We can highlight that, when running the data set in Spanish, the algorithm with 15 features results in a MAE = 0.1742, while that, when executing the data set confirmed by the 23 features and the encodings, the algorithm presents a better performance, resulting in a MAE = 0.1592. For the English data set, applying the 23 features plus the encodings executed in different algorithms, with the GBR algorithm a MAE = 0.0688 is achieved.

Table 5.1 presents the best results in terms of MAE about our applied approach, where we have integrated both linguistic features and encodings. We carry out various experiments with

data sets in different languages and texts from various origins, running them on supervised learning algorithms. The most notable results obtained in our research are detailed below.

Our hypothesis has been confirmed in the present thesis research, given that the wealth of knowledge present in transformer-based models can help extract complementary clues of contextual complexity (Paetzold, 2021). We highlight a significant finding: *the incorporation of linguistic features in conjunction with vector representations of words, and the increase in the number of epochs during the executions of large language models, leads to a substantial improvement in the performance of the algorithms to predict the lexical complexity*, as evidenced in our studies (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2022a), (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023a). It is graphically shown in Figure 6.2.

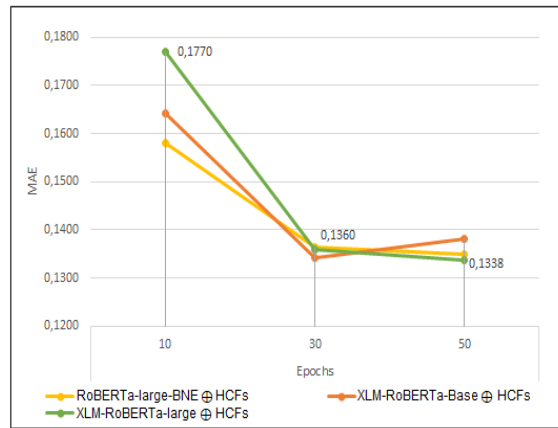


FIGURE 5.1: Integration of HCF and increase in times.

Another outstanding result that confirms our hypothesis is the solution presented in the proposed task “Task 2: What is unclear? Difficult concept identification and explanation (definitions, abbreviation deciphering, context, applications,...)” by the international workshop SimpleText@CLEF-2023. Our results in task 2.1 showed a significant performance of the proposed approach close to top performing systems. It is interesting to note that, in general, zero-shot learning prompts are better at detecting a difficult word, while few-shot learning prompts are better at determining a difficulty score. The use of the generative language GPT-3 by applying the Zero-Shot learning technique generated extremely beneficial results for identifying and predicting words (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023b) by giving us the first place among the participants (Ermakova et al., 2023). (It is important to note that the detailed results are presented exhaustively in Chapter IV).

Features of different nature with Classical Machine Learning algorithms				
Best Results				
Corpora Language	CompLex <i>English</i>		CLexIS2 <i>Spanish</i>	
Classical ML Algorithms	RFR	GBR	RFR	RFR
Encodings	No	Si	No	Si
MAE	0.0875	0.0688	0.1742	0.1592
MSE	0.0131	0.0118	0.0442	0.0379
RMSE	-	0.2103	0.0992	0.1982
Accorging MAE Winning Team Scientific event	0.0609	0.0609	-	-
	<i>SemEval</i> 2021	<i>Electronics</i> 2023	<i>RaNLP</i> 2021	<i>SEPLN</i> 2022
Authors	(Zambrano and Montejo-Ráez, 2021)	(Ortiz- Zambrano, Espin-Riofrio, and Montejo- Ráez, 2022a)	(Ortiz- Zambrano and Montejo- Ráez, 2021)	(Ortiz- Zambrano, Espin-Riofrio, and Montejo- Ráez, 2022b)

TABLE 5.1: Best results applying features with Classical Machine Learning algorithms presented in various scientific events.

5.2 Discussion

We carried out exhaustive research on complex words in English and Spanish, obtaining very promising results. The inclusion of linguistic features as part of the data set proved to be particularly effective, significantly improving the accuracy of our analyses. Regarding task 1 of SemEval 2021, the organization provided continuous annotations, allowing them to classify the words not only in terms of whether they are complex or not, but also to what extent they are, achieving a more nuanced evaluation of the lexical difficulty. The competition provided a data set spanning three diverse genres, also incorporating multi-word expressions (MWE) and individual tokens, using a Likert scale that allowed detailed and accurate assessment of lexical complexity in various contexts (Shardlow et al., [2021a](#)). We observe that there is a limitation of the task is its exclusive focus on the English language, which may restrict the generalization of the findings to other languages, as is the case of our research that is developed for the Spanish

language.

According to Shardlow et al., [2021a](#), the results of SemEval 2021, deep learning-based systems achieved the highest Pearson correlation in the subtask 1 of Lexical Complexity Prediction, occupying the first two places like those proposed by the teams ran the datasets with the BERT and RoBERTa models, such as JUST BLUE, RG PA, Andi, CS-UM6P, and OCHADAI-KYOTO, just to name a few, while feature-based systems are not far behind, reaching third and fourth place with a similar score to the best systems. The feature-based systems approach uses a variety of features that are known to be useful for lexical complexity. In particular, lexical frequency and word length figure heavily in many different ways of calculating these metrics. Psycholinguistic features that shape people's perception of words were understandably popular in this task, since complexity is a perceived phenomenon.

Semantic features taken from WordNet that model the meaning of the word and its ambiguity or abstraction have been widely used, as well as sentence-level features with the aim of modeling the context around the target words. Finally, word embeddings are a popular feature, coming from static resources like GLoVE or Word2Vec, but also derived from using Transformer models like BERT, RoBERTa, XLNet or GPT-2, which provide context-dependent embeddings suitable for our task.

We proposed a simple approach, calculated a total of 15 linguistic features and added them as part of the dataset, then applied machine learning and built the model using random forest regression algorithm, relying on contextual and known word-based features. We train the algorithm, and predict the results whose results obtained with a focus on simple word prediction are: $MAE = 0.0875$, $MSE = 0.0131$ and $R^2 = 0.1930$ (Zambrano and Montejo-Ráez, [2021](#)). Considering that the number of competitors is quite large (54 teams) and that the result presented by the first place winner is a MAE of 0.0609, which allows us to conclude that there is a small difference despite being a very simple system.

We thus highlight their importance as a relevant role in predicting the complexity of words, since they have shown that they are not only essential to improve the accuracy of predictive models, but are also highly valued and requested by numerous researchers to be included in their data sets as demonstrated by the results of the competition, becoming essential in any rigorous linguistic analysis in Lexical Simplification tasks.

With the aim of achieving optimal performance of classical supervised learning algorithms regarding word complexity prediction focused on the English language, we carry out our research by applying the approaches carried out by a large number of researchers who executed this approach in their proposals in task 1 of SemEval 2021, we refer to the encodings generated by transformer-based models, specifically BERT and RoBERTa (Shardlow et al., 2021a). Integrating linguistic features with the encodings derived from these pre-trained models was shown to consistently improve results.

This positive effect is evident in the exploration of features at both the word and sentence level using transformers, applied to several supervised learning algorithms. We obtained a MAE = 0.06897 when combining the BERT-W $\oplus$ BERT-S $\oplus$ LF encodings, and a very close score with XLMR-W $\oplus$ XLMR-S $\oplus$ LF reaching a MAE = 0.06897. It is important to note that both models, in the data set, include linguistic features and encodings processed using the same regression algorithm, Gradient Boosting Regressor (GBR) being the best result so far (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2022a).

These results conclusively highlight the relevance and effectiveness of transformer-based models in integrating advanced coding techniques with linguistic features. This combination not only significantly improved the accuracy of the predictive models in diverse and complex linguistic contexts, but also resulted in a MAE = 0.0689. This value represents a substantial improvement of 0.0186 units compared to previous results. Furthermore, this improvement was achieved using a different model, the Gradient Boosting Regressor (GBR), whose effectiveness was demonstrated in the BERT and RoBERTa models. These findings reaffirm the importance of the linguistic features proposed by numerous researchers within the framework of Task 1 of SemEval 2021 (Shardlow et al., 2021a).

For the performance analysis of Classical Machine Learning algorithms in Spanish, we used the CLexIS² corpus. Initially, we added 15 linguistic features to the dataset. During testing with various algorithms, we found that the Random Forest Regressor (RFR) performed the best, achieving an MAE = 0.1742 (Ortiz-Zambrano and Montejo-Ráez, 2021). This result indicates that RFR is not only effective on English data (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2022a), but also offers superior performance on Spanish data when using the same linguistic features. This consistency across different languages highlights the robustness and versatility of

RFR in processing linguistic data.

To improve the results in predicting complex words, we implemented the BERT, XLM-RoBERTa, and RoBERTa-large-BNE models in the analysis of the CLeXIS² corpus for the Spanish language. These models, widely used in the SemEval 2021 LCP competition for the Lexical Complexity Prediction (LCP) task (Shardlow et al., 2021a), have proven to be highly effective.

In our research, we integrated the encodings generated by these models into our datasets, which initially contained 15 linguistic features. We added 8 additional features derived from the models' encodings, reaching a total of 23 features.

The results of our final evaluation showed that the embeddings provided by the Spanish language models significantly improved the prediction of lexical complexity. With this integration, we managed to reduce the mean absolute error (MAE) from 0.1742 to 0.1592 (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2022b), thus surpassing previous results. This breakthrough demonstrates that the combination of linguistic features and language model encodings significantly increases the performance of supervised learning algorithms in the task of lexical complexity prediction.

In the experiments conducted so far, no fine-tuning of the deep learning models has been carried out. Instead, the language models were used solely to generate feature vectors. These vectors were sometimes combined with additional linguistic features and fed into Classical Machine Learning algorithms to carry out the classification tasks.

In the latter, we explored a different approach by adding to our research a set of experiments, where we performed fine-tuning where the entire network could be fine-tuned by injecting the linguistic features. This process allowed the network to be holistically optimized, potentially improving its ability to handle linguistic features and thus improving the overall performance of the model. Based on running different feature combinations, we obtained a MAE of 0.1323 with the Spanish CLeXIS² corpus, the best performance being achieved using the fine-tuned version of the BERT model in Spanish. In the case of the english *CompLex* corpus, with a MAE = 0.0683, BERT fine-tuning was again the best choice, but only if linguistic features are introduced into the model (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024a).

We applied fine-tuning to the new Spanish *LegalEc* corpus where the performance improvement seems to be related to the number of epochs plus the inclusion of additional features. Using these features together with encodings from pre-trained models proved favorable for performance. The best result was obtained by applying the XLM-RoBERTa-large + LF model with a MAE = 0.1338, after training for 50 epochs. See Table 5.2. The other models also reported higher performance during the different runs. In this research we have shown how the inclusion of linguistic features and increasing epochs substantially improve the prediction of lexical complexity, even for smaller models such as XLM-RoBERTa-base (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023a).

Final results of the fine-tuned end-to-end approach				
Corpora	Model	MAE	MSE	RMSE
CLexIS ²	BERT	0.1323	0.0368	0.1918
LegalEc	XLM-RoBERTa-large $\oplus$ LF	0.1338	0.0252	0.1587

TABLE 5.2: Running the pre-trained models on the Spanish datasets.

The results obtained by the GPT-3 auto-regressive language model on Simple@Text 2023 demonstrated significant performance, which closely matches that of the best-performing systems. It is noteworthy that, overall, version 1 of the message yielded better results. Zero-shot learning messages were more effective in detecting difficult words, while few-shot learning messages excelled in determining difficulty scores. We achieved high values in semantic matching. For providing definitions or explanations of difficult terms, few-shot learning messages consistently outperformed zero-shot learning messages.

5.3 Scientific Evaluation And Future Work

In the final phase of this research, we presented the main conclusions derived from our study on the identification of complex words in the academic domain of the Spanish language. These findings were presented at the Doctoral Symposium on Natural Language Processing within the ILENIA Project (PLN-DS-2023), held as part of the XXXIX International Congress of the Spanish Society for Natural Language Processing - SEPLN 2023 (Ortiz-Zambrano, 2023). We evaluated the validity of the hypothesis proposed in Section 1.2 examining various methodological approaches and evaluating multiple techniques to enhance the accuracy of complex term detection within academic texts. This study has not only deepened our understanding of the challenges and opportunities in this field but has also provided valuable insights into how to optimize

learning processes and improve accessibility in academic environments.

The results obtained in this study strongly support our hypotheses 1.2, which maintains that the implementation of large-scale language models complemented with linguistic features is a beneficial solution to identify and predict highly complex words in academic texts in Spanish. Through a series of detailed experiments and analyses, we demonstrate that the application of language models and linguistic features significantly improves performance on complex word identification and prediction tasks.

Additionally, the limitations of the study are presented and how these can be addressed in directions for future research to continue advancing the development of tools that facilitate reading comprehension in academic contexts. The purpose of these conclusions is to synthesize the knowledge gained, highlight the contribution of this thesis to the field of computational linguistics and education, and offer concrete recommendations that can be implemented to improve the identification of complex words and ultimately support success academic of Spanish-speaking students.

The proposed hypothesis:

H1. *The implementation of large language models complemented with linguistic features, emerges as an beneficial solution to identify and predict words of high complexity in academic texts in Spanish.*

Based on the data analysis we formulate conclusions about the validity of the hypothesis:

We present and analyze the conclusions obtained in relation to the specific objectives set out in Section 1.2, highlighting how the findings of this research respond to each of them.

First Specific Objective: Develop new benchmark corpora as essential resources for the development of research on linguistic complexity in academic texts in Spanish.

We have managed to meet the first specific objective of developing new reference corpora as fundamental resources for the study of linguistic complexity in academic texts in Spanish, for which several essential activities were carried out. First, a wide variety of academic texts from various disciplines were collected, including scientific articles, theses and textbooks, articles

of the Ecuadorian constitution, final degree projects, public texts, among others. These texts were carefully annotated to identify and classify complex words, a task that was carried out by students and people in general. In addition, automatic annotation tools were developed using machine learning and natural language processing algorithms, trained with the manually annotated data to improve their accuracy and efficiency.

The new corpora created as own resources and detailed extensively in the section 4.2 constituted an essential component in the investigation of complex words, becoming a solid base as a starting point, as well as the empirical data necessary to identify , analyze and understand the complexity of texts in the Spanish language in a specific context such as the academic domain.

These resources were presented at various international scientific conferences on semantic evaluation and recent advances in the field of Natural Language Processing (NLP). We highlight the *VYTEDU* corpus: A corpus of videos and their transcriptions for educational research (Ortiz-Zambrano and Montejo-Ráez, 2017). *VYTEDU* marked the starting point of this research. For its evaluation and validation, lexical complexity metrics were applied for Spanish (Ortiz-Zambrano and Varela Tapia, 2018). It was presented at the XXXIII International Congress of the Spanish Society for Natural Language Processing (SEPLN 2017).

The *VYTEDU-CW* corpus (Ortiz-Zambrano et al., 2019) was developed from the labeling process of the *VYTEDU* corpus, with the aim of preparing it for evaluation through its application with language models based on transformers, since these models have proven to be highly effective in identifying and predicting lexical complexity. The analysis of the words labeled as complex allowed a detailed characterization of these words in the *VYTEDU-CW* (Ortiz-Zambrano and Montejo-Ráez, 2020) corpus. This corpus was consolidated as an essential resource for new research in the field of Lexical Simplification, being an integral part of our workshop ALEXS 2020 - Lexical Analysis at SEPLN 2020 (Ortiz-Zambrano and Montejo-Ráez, 2020) in an active participation as part of the Iberian Forum for the Evaluation of Languages - IberLEF 2020.

The *CLexIS²* corpus: A new collection for research on the identification of complex words in computer science studies (Ortiz-Zambrano and Montejo-Ráez, 2021), is made up of texts from Computational Science. This corpus was developed with the purpose of implementing new evaluations, which go beyond conventional ones, including classic machine learning algorithms.

*CLexIS*² was presented at the Recent Advances in Natural Language Processing (RANLP 2021) conference, highlighting its innovative contribution to the field of natural language processing and lexical simplification.

The *LegalEc* corpus: An innovative resource for research into the identification of complex words in legal studies of Ecuadorian Spanish (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023a), has been created with the purpose of advancing the exploration and understanding of lexical complexity in the field of law . Presented at the prestigious XXXIX International Congress of the Spanish Society of Natural Language Processing - SEPLN 2023, LegalEc constitutes a valuable contribution to the study of legal language, facilitating the development of tools and methods to improve the accessibility and understanding of legal texts in Spanish Ecuadorian. Its evaluation process was carried out applying models based on transformers.

The resulting corpora were validated and evaluated using rigorous methods, ensuring their reliability and consistency. The application of lexical complexity and common error metrics allowed us to exhaustively evaluate the complexity of the content of the texts of the various corpora. This analysis revealed to us that the lexicon used by teachers when teaching their classes includes complex sentences, as well as technical and sophisticated language, which makes it difficult for students to understand. Likewise, we observe that the texts used by students for learning contain similar elements that complicate their understanding. These findings are just a few notable examples from our study.

GovAIEc: 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.

These resources were made available to the scientific community, you can follow this link to see them <https://repositorybeta.github.io/plnec/>, thus facilitating their use by the academic and research community. These resources not only support the present research, but also provide a robust pillar for future studies on linguistic complexity in the Spanish academic field. In this way, the precision of language models is improved and contributes to the development of more effective and accessible educational tools, promoting greater reading comprehension in university and government contexts.

Second specific objective: Explore advanced strategies for Complex Word Identification (CWI) and Lexical Complexity Prediction (LCP) in data sets, such as the application of artificial neural networks, in particular deep learning architectures such as Transformers-based language models, various Natural Language Processing (NLP) techniques, classic machine learning algorithms, and linguistic features calculated at the word level. Subsequently, evaluate the performance of the models by analyzing the generated results.

This specific objective has been widely addressed throughout the research, exploring various advanced strategies and technologies for the identification of complex words and the prediction of lexical complexity in Spanish academic texts. Special emphasis has been placed on the use of artificial neural networks, particularly Transformer-based architectures, recognized for their ability to capture and process complex contextual information effectively. These models have shown promise in the task of CWI and LCP, allowing for a more accurate and detailed identification of linguistically complex terms.

Not only Transformer-based language models have been explored, but also various Natural Language Processing (NLP) techniques have been applied, including classical machine learning algorithms and word-level linguistic features (Zambrano and Montejo-Ráez, 2021), (Ortiz-Zambrano and Montejo-Ráez, 2021), a combination of these techniques (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2022b), (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2022a), and end-to-end approach (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023a), (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024a), which has enriched the analysis and prediction of lexical complexity, thus providing significant results about the difficulty of Spanish academic texts.

The evaluation of the performance of the models has been a very rigorous process, analyzing

the results generated in terms of precision, recall, and other relevant metrics with those of common error. These analyses allowed us to identify strengths and areas of improvement in the proposed models, providing recommendations for future research and practical applications in the educational and academic field. The exploration of advanced strategies has contributed significantly to the advancement of knowledge in CWI and LCP, highlighting the importance of technological innovation and the application of sophisticated methods in computational linguistics.

Third specific Objective: Integrate linguistic features into classical machine learning algorithms and deep neural network models for performance evaluation on different data sets in both Spanish and English.

This specific objective has been addressed through a systematic approach that integrates linguistic features into classical machine learning algorithms as well as deep neural network models.

It is pertinent to emphasize that linguistic features has a highly relevant role in predicting lexical complexity. Not only are they essential to optimize the accuracy of predictive models, but they also enjoy high esteem and demand among numerous researchers, who seek to incorporate them into their data sets. The importance of these features in the field of lexical simplification highlights the need to consider them in any exhaustive linguistic analysis.

We started by identifying and calculating several relevant linguistic features at the word level, such as word length, frequency of use, morphological complexity, and syntactic features. These features were chosen based on their potential to capture relevant aspects of lexical complexity in academic texts in Spanish and English. The computed linguistic features were run on several classical machine learning algorithms, including Support Vector Machines (SVM), Decision Trees, and Random Forest, among others. These algorithms were trained and evaluated using Spanish (Ortiz-Zambrano and Montejo-Ráez, 2021), (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2022b), (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024a) and English datasets (Zambrano and Montejo-Ráez, 2021), (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2022a), (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024a). The integration of linguistic features significantly improved the performance of these models in identifying complex words and predicting their lexical complexity.

In addition, we executed the BERT, XLM-RoBERTa, XLM-RoBERTa-base, and RoBERTa-large-BNE models on the English and Spanish datasets using the end-to-end approach. The best results were demonstrated by the BERT model in combination with linguistic features for both the English and Spanish datasets (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024a). It is important to note that the performance of the models depends on the number of epochs (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023a). However, deep learning models function as 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 information related to syntax, dependencies, grammar, gender, negation, semantics, among others, within their layers.

The use of linguistic features and transformer-based models applying the end-to-end approach appeared to consistently improve the results. The performance gain appears to be related to language, which could be related to a higher quality of performance in the English and Spanish (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024a) transformer models or to a problem of task difficulty depending on the language. These findings underscore the relevance and effectiveness of combining advanced coding techniques with traditional linguistic features to improve the accuracy of predictive models in multilingual and multifaceted contexts.

It is worth emphasizing the relevance of the BERT and RoBERTa models have proven to be an essential type of neural network architecture in this area. Their implementation in our research with Spanish datasets has been particularly valuable. The results of the final evaluation reveal that the Spanish language models contributed significantly to the prediction of lexical complexity, outperforming the initial proposals that only used linguistic features. The combination of these features with the encodings generated by the models, applied in classical machine learning algorithms, resulted in a substantial improvement in performance (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2022b). As future work, we plan to study ablation tests to realize how different features contribute to better performance.

In conclusion, running different feature combinations, with a MAE = 0.1323 on the Spanish CLexIS² corpus, the best performance was obtained using the fine-tuned version of the Spanish BERT model. For untuned approaches, the AdaBoost algorithm with XLM-RoBERTa sentence and word encodings together with linguistic features reported high performance with a MAE

= 0.1623 and for the English CompLex corpus, with a MAE of 0.0683, fine-tuning BERT was again the best choice, but only if linguistic features are introduced into the model and SVM was the second best algorithm considering the pre-trained model and taking the target word embeddings from BERT with the proposed hand-crafted (linguistic) features (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024a).

Fourth Specific Objective: Explore and employ unsupervised techniques, like auto-regressive models, for the analysis of Lexical Complexity Prediction (LCP) to examine the capabilities of generative pre-trained large language models to solve this task.

The proposed hypothesis:

H2. *Generative models, like auto-regressive language models, can accomplish the task of CWI.*

This specific objective has been addressed by exploring and applying unsupervised techniques, particularly auto-regressive models, such as GPT-3, GPT-3.5, GPT-4-turbo and GPT-4o for the analysis of Lexical Complexity Prediction (LCP).

Based on the data analysis we formulate conclusions about the validity of the hypothesis:

The GPT-3 autoregressive model and its versions davinci-002, davinci-003 demonstrated their performance on English data sets when analyzing. An unsupervised approach was used, which allowed the model to learn the complexities of the language from a few examples, where the model has been able to interpret and generate its responses based on a few examples and complex instructions, demonstrating that the text-based version Davinci-003 provides better results than text-Davinci-002 (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024b)

The results obtained by the model demonstrated solid performance. These results were significant in overcoming the challenges associated with text simplification methods, with the aim of facilitating access to scientific information (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023b). The applied evaluation metrics such as BLEU, precision, recall, measurement and common error were applied to evaluate the effectiveness of the autoregressive models using. The results indicated that generative models such as GPT-3 and its later versions are capable of

capturing contextual nuances that are critical for accurate prediction of lexical complexity.

The model demonstrated strong performance in obtaining results in experiments applying zero-shot and few-shot approaches, generating high-quality samples. Furthermore, the model shows that learning performed with few samples has the ability to generalize from limited information. Few-shot learning has been applied to a task never performed before (Automatic Simplification of Scientific Texts) where it has been shown to overcome the limitation of large training data sets by allowing the model to learn quickly and effectively with few examples.

We have applied the GPT-4-Turbo model in the construction of several solutions for sub-tasks 2.1 and 2.2 in SimpleText@CLEF 2024 (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024), where the model has demonstrated robust performance in executing natural language processing tasks, specifically in lexical simplification and the identification of complex words. It has not only successfully addressed these tasks but has also been able to generate detailed illustrations by creating definitions, explanations, and use cases based on the context of the sentence (text). By identifying complex words and phrases and offering simpler versions, it facilitates access to scientific sources in a comprehensible manner, helping readers understand the complexity of terms across different user groups, making it very useful for a general audience and especially valuable for readers with reading difficulties or cognitive disabilities. We confirm our hypothesis that generative models, such as autoregressive language models, can perform the task of complex word identification (CWI) due to their ability to generate sequences of text and model the complexity of words based on context.

As a product of the study carried out, an innovative application was developed that converts complex texts into more accessible versions, you can visit the website at the following link <https://www.govaieasy.com/>. This tool is designed to benefit college students and the general public, with a particular focus on people with disabilities or reading comprehension difficulties. By significantly improving the accessibility and comprehension of texts, the application contributes to the inclusion and academic success of a wide diversity of users.

5.4 Limitations

The BERT and RoBERTa Transformer-based language models were applied to a set of features from a training corpus of simple English words. Our solution is not geared towards multi-words. We see it convenient to apply other language models that allow us to evaluate the results as a possible improvement in the level of word prediction. The value of R^2 obtained in our experiments tells us that only 19% of the variation of the complexity can be explained from our features, which is minimal. This suggests that further experimentation must be done with additional features and ablation tests to identify which kinds of features (embeddings, complexity related, stylistically related, among others.) contribute more to the identification of difficult words.

The experimentations made and the proposed systems for lexical simplification can help other researchers in their work to explore such a design. Some language models and most classical algorithms do not demand high computing resources and could be a solution in low-resourced environments, such as mobile devices and the like.

Capítulo VI

Identificación de las Palabras complejas en el Dominio Académico en Español

Resumen escrito en Español

6.1 Introducción

La legibilidad se define como la cualidad que hace que un texto sea más accesible y fácil de leer (DuBay, 2004), pero para muchas personas, la forma en que está escrito un texto puede ser una barrera para comprender su contenido (Saggion et al., 2015a) debido a la presencia de palabras poco frecuentes o desconocidas, y frases con complejidad léxica y semántica que complican drásticamente la comprensión del lector (North, Zampieri, and Shardlow, 2023b); esto es especialmente evidente en casos como niños, hablantes no nativos (Segura-Bedmar and Martinez, 2017) y personas con diversas capacidades cognitivas o discapacidades de lectura (Saggion et al., 2015b). El éxito o fracaso en la comprensión de un texto dependerá del conocimiento previo que tenga el lector del significado de las palabras (Anula, 2008).

La Ley de Zipf implica que las palabras que aparecen con menos frecuencia en un texto tienden a ser más largas y, por lo tanto, es probable que se consideren más complejas que las palabras que son más frecuentes y más cortas (Quijada and Medero, 2016). Además, las palabras

que tienen un gran número de sílabas tienden a ser difíciles de pronunciar, lo que puede complicar la experiencia de lectura (Mukherjee et al., 2016). Estos factores proporcionan métricas sólidas pero simples para comprender el nivel de complejidad asociado con una palabra (Desai et al., 2021a).

Predecir qué palabras presentan dificultades de comprensión para una población objetivo específica es un paso importante en muchas aplicaciones de Procesamiento del Lenguaje Natural (PNL). Esto es particularmente esencial en la simplificación de textos, un área que tradicionalmente ha dirigido su atención hacia estudiantes de una segunda lengua, hablantes nativos con bajos niveles de alfabetización y personas con dificultades lingüísticas en la lectura (Saggion et al., 2015b). La tarea de determinar qué palabras son más difíciles para el lector también se conoce como Identificación de Palabras Complejas (CWI - sus siglas en inglés).

La Identificación de las Palabras Complejas (CWI) y la Predicción de la Complejidad Léxica (LCP - sus siglas en inglés) pueden considerarse dos aspectos complementarios de una misma perspectiva. Mientras que CWI se centra en la identificación de palabras complejas mediante una tarea de clasificación, LCP aborda la predicción del grado de complejidad léxica de las palabras mediante una tarea de regresión. En esencia, CWI nos permite identificar palabras que podrían considerarse complejas y LCP nos proporciona una medida cuantitativa de su complejidad. La predicción precisa de la complejidad léxica puede permitir que los textos se adapten a las necesidades de los lectores (Shardlow, Cooper, and Zampieri, 2020b).

Rico-Sulayes, 2020 define CWI como la tarea de detectar en el contenido de documentos las palabras que resultan difíciles o complejas de entender por las personas de un determinado grupo. La aplicación de CWI y la posterior sustitución de aquellas palabras identificadas como difíciles puede generar mejoras significativas en la legibilidad y comprensión de un texto específico (Zotova et al., 2020).

Al ser la lengua española una de las lenguas más habladas del mundo (la cuarta en número de hablantes), suele contar con menos recursos terminológicos en comparación con otras lenguas como el inglés (Segura-Bedmar and Martínez, 2017) que ha sido objeto de un elevado número de investigaciones en el área de simplificación léxica, como lo demuestran las tareas compartidas de CWI en SemEval 2016 (Fernández et al., 2009), NAACL-HTL 2018 (Yimam et al., 2018b), la tarea ALexS en IberLEF 2020 (Ortiz-Zambrano and Montejo-Ráezb, 2020), la 15.^a edición de

SemEval y la primera tarea de predicción de la complejidad léxica (Shardlow et al., 2021b), taller TSAR-2022 sobre simplificación, accesibilidad y legibilidad del texto (Saggion et al., 2023), CLEF 2023 SimpleText Track (Ermakova et al., 2023), LREC-COLING 2024 - DeTermIt! Taller de Evaluación de la Dificultad del Texto en un Contexto Multilingüe (Nunzio et al., 2024), entre otras iniciativas en este sentido.

En resumen, el estudio de LCP es de gran interés por las razones indicadas. Hay mucho trabajo en marcha y campañas de evaluación, que requieren, sobre todo, corpus de referencia. En el caso del español, los corpus de referencia existentes son muy escasos o inexistentes, lo que dificulta el progreso. Además, la escasez de estos recursos puede afectar negativamente a la calidad de los análisis lingüísticos, lo que pone de relieve la urgente necesidad de iniciativas que promuevan la creación de corpus de lengua española para impulsar significativamente la investigación en este campo.

En términos de algoritmos y técnicas, hoy conviven soluciones basadas en ingeniería de características con algoritmos clásicos de aprendizaje automático, así como soluciones basadas en Deep Learning, principalmente a través de la clasificación a partir de representaciones codificadas de textos en incrustaciones de arquitecturas de codificación Transformer, y el prometedor uso de grandes modelos lingüísticos (generadores), que aún están poco explorados. La fusión de estas metodologías proporciona un amplio espectro de posibilidades que permitirían abordar de manera más completa y efectiva la complejidad de los temas asociados al LCP, promoviendo la sinergia entre técnicas tradicionales e innovadoras, abriendo así un horizonte prometedor para el desarrollo futuro en esta área de investigación.

Esta tesis proporciona un avance en el estudio del LCP para el español, creando nuevos conjuntos de datos y explorando nuevas soluciones para la predicción de la complejidad léxica, con aportaciones que han representado un nuevo estado del arte en algunos casos. Además, se ha realizado una validación experimental exhaustiva, destacando la robustez y generalización de las propuestas presentadas. Estos resultados no sólo fortalecen las investigaciones recientes sobre la complejidad léxica en español, sino que también abren el camino a nuevas perspectivas y desafíos para futuras investigaciones, marcando así un hito significativo en la evolución de esta disciplina en desarrollo.

6.2 Hipótesis de Investigación

El uso de redes neuronales artificiales, en particular arquitecturas de aprendizaje profundo, como los modelos lingüísticos basados en la arquitectura Transformers, ha demostrado ser altamente efectivo en la ejecución de tareas de Procesamiento del Lenguaje Natural, específicamente en tareas de regresión donde se combinan características de distinta naturaleza (Shardlow et al., 2021a). Partiendo de esta premisa, proponemos aplicar estos enfoques en el análisis de textos académicos en español, con el objetivo de abrir nuevas líneas de investigación para la identificación y evaluación automática de palabras complejas en dicha lengua dentro del ámbito académico, que permitan mejorar la comprensión de los desafíos lingüísticos a los que se enfrentan los usuarios al interactuar con textos especializados, facilitando así el desarrollo de herramientas que optimicen la accesibilidad y comprensión de contenidos académicos en español. De este modo, definimos nuestras hipótesis:

H1. *La implementación de grandes modelos lingüísticos complementados con características lingüísticas, surge como una solución beneficiosa para identificar y predecir palabras de alta complejidad en textos académicos en español.*

H2. *Los modelos generativos, como los modelos de lenguaje autorregresivos, pueden realizar la tarea de CWI.*

6.3 Justificación de la Investigación

La presencia de palabras desconocidas en un texto puede tener un impacto significativo en la capacidad del lector para comprenderlo. En algunos casos, el lector puede malinterpretar el texto, pasarlo por alto o incluso sentirse desanimado. Un lector dedicado puede optar por dedicar tiempo a buscar la definición de una palabra desconocida, enriqueciendo así su vocabulario mediante la comprensión y el aprendizaje de la palabra. Sin embargo, este enfoque conduce a una interrupción en el proceso de lectura, lo que a su vez puede afectar la concentración del lector (Shardlow et al., 2021a). Por lo tanto, las palabras complejas pueden imponer serias dificultades, considerando que su significado muchas veces depende de su contexto y no puede inferirse fácilmente (Zaharia, Cercel, and Dascalu, 2021).

La complejidad dentro de LCP se utiliza como “sinónimo de dificultad” (Malmasi and Zampieri, 2016). Muchos factores afectan la predicción de la complejidad léxica, uno de los cuales es el contexto en el que se encuentra una palabra, lo que afectará a su comprensibilidad. A veces, el significado de una palabra puede inferirse del contexto de las palabras conocidas que la rodean; en cambio, una palabra que está en el contexto de otras desconocidas es más difícil de entender. De manera similar, el género al que pertenece un texto puede afectar la comprensión del lector (Shardlow, Cooper, and Zampieri, 2020b). La complejidad de la palabra cambia ligeramente según la complejidad de la oración (Yaseen et al., 2021).

Según las investigaciones realizadas, actualmente no existen estudios realizados que investiguen el fenómeno de la complejidad léxica en el ámbito académico de la lengua española. Sumado a esto, hay poca investigación sobre herramientas de procesamiento del lenguaje natural (PNL) para apoyar a estudiantes y profesores de español, así como el desarrollo de aplicaciones efectivas de PNL dirigidas a la enseñanza, dado que los recursos para español que aún están disponibles no contienen anotaciones que faciliten el aporte de posibles soluciones (Davidson et al., 2020), lo que constituye un vacío importante, y una importante oportunidad de investigación para desarrollar nuevas herramientas y enfoques que apoyen a los estudiantes en la comprensión de textos académicos en español.

El aprendizaje profundo y sus tecnologías innovadoras representan nuevas tecnologías de vanguardia para diversas tareas de procesamiento del lenguaje natural (NLP) (Singh and Mahmood, 2021). Aquí, la predicción de la complejidad del vocabulario (LCP) no es una excepción (Nandy et al., 2021). Después de comparar y analizar los enfoques de aprendizaje profundo con otros enfoques, las posibles soluciones son factibles para el inglés y los idiomas de aprendizaje profundo con pocos recursos, donde los modelos profundos no siempre están disponibles o no son funcionales (Bender et al., 2021).

La falta de investigaciones sobre la identificación de palabras complejas en el contexto académico universitario en español representa una oportunidad importante para abordar esta necesidad mediante la aplicación de técnicas de PNL. El proyecto constituye una contribución significativa al campo de la educación superior, al centrarse en la solución de un problema específico que podría impactar significativamente la calidad de la educación y la igualdad de oportunidades para todos los estudiantes. La aportación de este proyecto también tiene un impacto en el ámbito social ya que facilitar la identificación de palabras complejas puede mejorar

la comprensión de textos académicos, lo que beneficia tanto a estudiantes como a académicos al abordar posibles barreras lingüísticas y mejorar el acceso a la información. Su contribución no se limita sólo a la población estudiantil específica, sino que también puede integrarse a otras instituciones de educación superior como un recurso valioso para apoyar el aprendizaje de los estudiantes universitarios.

Abordar la identificación de palabras complejas en el ámbito académico en español puede tener un impacto positivo en la equidad educativa, la accesibilidad a la información y el desarrollo de recursos que beneficien a los académicos de los estudiantes en el contexto hispanohablante y en la comunidad científica. La identificación de palabras complejas puede ser fundamental para diseñar intervenciones y recursos específicos dirigidos a estudiantes con dificultades lingüísticas, discapacidades, bajos niveles de comprensión lectora y problemas de lenguaje o construcción de vocabulario, brindándoles el apoyo necesario para superar los desafíos en el ámbito académico. Los resultados de la investigación podrían contribuir al desarrollo de recursos educativos especializados y estrategias pedagógicas que mejoren la enseñanza y el aprendizaje en contextos académicos. Mejorar la identificación de palabras complejas puede tener un impacto directo en la claridad y accesibilidad de la comunicación científica, facilitando la transmisión efectiva de conocimientos y descubrimientos en la comunidad académica.

6.4 Motivación Científica

La motivación científica que promueve esta tesis doctoral tiene su origen en la necesidad imperante de profundizar en la investigación sobre la complejidad léxica, específicamente en el ámbito académico de la lengua española. Este impulso se logrará mediante la exploración y aplicación de algoritmos de aprendizaje profundo, así como la ejecución de diversas tareas de procesamiento del lenguaje natural (PLN) enfocadas al análisis de palabras complejas. Esta elección se justifica por el potencial inherente de los grandes modelos de lenguaje basados en la arquitectura Transformer, que pueden entrenarse de manera eficiente utilizando extensos conjuntos de datos. Esta estrategia permitirá no sólo una investigación más profunda en el área de la complejidad léxica sino también la identificación de patrones lingüísticos específicos en el contexto académico, impulsando así el conocimiento y la aplicación práctica en este campo de investigación.

Los resultados de esta investigación pueden ayudar a la implementación de medidas y prácticas que hagan la información fácilmente comprensible para la población universitaria. Mejorar el aprendizaje durante el desarrollo de sus estudios académicos mediante la comprensión de textos académicos accesibles no solo beneficia a quienes tienen discapacidad, sino también a quienes tienen distintos niveles de habilidad lingüística, o bajos niveles de alfabetización; pero también contribuye a crear entornos educativos más inclusivos y equitativos para todos.

Como caso de uso, la aplicación de diversas técnicas de PNL nos brinda la capacidad de anticipar aquellas palabras complejas que muchas veces se convierten en obstáculos para la comprensión lectora de los estudiantes de pregrado de la Universidad de Guayaquil. La identificación temprana de palabras complejas, específicamente en los textos utilizados por los estudiantes de educación superior durante el proceso de aprendizaje, puede tener un impacto significativo en la mejora de la comprensión lectora y, por tanto, en el rendimiento académico de los estudiantes universitarios. Al realizar este estudio, no solo estamos explorando formas innovadoras de apoyar a los estudiantes, sino que también estamos contribuyendo al avance de la investigación en el campo de la educación y la PNL en un contexto académico específico, que es la Universidad de Guayaquil.

6.5 Objetivos de la Investigación

6.5.1 Objetivo General

Aplicar técnicas avanzadas de procesamiento del lenguaje natural para identificar y reducir la complejidad de palabras en textos académicos en español, con el fin de mejorar la accesibilidad y comprensión de dichos textos, y evaluar la eficacia de los algoritmos utilizados mediante un análisis exhaustivo de los resultados obtenidos.

6.5.2 Objetivos Específicos

1. Desarrollar nuevos corpus de referencia como recursos esenciales para el desarrollo de investigaciones sobre la complejidad lingüística en textos académicos en español.
2. Explore estrategias avanzadas para la identificación de palabras complejas y la predicción de la complejidad léxica en conjuntos de datos, como la aplicación de redes neuronales

artificiales, en particular arquitecturas de aprendizaje profundo, como modelos de lenguaje basados en Transformers, varias técnicas de procesamiento del lenguaje natural (PLN), algoritmos clásicos de aprendizaje automático y características lingüísticas calculadas a nivel de palabra. Posteriormente, evaluar el desempeño de los modelos analizando los resultados generados.

3. Integración de características lingüísticas en algoritmos clásicos de aprendizaje automático y modelos de redes neuronales profundas para la evaluación del desempeño en diferentes conjuntos de datos tanto en español como en inglés.
4. Explorar y aplicar técnicas no supervisadas, como modelos autorregresivos, para el análisis de la predicción de la complejidad léxica (LCP) para examinar las capacidades de los modelos generativos de lenguaje grande previamente entrenados para resolver esta tarea.

6.6 Metodología

Nos hemos basado en la línea de tiempo de Procesamiento del Lenguaje Natural propuesta por Kochmar, 2022 como marco de referencia para representar la secuencia de actividades sistemáticas y organizadas realizadas en nuestra investigación, garantizando la obtención de resultados confiables, así como la formulación de conclusiones válidas. La línea de tiempo proporciona información sobre el creciente nivel de sofisticación de los algoritmos de PNL, y en nuestro estudio hemos explorado las tres etapas mencionadas en la línea de tiempo. Para obtener una representación visual, Ver la Figura 6.1.

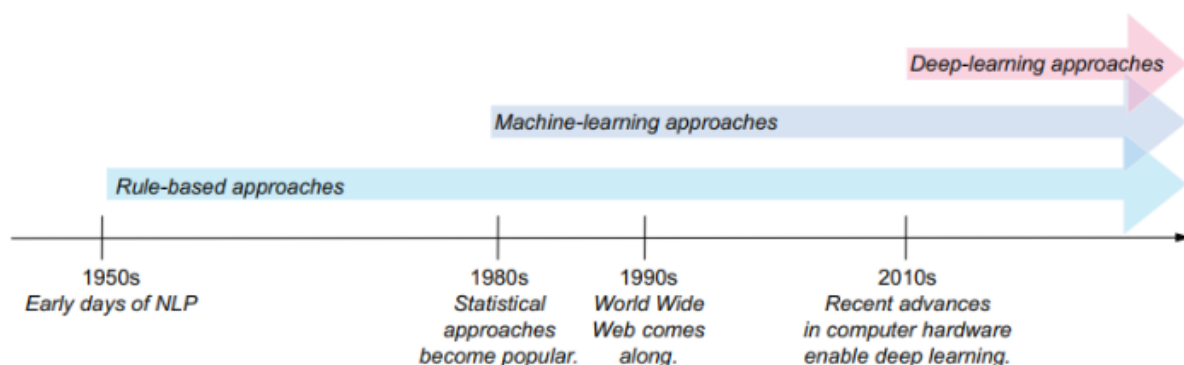


FIGURE 6.1: Cronología de la PNL que muestra tres tipos diferentes de enfoques de Kochmar, 2022.

A continuación, detallamos las principales etapas de nuestro proyecto:

1. *SOTA* - Estudio del estado del arte en la Identificación de Palabras Complejas.
2. *LF* - Identificación de características lingüísticas relevantes para mejorar el rendimiento de los algoritmos mediante el uso de las características lingüísticas extraídas, y medir el impacto de estas (de las características lingüísticas extraídas) en comparación con modelos basados en aprendizaje profundo (Gooding and Kochmar, 2018; Gooding and Kochmar, 2019), (North, Zampieri, and Shardlow, 2023a), (Aziz et al., 2023).
3. *Creación de los corpus* - Construcción de nuevos corpus y recursos.
4. *Enfoques basados en reglas* - El diseño de sistemas de evaluación se centró en evaluar la complejidad de las palabras según diversos criterios como la frecuencia de las palabras, el cual se basó en CREA¹ (Saggion et al., 2015a).
5. *Evaluación de diferentes modelos y combinaciones de características lingüísticas para CWI mediante la evaluación de algoritmos de aprendizaje automático y profundo* (Abramov, Ivanov, and Solovyev, 2023), (North et al., 2024).
 - (a) *Enfoques de Aprendizaje Automático* - Investigación y diseño de una solución y su evaluación mediante la ejecución de algoritmos clásicos de aprendizaje supervisado sobre conjuntos de datos en español e inglés.
 - (b) *Enfoques de Aprendizaje Profundo* - Investigación y diseño de una solución y su evaluación mediante la ejecución de los modelos basados en la arquitectura Transformer en los conjuntos de datos español e inglés.

6.7 Resultados

A continuación, se detallan los resultados obtenidos en cada etapa del desarrollo de la investigación:

¹CREA - Corpus de Referencia del Español Actual - Lista de frecuencias <https://corpus.rae.es/lfrecuencias.html>

6.7.1 SOTA - Estudio del Estado del Arte en Identificación de Palabras Complejas (CWI)

Esta parte de la investigación ha sido de vital importancia en la exploración y análisis de las fuentes documentales relevantes necesarias para sustentar las teorías y fórmulas que subyacen al estudio de la complejidad de las palabras en la lengua española y su aplicación desde diversos enfoques como el basado sobre frecuencia de palabras, aprendizaje automático y enfoques de aprendizaje profundo. Este estudio nos ha permitido reflexionar sobre el problema e identificar las fronteras desde donde exploramos nuestros enfoques para ir más allá y contribuir a un mayor conocimiento sobre la tarea del LCP.

6.7.2 Identificación de Características Lingüísticas Relevantes

Según la revisión de investigaciones sobre el tema de la complejidad léxica, hemos comprobado que tras las ejecuciones de las diferentes tecnologías aplicadas, en las que los rasgos lingüísticos han formado parte del conjunto de datos, estos juegan un papel fundamental en la predicción de palabras complejas; por lo tanto, han sido parte del diseño y compilación de conjuntos de datos en español como en CLeXIS² (Ortiz-Zambrano and Montejo-Ráez, 2021), LegalEc (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023a), y también en conjuntos de datos en inglés para el desarrollo de propuestas en Predicción de Complejidad Léxica como tarea 1 del 15° Taller Internacional sobre Evaluación Semántica - SemEval-2021² donde nuestra solución contenía un total de 15 características lingüísticas agregadas al conjunto de datos en inglés proporcionado a los participantes como parte del taller para el desarrollo de la tarea (Zambrano and Montejo-Ráez, 2021).

6.7.3 Creación de Corpus - La Creación de Recursos

Estos recursos constituyen la base fundamental para llevar a cabo tareas en nuestro estudio de las palabras complejas en diversas ramas de los estudios universitarios en español, dado que los recursos para el español en general y para la educación en particular son escasos. La falta de recursos como corpus paralelos y recursos léxicos representa una dificultad para el avance de propuestas en el área de Simplificación Léxica, convirtiéndose en un desafío evidente en el caso del idioma español (Bott and Saggion, 2012).

²SemEval-2021. El 15° Taller Internacional sobre Evaluación Semántica. Disponible en <https://semeval.github.io/SemEval2021/tasks>

Los corpus creados en el transcurso de esta investigación son:

- VYTEDU: Un corpus de videos y sus transcripciones para la investigación educativa (Ortiz-Zambrano and Montejo-Ráez, 2017).
- VYTEDU-CW: Las palabras difíciles como barrera en la comprensión lectora de estudiantes universitarios (Ortiz-Zambrano et al., 2019).
- CLexIS²: Un nuevo corpus para la investigación de identificación de palabras complejas en estudios de informática (Ortiz-Zambrano and Montejo-Ráez, 2021).
- LegalEc: Un nuevo corpus para la investigación de la identificación de palabras complejas en estudios de Derecho en español ecuatoriano (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023a).
- GovAIEc: En proceso de revisión.

6.7.4 Línea de Base con CREA

Comenzamos a explorar las soluciones más simples. Este enfoque se basa en la frecuencia de uso de una palabra según una investigación realizada por Saggion et al., 2015a. Un recurso general como CREA sirve de base para identificar si la palabra es compleja o no. Construimos un sistema que permitió la detección de palabras complejas contenidas en el corpus CLexIS2 (Ortiz-Zambrano and Montejo-Ráez, 2021) mediante el cálculo de palabras de baja frecuencia, siendo aquellas palabras cuya frecuencia es menor a 1,000 (Rebollo, 2008).

6.7.5 Enfoques de Aprendizaje Automático

Diseñamos y evaluamos soluciones basadas en algoritmos clásicos de aprendizaje automático, realizando un análisis de diferentes características lingüísticas y su combinación con incrustaciones de modelos previamente entrenados (no afinados), como algunos modelos basados en BERT (Zambrano and Montejo-Ráez, 2021; Ortiz-Zambrano and Montejo-Ráez, 2021; Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2022b; Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2022a). En estos trabajos estudiamos estos enfoques tanto para inglés como para español.

6.7.6 Enfoques de Aprendizaje Profundo: la implementación de modelos basados en la arquitectura Transformer

En esta etapa, llevamos a cabo la exploración y aplicación de modelos basados en la arquitectura Transformer, incluyendo BERT, RoBERTa (“Deep Encodings vs. Linguistic Features in Lexical Complexity Prediction” - artículo enviado a la revista Neural Computing and Applications de Springer Publishing³, que se encuentra en etapa de revisión desde noviembre de 2023), y el modelo generador de lenguaje humano GPT-3 (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023b) y nuestra publicación reciente aplicando GPT-4-turbo (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024b). Los resultados obtenidos demostraron un papel importante en el análisis y evaluación del desempeño de estos modelos sobre conjuntos de datos en inglés y español, el cual se realizó con el propósito de contrastarlos con un enfoque particular en español, mismo que constituye el objeto. de nuestro estudio.

Nuestra hipótesis H1. *La implementación de grandes modelos de lenguaje complementados con características lingüísticas, surge como una solución beneficiosa para identificar y predecir palabras de alta complejidad en textos académicos en español* (Paetzold, 2021). Destacamos un hallazgo significativo: *la incorporación de características lingüísticas en conjunto con representaciones vectoriales de palabras, y el aumento en el número de épocas durante la ejecución de grandes modelos de lenguaje, conduce a una mejora sustancial en el desempeño de los algoritmos para predecir la complejidad léxica, como se evidencia en nuestros estudios* (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2022a), (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023a). Se muestra gráficamente en la Figura 6.2.

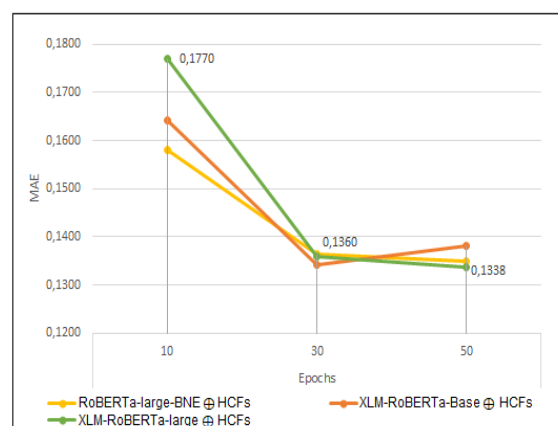


FIGURE 6.2: Integración de HCF e incremento de épocas.

³Computación neuronal y aplicaciones - Disponible en <https://link.springer.com/journal/521>

Otro resultado destacado que confirma nuestra hipótesis H2. *Los modelos generativos, cómo los modelos de lenguaje autorregresivos, pueden realizar la tarea de CWI*, es la solución presentada en la tarea propuesta “Tarea 2: ¿Qué no está claro? Identificación y explicación de conceptos difíciles (definiciones, descifrado de abreviaturas, contexto, aplicaciones,...)” por el taller internacional SimpleText@CLEF-2023 y SimpleText@CLEF-2024. Nuestros resultados en la tarea 2.1 en SimpleText@CLEF-2023 mostraron un desempeño significativo del enfoque propuesto cerca de los sistemas de mejor desempeño. Es interesante observar que, en general, las indicaciones de aprendizaje de cero disparos son mejores para detectar una palabra difícil, mientras que las indicaciones de aprendizaje de pocos disparos son mejores para determinar una puntuación de dificultad. El uso del lenguaje generativo GPT-3 mediante la aplicación de la técnica de aprendizaje Zero-Shot generó resultados extremadamente beneficiosos para identificar y predecir palabras al otorgarnos el primer lugar entre los participantes (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023b).

En nuestra propuesta en las subtareas 2.1 y 2.2 dentro de SimpleText@CLEF 2024 (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024), el modelo ha mostrado un rendimiento robusto en el área del procesamiento del lenguaje natural, particularmente en la identificación de palabras complejas. Aparte de realizar estas funciones de manera eficaz, también ha producido ilustraciones detalladas mediante la elaboración de definiciones, explicaciones y ejemplos contextualizados. Al identificar palabras simples y multipalabras complejas y ofrecer opciones más simples, facilita la comprensión de los textos científicos, beneficiando a diversos grupos de lectores, incluidas aquellas personas con dificultades de lectura o discapacidades cognitivas. Nuestra hipótesis se valida al evidenciar que los modelos generativos, como los modelos de lenguaje autorregresivos, son capaces de realizar la tarea de identificación de palabras complejas (CWI) gracias a su capacidad para generar texto y modelar la complejidad de las palabras en función del contexto.

6.8 Conclusiones

Presentamos las principales conclusiones derivadas de los resultados de la investigación sobre la identificación de palabras complejas en el ámbito académico español y presentadas en el Simposio Doctoral sobre Procesamiento del Lenguaje Natural del Proyecto ILENIA (PLN-DS-2023) celebrado en el marco del XXXIX edición del Congreso Internacional de la Sociedad

Española de Procesamiento del Lenguaje Natural (SEPLN 2023) (Ortiz-Zambrano, [2023](#)).

Los resultados obtenidos en este estudio apoyan firmemente nuestras hipótesis siendo H1. *La implementación de modelos lingüísticos a gran escala complementados con características lingüísticas es una solución beneficiosa para identificar y predecir palabras altamente complejas en textos académicos en español.* A través de una serie de experimentos y análisis detallados, demostramos que la aplicación de modelos lingüísticos y características lingüísticas mejora significativamente el rendimiento en tareas complejas de identificación y predicción de palabras. A partir del análisis de los resultados formulamos conclusiones sobre la validez de las hipótesis presentamos y analizamos las conclusiones obtenidas en relación con los objetivos específicos planteados destacando cómo los hallazgos de esta investigación responden a cada uno de ellos.

Primer Objetivo Específico: Desarrollar nuevos corpus de referencia como recursos esenciales para el desarrollo de investigaciones sobre la complejidad lingüística en textos académicos en español.

Hemos logrado cumplir con el primer objetivo específico de desarrollar nuevos corpus de referencia como recursos fundamentales para el estudio de la complejidad lingüística en textos académicos en español, para lo cual se llevaron a cabo varias actividades esenciales. En primer lugar, se recopiló una amplia variedad de textos académicos de diversas disciplinas, incluyendo artículos científicos, tesis y libros de texto, artículos de la constitución ecuatoriana, trabajos de fin de carrera, textos públicos, entre otros. Estos textos fueron cuidadosamente anotados para identificar y clasificar palabras complejas, tarea que fue realizada por estudiantes y personas en general. Además, se desarrollaron herramientas de anotación automática utilizando algoritmos de procesamiento de lenguaje natural y aprendizaje automático, entrenados con los datos anotados manualmente para mejorar su precisión y eficiencia.

Los nuevos corpus creados como recursos propios constituyeron un componente esencial en la investigación de palabras complejas, convirtiéndose en una base sólida como punto de partida, así como en los datos empíricos necesarios para identificar, analizar y comprender la complejidad de textos en lengua española en un contexto específico como es el ámbito académico.

Estos recursos fueron presentados en varios congresos científicos internacionales sobre evaluación semántica y avances recientes en el campo del Procesamiento del Lenguaje Natural (PNL).

Destacamos el corpus *VYTEDU*: Un corpus de videos y sus transcripciones para la investigación educativa (Ortiz-Zambrano and Montejo-Ráez, 2017). *VYTEDU* marcó el punto de partida de esta investigación. Para su evaluación y validación se aplicaron métricas de complejidad léxica para la lectura en español (Ortiz-Zambrano and Varela Tapia, 2018). Fue presentado en el XXXIII Congreso Internacional de la Sociedad Española de Procesamiento del Lenguaje Natural (SEPLN 2017).

El corpus *VYTEDU-CW* (Ortiz-Zambrano et al., 2019) fue desarrollado a partir del proceso de etiquetado del corpus *VYTEDU*, con el objetivo de prepararlo para su evaluación mediante su aplicación con modelos de lenguaje basados en transformadores, ya que estos modelos han demostrado ser altamente efectivo para identificar y predecir la complejidad léxica. El análisis de las palabras etiquetadas como complejas permitió una caracterización detallada de estas palabras en el corpus *VYTEDU-CW* (Ortiz-Zambrano and Montejo-Ráez, 2020). Este corpus se consolidó como un recurso esencial para nuevas investigaciones en el campo de la Simplificación Léxica, siendo parte integral de nuestro taller ALexS 2020 - Análisis Léxico en SEPLN 2020 (Ortiz-Zambrano and Montejo-Ráez, 2020) en una participación activa como parte del Foro Ibérico para la Evaluación de Idiomas - IberLEF 2020.

El corpus *CLexIS²*: Una nueva colección para la investigación sobre la identificación de palabras complejas en estudios de informática (Ortiz-Zambrano and Montejo-Ráez, 2021), está compuesto por textos de Ciencias Computacionales. Este corpus fue desarrollado con el propósito de implementar nuevas evaluaciones, que vayan más allá de las convencionales, incluyendo algoritmos clásicos de aprendizaje automático. *CLexIS²* se presentó en la conferencia Avances Recientes en el Procesamiento del Lenguaje Natural - RANLP 2021, destacando su contribución innovadora al campo del procesamiento del lenguaje natural y la simplificación léxica.

El corpus *LegalEc*: Un recurso innovador para la investigación en la identificación de palabras complejas en estudios jurídicos del español ecuatoriano (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023a), ha sido creado con el propósito de avanzar en la exploración y comprensión de la complejidad léxica en el campo de la ley. Presentado en el prestigioso XXXIX Congreso Internacional de la Sociedad Española de Procesamiento del Lenguaje Natural - SEPLN 2023, *LegalEc* constituye una valiosa contribución al estudio del lenguaje jurídico, facilitando el desarrollo de herramientas y métodos para mejorar la accesibilidad y comprensión

de los textos jurídicos en español. Ecuatoriano. Su proceso de evaluación se realizó aplicando modelos basados en la arquitectura Transformadors.

El corpus *GovAIEc* se encuentra en proceso de revisión.

Los corpus resultantes fueron validados y evaluados utilizando métodos rigurosos, asegurando su confiabilidad y consistencia. La aplicación de la complejidad léxica y métricas de error común nos permitió evaluar exhaustivamente la complejidad del contenido de los textos de los distintos corpus. Este análisis nos reveló que el léxico utilizado por los docentes al impartir sus clases incluye oraciones complejas, además de un lenguaje técnico y sofisticado, lo que dificulta su comprensión por parte de los estudiantes. Asimismo, observamos que los textos utilizados por los estudiantes para el aprendizaje contienen elementos similares que complican su comprensión. Estos hallazgos son sólo algunos ejemplos notables de nuestro estudio.

Segundo Objetivo Específico: Explorar estrategias avanzadas para la identificación de palabras complejas (CWI) y la predicción de la complejidad léxica (LCP) en conjuntos de datos, como la aplicación de redes neuronales artificiales, en particular arquitecturas de aprendizaje profundo como los modelos de lenguaje basados en Transformers, varias técnicas de procesamiento del lenguaje natural (PNL), algoritmos clásicos de aprendizaje automático y características lingüísticas calculadas a nivel de palabra. Posteriormente, evaluar el desempeño de los modelos analizando los resultados generados.

Este objetivo específico ha sido abordado ampliamente a lo largo de la investigación, explorando diversas estrategias y tecnologías avanzadas para la identificación de palabras complejas y la predicción de la complejidad léxica en textos académicos en español. Se ha puesto especial énfasis en el uso de redes neuronales artificiales, particularmente arquitecturas basadas en Transformer, reconocidas por su capacidad para capturar y procesar información contextual compleja de manera efectiva. Estos modelos se han mostrado prometedores en la tarea de CWI y LCP, permitiendo una identificación más precisa y detallada de términos lingüísticamente complejos.

No solo se han explorado modelos de lenguaje basados en Transformer, sino que también se han aplicado varias técnicas de procesamiento del lenguaje natural (NLP), incluidos algoritmos clásicos de aprendizaje automático y características lingüísticas a nivel de palabras (Zambrano

and Montejo-Ráez, 2021), (Ortiz-Zambrano and Montejo-Ráez, 2021), una combinación de estas técnicas (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2022b), (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2022a), y el enfoque de extremo a extremo (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023a), (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024a), que ha enriquecido el análisis y predicción de la complejidad léxica, proporcionando así importantes resultados sobre la dificultad de los textos académicos en español.

La evaluación del desempeño de los modelos ha sido un proceso muy riguroso, analizando los resultados generados en términos de precisión, recuperación y otras métricas relevantes con las de error común. Estos análisis nos permitieron identificar fortalezas y áreas de mejora en los modelos propuestos, brindando recomendaciones para futuras investigaciones y aplicaciones prácticas en el ámbito educativo y académico. La exploración de estrategias avanzadas ha contribuido significativamente al avance del conocimiento en CWI y LCP, destacando la importancia de la innovación tecnológica y la aplicación de métodos sofisticados en lingüística computacional.

Tercer Objetivo Específico: Integrar las características lingüísticas en algoritmos clásicos de aprendizaje automático y modelos de redes neuronales profundas para la evaluación del desempeño sobre diferentes conjuntos de datos tanto en español como en inglés.

Este objetivo específico se ha abordado mediante un enfoque sistemático que integra características lingüísticas en algoritmos clásicos de aprendizaje automático, así como en modelos de redes neuronales profundas.

Es pertinente enfatizar que las características lingüísticas tienen un papel muy relevante en la predicción de la complejidad léxica. No sólo son esenciales para optimizar la precisión de los modelos predictivos, sino que también gozan de gran estima y demanda entre numerosos investigadores, que buscan incorporarlos a sus conjuntos de datos. La importancia de estos rasgos en el campo de la simplificación léxica pone de relieve la necesidad de considerarlos en cualquier análisis lingüístico exhaustivo.

Comenzamos identificando y calculando varias características lingüísticas relevantes a nivel de palabra, como la longitud de la palabra, la frecuencia de uso, la complejidad morfológica y las características sintácticas. Estas características fueron elegidas en función de su potencial

para capturar aspectos relevantes de la complejidad léxica en textos académicos en español e inglés. Las características lingüísticas calculadas se ejecutaron en varios algoritmos clásicos de aprendizaje automático, incluidas máquinas de vectores de soporte (SVM), árboles de decisión y bosques aleatorios, entre otros. Estos algoritmos fueron entrenados y evaluados utilizando conjuntos de datos en español (Ortiz-Zambrano and Montejo-Ráez, 2021), (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2022b), (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024a) y en inglés (Zambrano and Montejo-Ráez, 2021), (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2022a), (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024a). La integración de características lingüísticas mejoró significativamente el rendimiento de estos modelos a la hora de identificar palabras complejas y predecir su complejidad léxica.

Además, ejecutamos los modelos BERT, XLM-RoBERTa, XLM-RoBERTa-base y RoBERTa-large-BNE en los conjuntos de datos en inglés y español utilizando el enfoque de un extremo a otro. Los mejores resultados los demostró el modelo BERT en combinación con características lingüísticas para los conjuntos de datos (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024a) en inglés y español. Es importante señalar que el rendimiento de los modelos depende del número de épocas (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023a). Sin embargo, los modelos de aprendizaje profundo funcionan como una caja negra, por lo que comprender cómo las características lingüísticas complementan las características profundas requiere trabajar en la explicabilidad del modelo profundo en sí, ya que los Transformers pueden codificar información relacionada con la sintaxis, las dependencias, la gramática, el género, la negación y la semántica, entre otras. otros, dentro de sus capas.

El uso de características lingüísticas y modelos basados en Transformer que aplican el enfoque de extremo a extremo pareció mejorar consistentemente los resultados. La ganancia de rendimiento parece estar relacionada con el lenguaje, lo que podría estar relacionado con una mayor calidad de desempeño en los modelos Transformer (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024a) inglés y español o con un problema de dificultad de la tarea dependiendo del idioma. Estos hallazgos subrayan la relevancia y efectividad de combinar técnicas de codificación avanzadas con características lingüísticas tradicionales para mejorar la precisión de los modelos predictivos en contextos multilingües y multifacéticos.

Vale la pena enfatizar la relevancia de los modelos BERT y RoBERTa, que han demostrado ser un tipo esencial de arquitectura de redes neuronales en esta área. Su implementación en

nuestra investigación con conjuntos de datos españoles ha sido particularmente valiosa. Los resultados de la evaluación final revelan que los modelos de lengua española contribuyeron significativamente a la predicción de la complejidad léxica, superando a las propuestas iniciales que sólo utilizaban rasgos lingüísticos. La combinación de estas características con las codificaciones generadas por los modelos, aplicadas en algoritmos clásicos de aprendizaje automático, resultó en una mejora sustancial en el rendimiento de (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2022b).

Es importante enfatizar que de acuerdo a los resultados, al ejecutar diferentes combinaciones de características, con un MAE = 0,1323 en el corpus español CLexIS², el mejor rendimiento se obtuvo utilizando la versión afinada del modelo BERT español. Para enfoques no afinados, el algoritmo AdaBoost con codificaciones de oraciones y palabras XLM-RoBERTa junto con características lingüísticas reportó un alto rendimiento con un MAE = 0.1623 (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024a). Para el corpus CompLex en inglés, con un MAE de 0,0683, el ajuste fino de BERT fue nuevamente la mejor opción, pero solo si se introducen características lingüísticas en el modelo y SVM fue el segundo mejor algoritmo considerando el modelo previamente entrenado y tomando la palabra objetivo. incrustaciones de BERT con las características (lingüísticas) hechas a mano propuestas (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024a).

Cuarto Objetivo Específico: Explorar y emplear técnicas no supervisadas, como modelos autorregresivos, para el análisis de la predicción de la complejidad léxica (LCP) para examinar las capacidades de los modelos generativos de lenguaje grande previamente entrenados para resolver esta tarea.

Hemos confirmado nuestra hipótesis **H2**. *Los modelos generativos, como los modelos de lenguaje autorregresivos, pueden cumplir la tarea de CWI*, ya que los resultados de la exploración y aplicación de técnicas no supervisadas, en particular modelos autorregresivos, como GPT-3, GPT-3.5, GPT-4-turbo y GPT-4o para el análisis de la predicción de complejidad léxica (LCP) lograron demostrar un alto rendimiento en el desempeño del modelo.

El modelo autorregresivo GPT-3 y sus versiones text-davinci-002, text-davinci-003 demostraron su desempeño en conjuntos de datos en inglés durante el análisis. Se utilizó un enfoque no supervisado, que permitió al modelo aprender las complejidades del lenguaje a

partir de unos pocos ejemplos, donde el modelo pudo interpretar y generar sus respuestas basándose en unos pocos ejemplos e instrucciones complejas, lo que demuestra que la versión basada en texto Davinci-003 proporciona mejores resultados que text-Davinci-002 (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024b). Los resultados obtenidos por el modelo demostraron un desempeño sólido. Estos resultados fueron significativos para superar los desafíos asociados a los métodos de simplificación de textos, con el objetivo de facilitar el acceso a la información científica (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023b). Las métricas de evaluación aplicadas, como BLEU, precisión, recuperación, medición y error común, se aplicaron para evaluar la efectividad de los modelos autorregresivos utilizados. Los resultados indicaron que los modelos generativos como GPT-3 y sus versiones posteriores son capaces de capturar matices contextuales que son críticos para una predicción precisa de la complejidad léxica.

También aplicado el modelo GPT-4-Turbo en la construcción de varias soluciones para las subtareas 2.1 y 2.2 en SimpleText@CLEF 2024 (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2024), donde el modelo ha demostrado un rendimiento robusto en la ejecución de tareas de procesamiento del lenguaje natural, específicamente en la simplificación léxica y la identificación de palabras complejas. No solo ha abordado con éxito estas tareas, sino que también ha sido capaz de generar ilustraciones detalladas creando definiciones, explicaciones y casos de uso basados en el contexto de la oración (texto). Al identificar palabras y frases complejas y ofrecer versiones más sencillas, facilita el acceso a fuentes científicas de forma comprensible, ayudando a los lectores a comprender la complejidad de los términos en diferentes grupos de usuarios, lo que lo hace muy útil para un público general y especialmente valioso para lectores con dificultades de lectura o discapacidades cognitivas. Confirmamos nuestra hipótesis de que los modelos generativos, como los modelos de lenguaje autorregresivos, pueden realizar la tarea de identificación de palabras complejas (CWI) gracias a su capacidad para generar secuencias de texto y modelar la complejidad de las palabras en función del contexto.

Como producto del estudio realizado se desarrolló una aplicación innovadora que convierte textos complejos en versiones más accesibles, puede visitar el sitio web en el siguiente enlace <https://govaieasy.com>. Esta herramienta está diseñada para beneficiar a estudiantes universitarios y público en general, con especial enfoque en personas con discapacidad o dificultades de comprensión lectora. Al mejorar significativamente la accesibilidad y comprensión de los textos, la aplicación contribuye a la inclusión y el éxito académico de una amplia diversidad de usuarios.

6.9 Limitations

Los modelos de lenguaje basados en BERT y RoBERTa Transformer se aplicaron a un conjunto de características de un corpus de entrenamiento de palabras simples en inglés. Nuestra solución no está orientada a varias palabras. Vemos conveniente aplicar otros modelos de lenguaje que nos permitan valorar los resultados como una posible mejora en el nivel de predicción de palabras. El valor de R^2 obtenido en nuestros experimentos nos dice que sólo el 19% de la variación de la complejidad puede explicarse a partir de nuestras características, lo cual es mínimo. Esto sugiere que se debe realizar más experimentación con características adicionales y pruebas de ablación para identificar qué tipos de características (incrustaciones, relacionadas con la complejidad, relacionadas estilísticamente, etc.) contribuyen más a la identificación de palabras difíciles.

Las experimentaciones realizadas y los sistemas propuestos para la simplificación léxica pueden ayudar a otros investigadores en su trabajo a explorar dicho diseño. Algunos modelos de lenguaje y la mayoría de los algoritmos clásicos no exigen grandes recursos informáticos y podrían ser una solución en entornos de bajos recursos, como dispositivos móviles y similares.

Bibliography

- Abramov, Aleksei V, Vladimir V Ivanov, and Valery D Solovyev (2023). "Lexical Complexity Evaluation based on Context for Russian Language". In: *Computación y Sistemas* 27.1, pp. 127–139.
- Abreu, José Luis (2014). "El método de la investigación Research Method". In: *Daena: International journal of good conscience* 9.3, pp. 195–204.
- Adamssen, John (2020). *Inteligencia artificial: Aprender sobre chatbots, robótica y otras aplicaciones comerciales*. Efalón Acies.
- Agrawal, Sweta and Marine Carpuat (2024). "Do Text Simplification Systems Preserve Meaning? A Human Evaluation via Reading Comprehension". In: *Transactions of the Association for Computational Linguistics* 12, pp. 432–448.
- Ahmad, Iftikhar et al. (2020). "Fake news detection using machine learning ensemble methods". In: *Complexity* 2020, pp. 1–11.
- Al-Ajlan, Amani A, Hend S Al-Khalifa, and AbdulMalik S Al-Salman (2008). "Towards the development of an automatic readability measurements for Arabic language". In: *2008 Third international conference on digital information management*. IEEE, pp. 506–511.
- Alammary, Ali Saleh (2022). "BERT Models for Arabic Text Classification: A Systematic Review". In: *Applied Sciences* 12.11. ISSN: 2076-3417. DOI: [10.3390/app12115720](https://doi.org/10.3390/app12115720). URL: <https://www.mdpi.com/2076-3417/12/11/5720>.
- Alarcón, Rodrigo, Lourdes Moreno, and Paloma Martínez (2020). "Hulat-AlexS CWI Task-CWI for Language and Learning Disabilities Applied to University Educational Texts". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR-WS, Malaga, Spain*. Malaga, Spain: CEUR-WS.
- (2021). "Exploration of Spanish Word Embeddings for Lexical Simplification." In: *CTTS@SEPLN*.
- Alarcon, Rodrigo et al. (2019). "Lexical simplification approach using easy-to-read resources". In: *Procesamiento del Lenguaje Natural* 63, pp. 95–102.

- Ali, Md Shahin et al. (2023). "A novel approach for best parameters selection and feature engineering to analyze and detect diabetes: Machine learning insights". In: *BioMed Research International* 2023.
- Ali, S. M. et al. (2024). "Improving Scientific Text Comprehension: A Multi-Task Approach with GPT-3.5 Turbo and Neural Ranking". In: *Working Notes of CLEF 2024: Conference and Labs of the Evaluation Forum*. Ed. by G. Faggioli et al. CEUR Workshop Proceedings. CEUR-WS.org, [24].
- Alsalman, Alanoud (2018). "Machine learning algorithms and natural language processing techniques for crime prediction with geo-tagged tweets". PhD thesis. University of Colorado.
- Amade, Divya et al. (2024). "Automatic Text Summarization Using NLTK & Spacy". In: *Available at SSRN* 4742012.
- Amiel Pérez, José (2007). "Las variables en el método científico". In: *Revista de la Sociedad Química del Perú* 73.3, pp. 171–177.
- Anggia, Helta and Anita Habók (2023). "Textual complexity adjustments to the English reading comprehension test for undergraduate EFL students". In: *Heliyon* 9.1.
- Anula, Alberto (2008). "Lecturas adaptadas a la enseñanza del español como L2: variables lingüísticas para la determinación del nivel de legibilidad". In: *La evaluación en el aprendizaje y la enseñanza del español como LE L 2*, pp. 162–170.
- Aranzabe, Jesús, Arantza Díaz de Ilarraza, and Itziar Gonzalez-Dios (2012). "First Approach to Automatic Text Simplification in Basque Maria". In.
- Arias, Fidas G (2012). *El proyecto de investigación. Introducción a la metodología científica. 6ta.* Fidas G. Arias Odón.
- Armadas, ESPE (2018). *Introducción a la metodología de la investigación científica.*
- Aziz, Abdul et al. (2023). "Leveraging contextual representations with BiLSTM-based regressor for lexical complexity prediction". In: *Natural Language Processing Journal* 5, p. 100039. ISSN: 2949-7191. DOI: <https://doi.org/10.1016/j.nlp.2023.100039>. URL: <https://www.sciencedirect.com/science/article/pii/S2949719123000365>.
- Bakker, J., G. Yüksel, and J. Kamps (2024). "University of Amsterdam at the CLEF 2024 SimpleText Track". In: *Working Notes of CLEF 2024: Conference and Labs of the Evaluation Forum*. Ed. by G. Faggioli et al. CEUR Workshop Proceedings. CEUR-WS.org.
- Bender, Emily M. et al. (2021). "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. Virtual Event, Canada: Association for Computing Machinery,

- 610–623. ISBN: 9781450383097. DOI: [10.1145/3442188.3445922](https://doi.org/10.1145/3442188.3445922). URL: <https://doi.org/10.1145/3442188.3445922>.
- Bhattacharjee, Amrita et al. (2024). *Zero-shot LLM-guided Counterfactual Generation for Text*. arXiv: 2405.04793 [cs.CL].
- Bott, Stefan and Horacio Saggion (2012). “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, pp. 527–534.
- Bott, Stefan et al. (2012). “Can spanish be simpler? lexis: Lexical simplification for spanish”. In: *Proceedings of COLING 2012*, pp. 357–374.
- Bottou, Léon (2010). “Large-scale machine learning with stochastic gradient descent”. In: *Proceedings of COMPSTAT’2010*. Springer, pp. 177–186.
- Briskilal, J and CN Subalalitha (2022). “An ensemble model for classifying idioms and literal texts using BERT and RoBERTa”. In: *Information Processing & Management* 59.1, p. 102756.
- Bronte-Stewart, Helen et al. (2009). “The STN beta-band profile in Parkinson’s disease is stationary and shows prolonged attenuation after deep brain stimulation”. In: *Experimental neurology* 215.1, pp. 20–28.
- Brown, Tom et al. (2020). “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33, pp. 1877–1901.
- Burešová, Karolína (2017). “Text simplification in Czech”. In: .
- Camposa, Randall Araya et al. (2020). “Estudio de la complejidad del Español para la simplificación textual”. In: *Revista Tecnología en Marcha*, ág–45.
- Canete, José et al. (2020). “Spanish pre-trained bert model and evaluation data”. In: *Pml4dc at iclr 2020*, p. 2020.
- Carletta, Jean (2007). “Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus”. In: *Language Resources and Evaluation* 41, pp. 181–190.
- Carroll, John et al. (1998). “Practical simplification of English newspaper text to assist aphasic readers”. In: *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*. Association for the Advancement of Artificial Intelligence, pp. 7–10.
- Chan, Anastasia (2023). “GPT-3 and InstructGPT: technological dystopianism, utopianism, and “Contextual” perspectives in AI ethics and industry”. In: *AI and Ethics* 3.1, pp. 53–64.
- Chandrasekar, Raman, Christine Doran, and Srinivas Bangalore (1996). “Motivations and methods for text simplification”. In: *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

- Chang, Yupeng et al. (2023). *A Survey on Evaluation of Large Language Models*. arXiv: [2307.03109 \[cs.CL\]](#).
- Chen, Ping et al. (2017). "Automatic text simplification for people with intellectual disabilities". In: *Artificial Intelligence Science and Technology: Proceedings of the 2016 International Conference (AIST2016)*. World Scientific, pp. 725–731.
- Conneau, Alexis et al. (July 2020). "Unsupervised Cross-lingual Representation Learning at Scale". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 8440–8451. DOI: [10.18653/v1/2020.acl-main.747](#). URL: <https://aclanthology.org/2020.acl-main.747>.
- Crammer, Koby et al. (2006). "Online Passive-Aggressive Algorithms". In: *Journal of Machine Learning Research* 7, pp. 551–585.
- Crossley, Scott et al. (2023). "A large-scaled corpus for assessing text readability". In: *Behavior Research Methods* 55.2, pp. 491–507.
- Crossley, Scott A., Tom Salsbury, and Danielle S. McNamara (2012). "Predicting the proficiency level of language learners using lexical indices". In: *Language Testing* 29.2, pp. 243–263.
- Davidson, Sam et al. (2020). "Developing NLP tools with a new corpus of learner Spanish". In: *Proceedings of the 12th language resources and evaluation conference*, pp. 7238–7243.
- De Belder, Jan and Marie-Francine Moens (2010). "Text simplification for children". In: *Proceedings of the SIGIR workshop on accessible search systems*. ACM; New York, pp. 19–26.
- De Jesús Torres, Jessika Anabel (2017). *Diseño de un software educativo para el aprendizaje de lengua y literatura en los signos de puntuación de los y las estudiantes de primer año de bachillerato general unificado sección matutina, paralelo F del instituto tecnológico superior central técnico D.M. Quito, periodo 2016*. Trabajo de investigación. Disponible en: <http://www.dspace.uce.edu.ec/handle/25000/11181>.
- Dean, Jared (2014). *Big data, data mining, and machine learning: value creation for business leaders and practitioners*. John Wiley & Sons.
- Degraeuwe, Jasper and Patrick Goethals (May 2024). "LexComSpaL2: A Lexical Complexity Corpus for Spanish as a Foreign Language". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari et al. Torino, Italia: ELRA and ICCL, pp. 10432–10447. URL: <https://aclanthology.org/2024.lrec-main.912>.

- Dell'Orletta, Felice, Simonetta Montemagni, and Giulia Venturi (2011). "READ-IT: Assessing readability of Italian texts with a view to text simplification". In: *Proceedings of the second workshop on speech and language processing for assistive technologies*, pp. 73–83.
- Desai, Abhinandan et al. (2021a). "LCP-RIT at SemEval-2021 Task 1: Exploring Linguistic Features for Lexical Complexity Prediction". In: *arXiv preprint arXiv:2105.08780*.
- Desai, Abhinandan Tejalkumar et al. (Aug. 2021b). "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)*. Online: Association for Computational Linguistics, pp. 548–553. DOI: [10.18653/v1/2021.semeval-1.67](https://doi.org/10.18653/v1/2021.semeval-1.67). URL: <https://aclanthology.org/2021.semeval-1.67>.
- Devlin, Jacob et al. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. DOI: [10.48550/ARXIV.1810.04805](https://doi.org/10.48550/ARXIV.1810.04805). URL: <https://arxiv.org/abs/1810.04805>.
- (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423). URL: <https://aclanthology.org/N19-1423>.
- Di Nunzio, Giorgio Maria, Elena Gallina, and Federica Vezzani (2024). "UNIPD@ SimpleText2024: A Semi-Manual Approach on Prompting ChatGPT for Extracting Terms and Write Terminological Definitions". In: *CEUR Workshop Proceedings*. Ed. by G. Faggioli et al., [24].
- Di Nunzio, Giorgio Maria et al. (2024). "Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context@ LREC-COLING 2024". In: *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context@ LREC-COLING 2024*.
- Dietterich, Thomas G (2000). "Ensemble methods in machine learning". In: *International workshop on multiple classifier systems*. Springer, pp. 1–15.
- Dobransky, Kerry and Eszter Hargittai (2006). "The disability divide in internet access and use". In: *Information, Communication & Society* 9.3, pp. 313–334.
- Döring, Matthias (2021). "How-to bureaucracy: A concept of citizens' administrative literacy". In: *Administration & Society* 53.8, pp. 1155–1177.
- DuBay, William H (2004). "The Principles of Readability." In: *Online Submission*.

- Duboue, Pablo (2020). *The Art of Feature Engineering: Essentials for Machine Learning*. Cambridge University Press. DOI: [10.1017/9781108671682](https://doi.org/10.1017/9781108671682).
- Elagina, R. and P. Vučić (2024). “AI Contributions to Simplifying Scientific Discourse in SimpleText 2024”. In: *Working Notes of CLEF 2024: Conference and Labs of the Evaluation Forum*. Ed. by G. Faggioli et al. CEUR Workshop Proceedings. CEUR-WS.org.
- Ermakova, Liana et al. (2022a). “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*.
- Ermakova, Liana et al. (2022b). “Overview of the CLEF 2022 SimpleText Lab: Automatic Simplification of Scientific Texts”. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Ed. by Alberto Barrón-Cedeño et al. Cham: Springer International Publishing, pp. 470–494. ISBN: 978-3-031-13643-6.
- Ermakova, Liana et al. (2023). “Overview of the CLEF 2023 SimpleText Task 2: Difficult Concept Identification and Explanation”. In.
- Ermakova, Liana et al. (2024). “Overview of CLEF 2024 SimpleText Track on Improving Access to Scientific Texts”. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024)*. Ed. by Lorraine Goeuriot et al. Lecture Notes in Computer Science. Springer.
- Espinosa-Zaragoza, Isabel et al. (2023). “A review of research-based automatic text simplification tools”. In: *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pp. 321–330.
- Europe, Inclusion (1998). *Información para todos. Las reglas europeas para hacer información fácil de leer y comprender*.
- Faggioli, G. et al., eds. (2024a). *Working Notes of CLEF 2024: Conference and Labs of the Evaluation Forum*. CEUR Workshop Proceedings. CEUR-WS.org.
- Faggioli, Guglielmo et al., eds. (2024b). *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*. CLEF 2024. Grenoble, France: CEUR Workshop Proceedings, CEUR-WS.org.
- Fellbaum, Christiane (2010). “WordNet”. In: *Theory and applications of ontology: computer applications*. Springer, pp. 231–243.
- Fernández, Antonio et al. (Jan. 2009). “Sistema para el pre-procesamiento de textos para el procesamiento del lenguaje natural”. In.
- Ferrés, Daniel et al. (2016). “YATS: yet another text simplifier”. In: *Natural Language Processing and Information Systems: 21st International Conference on Applications of Natural Language*

- to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016, *Proceedings* 21. Springer, pp. 335–342.
- Finnimore, Pierre et al. (2019). “Strong baselines for complex word identification across multiple languages”. In: *arXiv preprint arXiv:1904.05953*.
- García Ferrero, Iker (2018). “Estudio de Word Embeddings y métodos de generación de Meta Embeddings”. In.
- García Muñoz, Óscar (2014). “Lectura fácil: métodos de redacción y evaluación”. In.
- García Subies, Guillem (2021). “Modelos de Transformers para la Clasificación de Texto”. PhD thesis. ETSI_Informatica.
- Goeuriot, Lorraine (2024). *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9–12, 2024, Proceedings, Part I*. Springer Nature.
- Gonzalez-Dios, Itziar (2017). “Análisis de la complejidad y simplificación automática de textos. El análisis de las estructuras complejas en euskera”. In: *Procesamiento del Lenguaje Natural* 58, pp. 155–158.
- Gooding, Sian and Ekaterina Kochmar (2018). “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*, pp. 184–194.
- (2019). “Complex word identification as a sequence labelling task”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1148–1153.
- Granik, Mykhailo and Volodymyr Mesyura (2017). “Fake news detection using naive Bayes classifier”. In: *2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON)*. IEEE, pp. 900–903.
- Gutiérrez-Fandiño, Asier et al. (2021). “Spanish language models”. In: *arXiv preprint arXiv:2107.07253*.
- Hackeling, Gavin (2017). *Mastering Machine Learning with scikit-learn*. Packt Publishing Ltd.
- Haenlein, Michael and Andreas Kaplan (2019). “A brief history of artificial intelligence: On the past, present, and future of artificial intelligence”. In: *California management review* 61.4, pp. 5–14.
- Hernández, Roberto, Carlos Fernández, and Pilar Baptista (2014). “Metodología de la investigación sexta edición”. In: *México DF: McGraw-Hill*.
- Hernández Sampieri, Roberto, Carlos Fernández Collado, Pilar Baptista Lucio, et al. (2018). *Metodología de la investigación*. Vol. 4. McGraw-Hill Interamericana México.

- Hirano, Yuichiro et al. (2024). "GPT-4 Turbo with Vision fails to outperform text-only GPT-4 Turbo in the Japan Diagnostic Radiology Board Examination". In: *Japanese Journal of Radiology*, pp. 1–9.
- Hmida, Firas et al. (2018). "Assisted lexical simplification for French native children with reading difficulties". In: *The Workshop of Automatic Text Adaptation, 11th International Conference on Natural Language Generation*.
- Höfer, Antonia and Mina Mottahedin (2023). "Minanto at SemEval-2023 Task 2: Fine-tuning XLM-RoBERTa for Named Entity Recognition on English Data". In: *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pp. 1127–1130.
- Hopgood, Adrian A (2021). *Intelligent systems for engineers and scientists: a practical guide to artificial intelligence*. CRC press.
- Huertas Mora, Alexander (2020). "Algoritmos de aprendizaje supervisado utilizando datos de monitoreo de condiciones: un estudio para el pronóstico de fallas en máquinas." In.
- Jiménez Revuelta, José Carlos (2023). "Modelos Grandes de Lenguaje y aplicaciones a la generación automática de texto". In.
- Kadhim, Zahraa Saddi, Hasanen S Abdullah, and Khalil I Ghathwan (2023). "Automatically Avoiding Overfitting in Deep Neural Networks by Using Hyper-Parameters Optimization Methods." In: *International Journal of Online & Biomedical Engineering* 19.5.
- Kamilaris, Andreas and Francesc X. Prenafeta-Boldú (2018). "Deep learning in agriculture: A survey". In: *Computers and Electronics in Agriculture* 147, pp. 70–90. ISSN: 0168-1699. DOI: <https://doi.org/10.1016/j.compag.2018.02.016>. URL: <https://www.sciencedirect.com/science/article/pii/S0168169917308803>.
- Kenton, Jacob Devlin Ming-Wei Chang and Lee Kristina Toutanova (2019). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of naacL-HLT*. Vol. 1, p. 2.
- Kincaid, Peter et al. (1975). "Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel". In: URL: <https://api.semanticscholar.org/CorpusID:61131325>.
- Kochmar, Ekaterina (2022). "Getting started with Natural Language Processing". In: (No Title).
- Kublik, Sandra and Shubham Saboo (2022). "GPT-3: Building innovative NLP products using large language models". In: *O'Reilly Media*.
- Kuhn, Max and Kjell Johnson (2019). *Feature engineering and selection: A practical approach for predictive models*. Chapman and Hall/CRC.

- Landis, J Richard and Gary G Koch (1977). "The measurement of observer agreement for categorical data". In: *biometrics*, pp. 159–174.
- Largey, Nicholas et al. (2024). "AIIR Lab Systems for CLEF 2024 SimpleText: Large Language Models for Text Simplification". In: CEUR Workshop Proceedings. Ed. by G. Faggioli et al., [24].
- León, Rolando Alfredo Hernández and Sayda Coello González (2020). *El proceso de investigación científica*. Editorial Universitaria (Cuba).
- Li, Hang (2023). *Machine Learning Methods*. Springer Nature.
- Liebeskind, Chaya, Otniel Elkayam, and Shmuel Liebeskind (2021). "JCT at SemEval-2021 Task 1: Context-aware Representation for Lexical Complexity Prediction". In: *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pp. 138–143.
- Lipková, Martina (n.d.). "Lexical Density in Academic Writing: Lexical Features and Learner Corpora Analysis in L2 Tertiary Students' Essays and Didactic Implications". In: ().
- Liu, Yinhan et al. (2019). "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692*.
- Mahesh, Batta (2020). "Machine learning algorithms-a review". In: *International Journal of Science and Research (IJSR)*. [Internet] 9.1, pp. 381–386.
- Mahowald, Kyle et al. (2024). "Dissociating language and thought in large language models". In: *Trends in Cognitive Sciences*.
- Malik, Muhammad Shahid Iqbal et al. (2023). "Multilingual hope speech detection: A Robust framework using transfer learning of fine-tuning RoBERTa model". In: *Journal of King Saud University - Computer and Information Sciences* 35.8, p. 101736. ISSN: 1319-1578. DOI: <https://doi.org/10.1016/j.jksuci.2023.101736>. URL: <https://www.sciencedirect.com/science/article/pii/S1319157823002902>.
- Malmasi, Shervin and Marcos Zampieri (June 2016). "MAZA at SemEval-2016 Task 11: Detecting Lexical Complexity Using a Decision Stump Meta-Classifer". In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California: Association for Computational Linguistics, pp. 991–995. DOI: [10.18653/v1/S16-1153](https://doi.org/10.18653/v1/S16-1153). URL: <https://aclanthology.org/S16-1153>.
- Mann, R. and T. Mikulandric (2024). "CLEF 2024 SimpleText Tasks 1-3: Use of LLaMA-2 for Text Simplification". In: *Working Notes of CLEF 2024: Conference and Labs of the Evaluation Forum*. Ed. by G. Faggioli et al. CEUR Workshop Proceedings. CEUR-WS.org.
- McCulloch, Warren S and Walter Pitts (1943). "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5, pp. 115–133.

- Mehta, Rahul and Vasudeva Varma (2023). “LLM-RM at SemEval-2023 Task 2: Multilingual Complex NER using XLM-RoBERTa”. In: *arXiv preprint arXiv:2305.03300*.
- Miao, Jing et al. (2024). “Chain of thought utilization in large language models and application in nephrology”. In: *Medicina* 60.1, p. 148.
- Molin, Stefanie (2019). *Hands-On Data Analysis with Pandas: Efficiently perform data collection, wrangling, analysis, and visualization using Python*. Packt Publishing Ltd.
- Mosquera, Alejandro (Aug. 2021a). “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)*. Online: Association for Computational Linguistics, pp. 554–559. DOI: [10.18653/v1/2021.semeval-1.68](https://doi.org/10.18653/v1/2021.semeval-1.68). URL: <https://aclanthology.org/2021.semeval-1.68>.
- (2021b). “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)*, pp. 554–559.
- Mostert, Femke et al. (2022). “University of Amsterdam at the CLEF 2022 SimpleText track”. In: *Proceedings of the Working Notes of CLEF*.
- Mujahid, Muhammad et al. (2023). “Arabic ChatGPT Tweets Classification using RoBERTa and BERT Ensemble Model”. In: *ACM Transactions on Asian and Low-Resource Language Information Processing* 22.8, pp. 1–23.
- Mukherjee, Niloy et al. (2016). “Ju_nlp at semeval-2016 task 11: identifying complex words in a sentence”. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 986–990.
- Nandy, Abhilash et al. (2021). “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)*, pp. 678–682.
- Nokeri, Tshepo Chris (2021). *Data Science Revealed: With Feature Engineering, Data Visualization, Pipeline Development, and Hyperparameter Tuning*. Springer.
- Norman, Alan T (2019). *Aprendizaje automático en acción: Un libro para el lego, guía paso a paso para los novatos*. Tektime.
- North, Kai, Marcos Zampieri, and Matthew Shardlow (2023a). “Lexical Complexity Prediction: An Overview”. In: *ACM Comput. Surv.* 55.9. ISSN: 0360-0300. DOI: [10.1145/3557885](https://doi.org/10.1145/3557885). URL: <https://doi.org/10.1145/3557885>.

- North, Kai, Marcos Zampieri, and Matthew Shardlow (2023b). "Lexical complexity prediction: An overview". In: *ACM Computing Surveys* 55.9, pp. 1–42.
- North, Kai et al. (2024). *MultiLS: A Multi-task Lexical Simplification Framework*. arXiv: [2402.14972](https://arxiv.org/abs/2402.14972) [cs.CL].
- Nunzio, Giorgio Maria Di et al., eds. (May 2024). *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*. Torino, Italia: ELRA and ICCL. URL: <https://aclanthology.org/2024.determinit-1.0>.
- Ortíz, Jéssica Jasmín Herrera et al. (2024). "La inteligencia Artificial y su impacto en la comunicación: Recorrido y perspectivas". In: *Telos: Revista de Estudios Interdisciplinarios en Ciencias Sociales* 26.1, pp. 278–296.
- Ortiz-Zambrano, J. A., C. Espin-Riofrio, and A. Montejo-Ráez (2024). "SINAI Participation in SimpleText Task 2 at CLEF 2024: Zero-shot Prompting on GPT-4-Turbo for Lexical Complexity Prediction". In: *Working Notes of CLEF 2024: Conference and Labs of the Evaluation Forum*. Ed. by G. Faggioli et al. CEUR Workshop Proceedings. CEUR-WS.org, [24].
- Ortiz-Zambrano, Jenny (2023). "Identification of Complex Words in the Academic Domain in Spanish." In: *PLN-DS@ SEPLN*, pp. 77–85.
- Ortiz-Zambrano, Jenny, César Espin-Riofrio, and Arturo Montejo-Ráez (2022a). "Combining Transformer Embeddings with Linguistic Features for Complex Word Identification". In: *Electronics* 12.1, p. 120.
- (2022b). "Transformers for Lexical Complexity Prediction in Spanish Language". In: *Procesamiento del Lenguaje Natural* 69, pp. 177–188.
- (2023a). "LegalEc: A New Corpus for Complex Word Identification Research in Law Studies in Ecuatorian Spanish". In.
- (2023b). "SINAI Participation in SimpleText Task 2 at CLEF 2023: GPT-3 in Lexical Complexity Prediction for General Audience". In.
- (2024a). "Deep Encodings vs. Linguistic Features in Lexical Complexity Prediction (under review)". In: *Neural Computing and Applications*.
- (2024b). "Enhancing Lexical Complexity Prediction through Few-shot Learning with Gpt-3". In: *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pp. 68–76.
- Ortiz-Zambrano, Jenny and Arturo Montejo-Ráez (2017). "VYTEDU: Un corpus de vídeos y sus transcripciones para Investigación en el ámbito educativo". In.

- Ortiz-Zambrano, Jenny and Arturo Montejo-Raéz (2020). "Barriers in Reading Comprehension of University Students: Analysis of the Complicated Words Annotated in the VYTEDU-CW Corpus". In.
- Ortiz-Zambrano, Jenny 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)*, pp. 1075–1083.
- Ortiz-Zambrano, Jenny and Arturo Montejo-Raéz (2020). "Overview of ALexS 2020: First workshop on lexical analysis at SEPLN". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*. Vol. 2664, pp. 1–6.
- Ortiz-Zambrano, Jenny and Arturo Montejo-Raéz (2020). "Barriers in Reading Comprehension of University Students: Analysis of the Complicated Words Annotated in the VYTEDU-CW Corpus". In: *International Journal on Advanced Science, Engineering and Information Technology* 10.5, pp. 1798–1805. ISSN: 2088-5334. DOI: [10.18517/ijaseit.10.5.10809](https://doi.org/10.18517/ijaseit.10.5.10809). URL: http://ijaseit.insightsociety.org/index.php?Option=com_content&view=article&id=9&Itemid=1&article_id=10809.
- Ortiz-Zambrano, Jenny and Eleanor Varela Tapia (2018). "Reading comprehension in university texts: the metrics of lexical complexity in corpus analysis in Spanish". In: *International Conference on Computer and Communication Engineering*. Springer, pp. 111–123.
- Ortiz-Zambrano, Jenny et al. (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, pp. 167–176.
- Ortiz-Zambrano, Jenny A 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)*, pp. 1075–1083.
- Ozdemir, Sinan and Divya Susarla (2018). *Feature Engineering Made Easy: Identify unique features from your dataset in order to build powerful machine learning systems*. Packt Publishing Ltd.
- Paetzold, Gustavo (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)*, pp. 617–622.
- Paetzold, Gustavo and Lucia Specia (2016a). "Semeval 2016 task 11: Complex word identification". In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 560–569.

- Paetzold, Gustavo and Lucia 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)*, pp. 969–974.
- Pal, Aishika, Pranav, and Moumita Pradhan (2023). "Survey of fake news detection using machine intelligence approach". In: *Data Knowledge Engineering* 144, p. 102118. ISSN: 0169-023X. DOI: <https://doi.org/10.1016/j.datak.2022.102118>. URL: <https://www.sciencedirect.com/science/article/pii/S0169023X22001094>.
- Paszke, Adam et al. (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.
- Perkins, Jacob (2010). *Python text processing with NLTK 2.0 cookbook*. PACKT publishing.
- Petersen, Sarah E and Mari Ostendorf (2007). "Text simplification for language learners: a corpus analysis". In: *Workshop on speech and language technology in education*. Citeseer.
- Pilicita, Anabel and Enrique Barra (2023). "Using of Transformers Models for Text Classification to Mobile Educational Applications". In: *IEEE Latin America Transactions* 21.6, pp. 730–736.
- Pitkowski, Elena Fabiana and Javier Vásquez Gamarra (2009). "El uso de los corpus lingüísticos como herramienta pedagógica para la enseñanza y aprendizaje de ELE". In: *Tinkuy: boletín de investigación y debate* 11, pp. 31–51.
- Quijada, Maury and Julie Medero (2016). "Hmc at semeval-2016 task 11: Identifying complex words using depth-limited decision trees". In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 1034–1037.
- Raheja, Supriya et al. (2022). "A hybrid MCDM framework and simulation analysis for the assessment of worst polluted cities". In: *Simulation Modelling Practice and Theory* 118, p. 102540. ISSN: 1569-190X. DOI: <https://doi.org/10.1016/j.simpat.2022.102540>. URL: <https://www.sciencedirect.com/science/article/pii/S1569190X22000417>.
- Rayner, Keith and Susan A Duffy (1986). "Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity". In: *Memory & cognition* 14.3, pp. 191–201.
- Rebollo, Alberto Anula (2008). "Lecturas adaptadas a la enseñanza del español como L2: variables lingüísticas para la determinación del nivel de legibilidad". In: *La evaluación en el aprendizaje y la enseñanza del español como lengua extranjera/segunda lengua: XVIII Congreso Internacional de la Asociación para la Enseñanza del Español como lengua Extranjera (ASELE): Alicante, 19-22 de septiembre de 2007*. Servicio de Publicaciones, pp. 162–170.

- Requena, Begoña Esther Sampedro (2016). "Las TIC y la educación social en el siglo XXI". In: *Edmetic 5.1*, pp. 8–24.
- Reyes, José Gregorio Mesa (2016). *Procesamiento de Lenguaje Natural y su aplicación en servicios de hostelería*. Disponible en: <https://riull.ull.es/xmlui/bitstream/handle/915/2912/Procesamiento%20de%20Lenguaje%20Natural%20y%20su%20aplicacion%20en%20servicios%20de%20hosteleria.pdf?sequence=1>.
- Rico-Sulayes, Antonio (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*.
- Rodríguez, S. (2017). "Extraction of information from emails using natural language processing techniques". In.
- Rogers, Anna, Olga Kovaleva, and Anna Rumshisky (2021). "A primer in BERTology: What we know about how BERT works". In: *Transactions of the Association for Computational Linguistics* 8, pp. 842–866.
- Ronzano, Francesco, 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)*, pp. 1011–1016.
- Rosado Fernández, Eduardo (2020). "Aplicación de técnicas de NLP a la literatura científica biomédica: detección de bases de datos y análisis predictivo de la terminología MeSH". In.
- Rosas, Verónica Pérez, Rada Mihalcea, and Louis-Philippe Morency (2013). "Multimodal sentiment analysis of spanish online videos". In: *IEEE intelligent systems* 28.3, pp. 38–45.
- Rosenblatt, Frank (1958). "The perceptron: a probabilistic model for information storage and organization in the brain." In: *Psychological review* 65.6, p. 386.
- Rothman, Denis (2021). *Transformers for Natural Language Processing: Build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and more*. Packt Publishing Ltd.
- Rothman, Denis and Antonio Gulli (2022). *Transformers for Natural Language Processing: Build, train, and fine-tune deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, and GPT-3*. Packt Publishing Ltd.
- Rouhiainen, Lasse (2018). "Inteligencia artificial". In: *Madrid: Alienta Editorial*.
- Roundy, Philip T, John M Trussel, and Stephan A Davenport (2023). "The text complexity of local government annual reports". In: *Local Government Studies* 49.5, pp. 1135–1156.

- Rumelhart, David E et al. (1995). "Backpropagation: The basic theory". In: *Backpropagation: Theory, architectures and applications*, pp. 1–34.
- Saggion, Horacio et al. (2011). "Text simplification in simplext: Making texts more accessible". In: *Procesamiento del lenguaje natural* 47, pp. 341–342.
- Saggion, Horacio et al. (2015a). "Making it simplext: Implementation and evaluation of a text simplification system for spanish". In: *ACM Transactions on Accessible Computing (TACCESS)* 6.4, pp. 1–36.
- Saggion, Horacio et al. (May 2015b). "Making It Simplext: Implementation and Evaluation of a Text Simplification System for Spanish". In: *ACM Trans. Access. Comput.* 6.4. ISSN: 1936-7228. DOI: [10.1145/2738046](https://doi.org/10.1145/2738046). URL: <https://doi.org/10.1145/2738046>.
- Saggion, Horacio et al. (2023). "Findings of the tsar-2022 shared task on multilingual lexical simplification". In: *arXiv preprint arXiv:2302.02888*.
- Sagi, Omer and Lior Rokach (2018). "Ensemble learning: A survey". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8.4, e1249.
- Sagu, Amit et al. (2023). "Design of Metaheuristic Optimization Algorithms for Deep Learning Model for Secure IoT Environment". In: *Sustainability* 15.3, p. 2204.
- Sapru, Ashtosh and Hervé Bourlard (2015). "Automatic recognition of emergent social roles in small group interactions". In: *IEEE Transactions on Multimedia* 17.5, pp. 746–760.
- Segura-Bedmar, Isabel and Paloma Martínez (Sept. 2017). "Simplifying drug package leaflets written in Spanish by using word embedding". In: *Journal of Biomedical Semantics* 8. DOI: [10.1186/s13326-017-0156-7](https://doi.org/10.1186/s13326-017-0156-7).
- Segura-Bedmar, Isabel and Paloma Martínez (2017). "Simplifying drug package leaflets written in Spanish by using word embedding". In: *Journal of biomedical semantics* 8.1, pp. 1–9.
- Seng, K. and D. Simunovic (2024). "Simplify It Like It's Hot: Making Complex Texts Easy to Digest". In: *Working Notes of CLEF 2024: Conference and Labs of the Evaluation Forum*. Ed. by G. Faggioli et al. CEUR Workshop Proceedings. CEUR-WS.org, [24].
- Senter, RJ and Edgar A Smith (1967). *Automated readability index*. Tech. rep. Cincinnati Univ OH.
- Shahandashti, Kimya Khakzad et al. (2024). *Evaluating the Effectiveness of GPT-4 Turbo in Creating Defeaters for Assurance Cases*. arXiv: [2401.17991](https://arxiv.org/abs/2401.17991) [cs.SE].
- Shardlow, Matthew (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*, pp. 103–109.
- Shardlow, Matthew, Michael Cooper, and Marcos Zampieri (May 2020a). "CompLex — A New Corpus for Lexical Complexity Prediction from Likert Scale Data". English. In: *Proceedings of*

- the 1st Workshop on Tools and Resources to Empower People with Reading Difficulties (READI)*. Marseille, France: European Language Resources Association, pp. 57–62. ISBN: 979-10-95546-45-0. URL: <https://aclanthology.org/2020.readi-1.9>.
- (2020b). “Complex: A new corpus for lexical complexity prediction from likert scale data”. In: *arXiv preprint arXiv:2003.07008*.
- Shardlow, Matthew, Richard Evans, and Marcos Zampieri (2022). “Predicting lexical complexity in English texts: the Complex 2.0 dataset”. In: *Language Resources and Evaluation* 56.4, pp. 1153–1194.
- Shardlow, Matthew et al. (2021a). “SemEval-2021 Task 1: Lexical Complexity Prediction”. In: *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2021)*.
- Shardlow, Matthew et al. (Aug. 2021b). “SemEval-2021 Task 1: Lexical Complexity Prediction”. In: *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Ed. by Alexis Palmer et al. Online: Association for Computational Linguistics, pp. 1–16. DOI: [10.18653/v1/2021.semeval-1.1](https://aclanthology.org/2021.semeval-1.1). URL: <https://aclanthology.org/2021.semeval-1.1>.
- Shiroyama, Tomotaka (2022). “Comparing lexical complexity using two different VE modes: a pilot study”. In: *Intelligent CALL, granular systems and learner data: short papers from EUROCALL 2022*, p. 358.
- Siddharthan, Advait (2006). “Syntactic simplification and text cohesion”. In: *Research on Language and Computation* 4, pp. 77–109.
- Singh, Sushant and Ausif Mahmood (2021). “The NLP cookbook: Modern recipes for transformer based deep learning architectures”. In: *IEEE Access* 9, pp. 68675–68702.
- Song, Bingyan et al. (2021). “DeepBlueAI at SemEval-2021 Task 7: Detecting and Rating Humor and Offense with Stacking Diverse Language Model-Based Methods”. In: *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pp. 1130–1134.
- Spaulding, Seth (1956). “A Spanish readability formula”. In: *The Modern Language Journal* 40.8, pp. 433–441.
- Specia, Lucia, Sujay Kumar Jauhar, and Rada Mihalcea (2012). “Semeval-2012 task 1: English lexical simplification”. In: * *SEM 2012: The First Joint Conference on Lexical and Computational Semantics—Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pp. 347–355.
- Štajner, Sanja and Goran Glavaš (2017). “Leveraging event-based semantics for automated text simplification”. In: *Expert systems with applications* 82, pp. 383–395.

- Štajner, Sanja et al. (2012). "What can readability measures really tell us about text complexity". In: *Proceedings of workshop on natural language processing for improving textual accessibility*. Citeseer, pp. 14–22.
- Suter, Julia, Sarah Ebling, and Martin Volk (2016). "Rule-based automatic text simplification for German". In.
- Tacchini, Eugenio et al. (2017). *Some Like it Hoax: Automated Fake News Detection in Social Networks*. arXiv: 1704.07506 [cs.LG].
- Tamayo, Mario et al. (2001). *El proceso de la investigación científica*. Editorial Limusa.
- Tancara, Constantino (1993). "La investigación documental". In: *Temas sociales* 17, pp. 91–106.
- Taya, Yuki et al. (Aug. 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)*. Online: Association for Computational Linguistics, pp. 17–23. DOI: 10.18653/v1/2021.semeval-1.2. URL: <https://aclanthology.org/2021.semeval-1.2>.
- Todirascu, Amalia et al. (2022). "Hector: A hybrid text simplification tool for raw texts in french". In: *12th International Conference on Language Resources and Evaluation (LREC)*.
- Torra, Vicenç (2011). "La inteligencia artificial". In: *Lychnos* 7, pp. 14–19.
- Torunoglu-Selamet, Dilara, Tugba Pamay, and Gülsen Eryigit (2016). "Simplification of Turkish sentences". In: *The First International Conference on Turkic Computational Linguistics*, pp. 55–59.
- Tosi, Sandro (2009). *Matplotlib for Python developers*. Packt Publishing Ltd.
- Uluslu, Ahmet Yavuz (2023). *Automatic Lexical Simplification for Turkish*. arXiv: 2201.05878 [cs.CL]. URL: <https://arxiv.org/abs/2201.05878>.
- Valbuena, Roiman (2021). *Inteligencia Artificial: Investigación Científica Avanzada Centrada en Datos*.
- Varadi, D. P. and A. Bartulović (2024). "SimpleText 2024: Scientific Text Made Simpler Through the Use of AI". In: *Working Notes of CLEF 2024: Conference and Labs of the Evaluation Forum*. Ed. by G. Faggioli et al. CEUR Workshop Proceedings. CEUR-WS.org, [24].
- Vaswani, Ashish et al. (2017). "Attention is all you need". In: *Advances in Neural Information Processing Systems*, pp. 5998–6008.
- Vendeville, B., L. Ermakova, and P. De Loor (2024). "UBO NLP report on the SimpleText track at CLEF 2024". In: *Working Notes of CLEF 2024: Conference and Labs of the Evaluation Forum*. Ed. by G. Faggioli et al. CEUR Workshop Proceedings. CEUR-WS.org.
- Verdonck, Tim et al. (2024). "Special issue on feature engineering editorial". In: *Machine learning* 113.7, pp. 3917–3928.

- Vettigli, Giuseppe and Antonio Sorgente (2021). "CompNA at SemEval-2021 Task 1: Prediction of lexical complexity analyzing heterogeneous features". In: *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pp. 560–564.
- Wei, Jason et al. (2022). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., pp. 24824–24837. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.
- Wei, Jason et al. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. arXiv: 2201.11903 [cs.CL]. URL: <https://arxiv.org/abs/2201.11903>.
- Wolf, Thomas et al. (2020). "Transformers: State-of-the-art natural language processing". In: *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45.
- Wong, J (2016). "Almost all the traffic to fake news sites is from Facebook, new data show". In: *The Medium*.
- Yaseen, Tuqa Bani et al. (2021). "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)*, pp. 661–666.
- Yimam, Seid Muhie et al. (June 2018a). "A report on the complex word identification shared task 2018". In: *arXiv preprint arXiv:1804.09132*.
- Yimam, Seid Muhie et al. (June 2018b). "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*. Ed. by Joel Tetreault et al. New Orleans, Louisiana: Association for Computational Linguistics, pp. 66–78. DOI: [10.18653/v1/W18-0507](https://doi.org/10.18653/v1/W18-0507). URL: <https://aclanthology.org/W18-0507>.
- Yuan, Yun-Peng et al. (2023). "Government Digital Transformation: Understanding the Role of Government Social Media". In: *Government Information Quarterly* 40.1, p. 101775. ISSN: 0740-624X. DOI: <https://doi.org/10.1016/j.giq.2022.101775>. URL: <https://www.sciencedirect.com/science/article/pii/S0740624X22001113>.
- Zaharia, George-Eduard, Dumitru-Clementin Cercel, and Mihai Dascalu (2021). "UPB at SemEval-2021 Task 1: Combining Deep Learning and Hand-Crafted Features for Lexical Complexity Prediction". In: *arXiv preprint arXiv:2104.06983*.
- Zambrano, Jenny A Ortiz and Arturo Montejo-Ráez (2021). "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.

- Zeng, Jinshan et al. (2024). "InterpretARA: Enhancing Hybrid Automatic Readability Assessment with Linguistic Feature Interpreter and Contrastive Learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 17, pp. 19497–19505.
- Zečević, A. and F. Doljanin (2024). "Simplify It Like It's Hot: Making Complex Texts Easy to Digest". In: *Working Notes of CLEF 2024: Conference and Labs of the Evaluation Forum*. Ed. by G. Faggioli et al. CEUR Workshop Proceedings. CEUR-WS.org, [24].
- Zhang, Dongsong and Tianhua Chen (2024). "Scikit-ANFIS: A Scikit-Learn Compatible Python Implementation for Adaptive Neuro-Fuzzy Inference System". In: *International Journal of Fuzzy Systems*, pp. 1–19.
- Zhuang, Liu et al. (Aug. 2021). "A Robustly Optimized BERT Pre-training Approach with Post-training". English. In: *Proceedings of the 20th Chinese National Conference on Computational Linguistics*. Huhhot, China: Chinese Information Processing Society of China, pp. 1218–1227. URL: <https://aclanthology.org/2021.ccl-1.108>.
- Zotova, Elena et al. (2020). "Vicomtech at ALEXS 2020: Unsupervised Complex Word Identification Based on Domain Frequency". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR-WS, Malaga, Spain*. Malaga, Spain: CEUR-WS.

Annex A

Contributions and participations in workshops and international conferences

A.1 Contributions

Throughout the development of the doctoral project, we have managed to make significant contributions to the scientific field that highlight our active participation in the research community in the field of Natural Language Processing, specifically in the area of Lexical Simplification. The publications generated during this period not only reflect the rigor and quality of our work, but also demonstrate a constant commitment to excellence and innovation in our area of study.

Our participation in various international workshops and conferences has provided a valuable space to present and discuss our findings with experts in the area, receive constructive feedback and establish fruitful collaborations. Each intervention in these international forums has enriched our knowledge, allowed the dissemination of our results and contributed to the advancement of our field of research.

This constant effort and interaction with the international scientific community have been essential to consolidate our contributions and reflect the impact of our doctoral project. Figure

A.1 shows a visual summary of the scientific events in which we have had the opportunity to present our results with our intervention in the international research community.

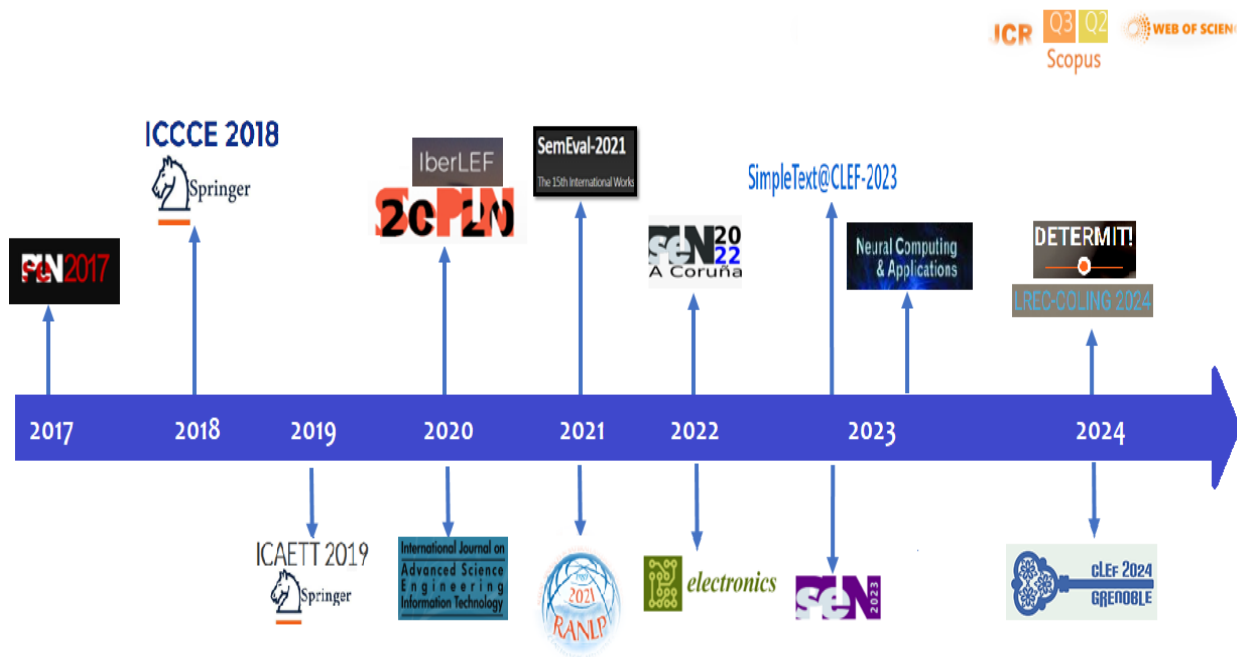


FIGURE A.1: .

A.1.1 JCR–Indexed Journal

- 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.

Impact Factor: 2.9 (2022); 5-Year Impact Factor: 2.9 (2022)

Journal Rank: JCR - Q2(Electrical and Electronic Engineering) CiteScore - Q2 (Electrical and Electronic Engineering)

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

- Ortiz-Zambrano, Jenny A.; Espin-Riofrio, César; Montejo-Ráez, Arturo (2024). Deep Encodings vs. Linguistic Features in Lexical Complexity Prediction. *Neural Computing and Applications*. (Accepted and pending for publication). Springer. JCR Impact Factor 4.5 (2023). Q2 (COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE).

A.1.2 International Conferences - Main Publications

- Revista: SEPLN

Procesamiento del Lenguaje Natural (edición impresa). ISSN: 1135-5948.

Procesamiento del Lenguaje Natural (edición electrónica). ISSN: 1989-7553.

Indexación: Emerging Sources Citation Index (nueva edición de Web of Science).

SCImago Journal Rankings: 2021: 0,270.

Computer Science Applications: Q3.

Language and Linguistics: Q2.

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

<http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5512>

- Ortiz Zambrano, J. A., Montejo Ráez, A. (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, Held Online. INCOMA Ltd..

<https://aclanthology.org/2021.ranlp-1.121/>

RANLP 2021

Deep Learning for Natural Language Processing Methods and Applications

Online ISBN 978-954-452-072-4

e-book site: www.acl-bg.org

Series Print ISSN 1313-8502

Series Online ISSN ISSN 2603-2813

Shoumen, BULGARIA

- 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.

<http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6438>

- 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.
<http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6557>

A.1.3 International Workshop

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

<https://ceur-ws.org/Vol-2664/>

IberLEF 2020 - Iberian Languages Evaluation Forum 2020

Co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020)

CEUR-WS.org, ISSN 1613-0073

- 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).

<https://aclanthology.org/2021.semeval-1.11/>

SemEval-2021 - co-located (virtually) with The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021).

Bangkok, Thailand (online)

ISBN 978-1-954085-70-1

- 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>

SimpleText@CLEF-2023 - co-located with CLEF 2023 Conference and Labs on Evaluation Forum.

Information Access Evaluation meets Multilinguality, Multimodality, and Visualization
Bangkok, Thailand (online)
ISBN 978-1-954085-70-1

- Ortiz-Zambrano, J. (2023). Identification of Complex Words in the Academic Domain in Spanish.

<https://ceur-ws.org/Vol-3625/>

<https://ceur-ws.org/Vol-3625/paper10.pdf>

Doctoral Symposium on Natural Language Processing from the Proyecto ILENIA held as part of the XXXIX edition of the International Conference of the Spanish Society for Natural Language Processing (SEPLN 2023)

Jaén, Spain - 28 September 2023

ISBN 978-2-493814-15-9 ISSN 2951-2093 (COLING); 2522-2686 (LREC)

- Ortiz-Zambrano, J., Espin-Riofrio, C., & Montejo-Ráez, A. (2023). Enhancing Lexical Complexity Prediction Through Few-Shot Learning with GPT-3.

<https://aclanthology.org/2024.determit-1.pdf>

DeTermIt! Workshop Evaluating Text Difficulty in a Multilingual Context - Co-Located with the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation LREC-COLING 2024.

Torino, Italia - 21 May, 2024

ISBN 978-2-493814-15-9 ISSN 2951-2093 (COLING); 2522-2686 (LREC)

- Ortiz-Zambrano, J., Espin-Riofrio, C., & Montejo-Ráez, A. (2024). SINAI Participation in SimpleText Task 2 at CLEF 2024: Zero-shot Prompting on GPT-4-Turbo for Lexical Complexity Prediction.

<https://ceur-ws.org/Vol-3740/>

CLEF 2024 - Conference and Labs of the Evaluation Forum Information Access Evaluation meets Multilinguality, Multimodality, and Visualization

Grenoble, France, 9-12 September, 2024.

ISSN 1613-0073

A.1.4 Other Publications

- 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.

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

DOI: https://doi.org/10.1007/978-3-030-12018-4_9

Online ISBN: 978-3-030-12018-4

Publisher Name: Springer, Cham

Print ISBN: 978-3-030-12017-7

- Ortiz Zambrano, J., MontejóRáez, 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.

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

DOI: https://doi.org/10.1007/978-3-030-32022-5_16

Online ISBN: 978-3-030-32022-5

Publisher Name: Springer, Cham

Print ISBN: 978-3-030-32021-8

- Ortiz, Jenny & Montejó-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.

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

DOI: 10.18517/ijaseit

ISSN: 2088-5334

Scopus CiteScore: 1.9

Scimago Journal Rank: Q3

SJR 2021: 0.25