CLexIS^2: A New Corpus for Complex Word Identification Research in Computing Studies

Jenny A. Ortiz-Zambrano
Universidad de Guayaquil
Guayaquil, Ecuador
jenny.ortizz@ug.edu.ec

Arturo Montejo-Raéz
CEATIC - Universidad de Jaén
Jaén, Spain
amontejo@ujaen.es

Abstract

Reading is a complex process not only because of the words or sections that are difficult for the reader to understand. Complex word identification (CWI) is 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. Annotated corpora for English learners are widely available, while they are less common for the Spanish language. In this article, we present CLexIS^2, a new corpus in Spanish to contribute to the advancement of research in the area of Lexical Simplification, specifically in the identification and prediction of complex words in computing studies. Several metrics used to evaluate the complexity of texts in Spanish were applied, such as LC, LDI, ILFW, SSR, SCI, ASL, CS. Furthermore, as a baseline of the primer, two experiments have been performed to predict the complexity of words: one using a supervised learning approach and the other using an unsupervised solution based on the frequency of words on a general corpus.

1 Introduction

Reading is a complex process not only because of the words or sections that are difficult for the reader to understand. Therefore, an adequate understanding of the content of the texts is required to be able to create coherent mental representations and this way to be able to capture their content (van den Broek, 2010).

Information technologies make it possible for people to access abundant information in different areas such as education, news, social, health, or government, among others. However, this information is not accessible to many, since some people face great reading barriers such as long sentences, unusual words, or complex linguistic structures that do not allow them to understand the content of the texts, with people with intellectual disabilities and people being directly affected in learning; including university students, who are people with a high educational level and specialized knowledge in different subjects of study but, still, could be part of groups of people with reading disabilities (Alarcón et al., 2020).

Complex Word Identification (CWI) is the task of detecting words in the contents of texts that are difficult or complex for people in a certain group to understand (Rico-Sulayes, 2020). CWI and the substitution of words identified as complex may significantly improve readability and understandability of a given text (Zotova et al., 2020).

In recent years, CWI has become an area of great interest for the scientific community and researchers in computational linguistics proposing development of computational semantic analysis systems as evidenced by the shared tasks of CWI by (Paetzold and Specia, 2016) in SemEval 2016, and NAACL-HTL 2018 by (Yimam et al., 2018), the task of the CWI of the ALexS 2020 contest, headquarters of IberLEF 2020 by (Ortiz-Zambranoa and Montejo-Ráezb, 2020), and 15th edition of SemEval but the first Lexical Complexity Prediction task. (Shardlow et al., 2021).

Annotated English Learner Corpus are widely available, Spanish Large Learner Corpus are far less common (Davidson et al., 2020). Although there are corpus for Natural Language Processing (NLP) research in Spanish, they do not contain the necessary annotations to develop reading comprehension tools for students in computing science.

Our aim is to begin to address the lack of data recorded in the corpora of written learner Spanish. This article introduces the creation of a new corpus in Spanish to contribute to the advances of research in the area of Lexical Simplification, specifically in the identification and prediction of complex words in computing studies.

The corpus is named *CLexIS*^{2}, and it is made

Proceedings of Recent Advances in Natural Language Processing, pages 1075–1083
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_121

up of a collection of academic texts from the degrees in Computer Systems Engineering and the Software degree of the Faculty of Mathematical Sciences of the University of Guayaquil (Ecuador), a public institution, and one of the largest and oldest in the country, with around 67,000 students (according to the census in 2019).

2 Related work

2.1 Corpora for CWI in Spanish.

(Pitkowski and Gamarra, 2009) insure that a corpus is a large collection of different types of texts, oral or written, in electronic format, made up of tens of thousands of words, and in some cases made up of several million words. The processing of these large amounts of electronic texts contributes significantly to its application in numerous areas of study in the field of linguistics such as learning a second language (L2), lexical and syntactic simplification, predictions, automatic translations, retrieval and information extraction, speech synthesis, language analysis, among others. (Davidson et al., 2020) state that few corpus written in Spanish are available to NLP researchers. Some corpus for Spanish do not usually include annotations that facilitate the development of NLP models.

(Ortiz-Zambranoa and Montejo-Ráezb, 2020) recently created a resource that can be used to test complex difficult word identification systems, built to adapt to the educational environment. It is a new annotated corpus of transcripts of teaching classes, called *VYTEDU-CW*. This resource was provided for the ALexS workshop (Task on Lexical Analysis at SEPLN 2020) as part of the second edition of IberLEF 2020 (Iberian Languages Evaluation Forum) that joined the efforts of the IberEval and TASS workshops where participants applied interesting approaches to address the CWI problem in an unsupervised or semi-supervised way.

(Davidson et al., 2020) generated the data corpus of Spanish students *Corpus of Written Spanish of L2 and Heritage Speakers*, or *COWS-L2H* built to help researchers better understand L2 development, examine practices teaching empirically and develop NLP tools and thus provide a better service for the community of Spanish teachers. This resource consists of 3.498 short essays written by students at an American university.

(Miaschi et al., 2020) presented an NLP-based approach to track the evolution of written language proficiency in L2 Spanish students using a wide range of linguistic characteristics automatically drawn from students' written productions. To carry out their purpose, they analyzed the development of students' writing from the *COWS-L2H* (Davidson et al., 2020).

The Complex Word Identification (*CWI*) Shared Task organized as part of the 13th Workshop on Innovative Use of NLP for Creating Educational Applications (BEA), hosted in conjunction with NAACL-HLT'2018, focused on multilingualism and provided data sets containing four languages: English, German, French, and Spanish. According to (Yimam et al., 2018) the goal of the CWI task was to predict which words challenge non-native speakers based on annotations collected from native and non-native speakers.

(Parodi, 2015) proposed the *Corpus of Spanish Learners (CAES - acronym in Spanish)* (*Corpus of Spanish learners* in English). (Segura-Bedmar and Martinez, 2017) used the *EasyDPL* (Easy Drug Package Leaflets) corpus, a collection of 306 booklets written in Spanish and manually annotated with 1400 adverse drug effects and their simplest synonyms. The objective of this work was to improve the readability of leaflets by replacing the terms that describe the effects of drugs with simpler synonyms. They used a vector from a previously trained word embedding model.

2.2 Lexical Complexity Measures.

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 et al., 2012). The task of detecting in the content of the documents the words that are difficult or complex to the people of a certain group is known as complex word identification (CWI) (Rico-Sulayes, 2020). Replacing these words with their simplest synonym can improve the understandability and readability of a given text (Zotova et al., 2020). This process may be adapted for college students by making texts more readable (Alarcón et al., 2020).

(Schnur and Rubio, 2021) conducted a study that focused on the application of lexical complexity operationalized by three measures: lexical diversity, lexical density, and lexical sophistication using the 2.4 million-word written Spanish subsection of the*Corpus of Utah Dual Language Immersion*. The study investigated the effect of the three measures of lexical complexity where it was shown that a broad and deep lexical repertoire is a key charac-

teristic of the most advanced levels of proficiency.

In the research carried out by (Saggion et al., 2015) in the Implementation and Evaluation of a Text Simplification System for Spanish, they applied the Lexical Readability Measures based on the definitions of (Rebollo, 2008) for the calculation of low frequency words.

(Kajiwara and Komachi, 2018) introduced systems named *TMU* for the identification of complex words. *TMU* systems applied random forest classifiers and regressors whose characteristics were the number of characters and words and the frequency of target words in various corpus.

To characterize the corpus we applied in this work several metrics used to evaluate the complexity of texts in Spanish were applied, such as LC, LDI, ILFW, SSR, SCI, ASL, CS considered as an approach to validate the coherence of the manually annotated terms regarding their complexity.

3 A New Dataset

The creation of *CLexIS*[2] gave rise to a new data set in the scope of academic courses at a higher education level. The process of preparing the texts is detailed below:

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 Dictation[1] 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 subject corresponds to an average of 100 texts, and in turn, each text contains an average of 77.29 words. Table 1 shows the descriptive statistics of *CLexIS*[2], with a total of 3,887 texts. In Table 2 the definition of the variables is detailed.

[1]Dictation - https://dictation.io/speech

3.1 Manual Annotation.

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 their university expedient was 8.72 / 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.

3.2 Labeling Process.

An application was developed with free software tools Python, Fire-base, and Cloud Firestore for the creation and management of the database, and the texts were loaded into the system. The taggers had to begin to read 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 data set 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.

As can be seen in table 4, 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.
It is evident that the highest number of words labeled as complex is at the level of 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.

A similar behavior occurs in the Computer Systems Engineering career, the total number of words annotated by a single tagger is much higher than when the words are annotated by more than one

The Statistics of *CLexIS*^{2}

	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 1: Descriptive Statistics of different counters over documents in *CLexIS*^{2}.

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: Definition of the columns in Table 1

tagger, the amount corresponding to the number of difficult words annotated decreases as which increases the number of words annotated by more than one tagger.

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., 2015). These formulas were proposed by (Rebollo, 2008), with the exception of the SSR whose measurement was provided by (Spaulding, 1956). The detail is as follows:
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, the first, at the lexical complexity of reading texts for which the readability indices LC, LDI, ILFW, and SSR were applied; and the 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 that make up the Systems Engineering and Software degrees - 40 subjects in total. The table 3 shows the values that correspond to the lexical complexity metrics detailed in the previous paragraph.

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.

The lexical complexity of the words per semester according to the results, determine that, 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 these *Elective III*, *Legislation in Computing*, *Financial Mathematics*, *Microprocessors*,

	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 3: Results by subject of the application of lexical complexity metrics.

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: Results by semester of the careers of Software and Engineering in Computer Systems about the total number of complex words annotated by the taggers.

Simulation.

In 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, it decreases to 9,975.89, it is the last semester of studies for the student, it has subjects what has been learned is applied throughout their university stay, 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*.

5 The Experiments in CWI on the New Corpus

We carried out two experiments following two major approaches in CWI:

1. Detection of complex words based on the CREA resource. This is an unsupervised, lexicon based approach.

2. Prediction of complex words using a machine learning approach over different lexical features.

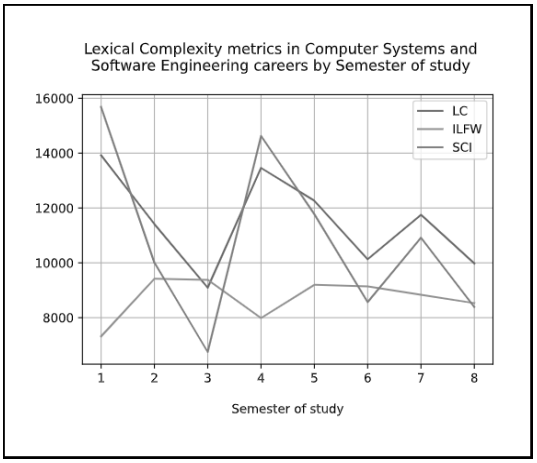

Figure 1: Lexical Complexity in the texts of the semesters of the Software Engineering and Computer Systems careers

5.1 Complex Word Detection System based on CREA.

Based on the definitions of the experiments for the identification of difficult words in Spanish carried out by (Saggion et al., 2015) 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 that allowed the detection of complex words contained in the *CLexIS*[2] corpus using CREA[2]. In Fig. 6 you can see the process flow implemented to automatically

[2]CREA - Royal Spanish Academy Corpus. Royal Spanish Academy frequencies - http://corpus.rae.es/lfrecuencias.html.

detect if a word is complex or not based on CREA. Next, the confusion matrix was performed to determine the effectiveness of the classification system used.

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* of 0.4394, a *precision* of 0.5165 that belong to the really correct positive identifications, the hit rate allowed us to obtain a *recall* of 0.4709, and the proportion of real negatives correctly identified whose measure is the specificity with a value of 0.3963, finally, an *F1* of 0.4927 was obtained that indicates the precision and robustness of the applied model.

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 Software Engineering, and Language and Communication.

5.2 Complex Word Identification using a Machine Learning Approach.

A supervised learning approach was applied using the Random Forest algorithm. The annotated data identifying the simple or multi-word complex word was necessary. Therefore, our system was developed following the process detailed below.

5.2.1 Training/Test Data.

Data from the annotated corpus *CLexIS*^{2}-CW were used. We follow the example of the corpus data model provided by Lexical Complexity Prediction (LCP) shared task, organized by the International Workshop on Semantic Evaluation - SemEval-2021 (Shardlow et al., 2021) for Task 1: Lexical Complexity Prediction on the Lexical semantics track.

The data set consisted of the fields: Id of the text from which the complex word comes, the sentence, the word labeled as complex, and a level of complexity (computed as the division between the number of taggers who scored the word as complex and the total number of taggers). See Table 5.

5.2.2 Features.

To feed the learning algorithm, a total number of 15 characteristics were generated per sample, as in the works of (Gooding and Kochmar, 2018) y (Finnimore et al., 2019) for the detection of complex words:

- Absolute frequency *(abs-frequency)*: the absolute frequency.
- Relative frequency *(rel-frequency)*: the relative frequency of the target word.
- Word length *(length)*: the number of characters of the token.
- Number of syllables *(number-syllables)*: the number of syllables.
- Target word position *(token-position)*: the position of the target word in the sentence.
- Number of words in the sentence *(n-words-sentences)*: number of words in sentence.
- Part Of Speech *(POS)*: the Part Of Speech category.
- Relative frequency of the previous the token *(freq-rel-word-before)*: the relative frequency of the word before the token.
- Relative frequency of the word after the token *(freq-rel-word-after)*: the relative frequency of the word after the token.
- Length of previous word *(len-word-before)*: the number of characters in the word before the token.
- Length of the after word *(len-word-after)*: the number of characters in the word after the token.
- Measure of Textual Lexical Diversity *(MTLD-diversity)*: the lexical diversity of the target word in the sentence using the metric proposed by (McCarthy and Jarvis, 2010), computed using this Python library.
- Number of synonyms *(number-synonyms)*.
- Number of hyponyms *(number-hyponyms)*.
- Number of hyperonyms *(number-hypernyms)*.

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 5: Examples of words tagged by the annotators in the texts of the *CLexIS*^{2} corpus

5.2.3 Applying Supervised Learning.

These numerical features were scaled to a standard range, because it has been proven that many machine learning algorithms when normalized are when they achieve the best results. Besides, a polynomial transformation with a degree value of 2 was applied to the characteristics which produced the creation of new characteristics. The Random Forest algorithm was selected as learning approach. To build the Random Forest regression model, the data set was divided into: the training set and the test set, where 10% of the data set was used as the test set and the remaining 90% was used as training set.

Several runs were performed with different configuration values to observe the performance of the algorithm and fine-tune the hyper-parameters of the model. The results on the test set lead the configuration with best scores, which consisted in 241 nodes and all the 15 features considered. The best results reported a MAE of 0.060970, MSE of 0.005889 and RMSE of 0.076739. See Table 6.

# Trees	MAE	MSE	RMSE
241	0.060970	0.005889	0.076739
241	0.060973	0.005888	0.076733
230	0.060980	0.005894	0.076771

Table 6: Best results obtained with the Random Forest algorithm.

6 Conclusions

A new corpus was created and made available. We believe that it becomes a fundamental resource for the identification of complex words in computer science studies, which means a very useful resource for the development of effective NLP tools for university students. The texts used as a central source of linguistic information reveal the difficulties faced by students of computer science studies.

The application of the lexical complexity metrics allowed evaluating 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.

Future works could propose solutions that involve the creation of tools applying lexical simplification that greatly contribute to the contribution of students with low reading comprehension or intellectual disabilities to better understand the content of texts in the area of computer science. A possible solution will be the creation of a system that transforms complex texts into accessible ones, benefiting mainly university students in computer science who have disabilities and those who have reading comprehension difficulties.

The best result obtained for the predictive value of the words for the data set was: MAE of 0.060970, MSE of 0.005889 and RMSE of 0.09687 in the configuration with 240 nodes and 15 selected characteristics. The resource is available and can be shared by contacting the authors.

7 Acknowledgments

We appreciate Darwin Fabricio Borbor Merejildo, Kevin Francisco Labre Hidalgo, graduates of the Computer Systems Engineering degree from the University of Guayaquil, for their valuable contribution to the development of our work.

This study is partially funded by the Spanish Government under the LIVING-LANG project (RTI2018-094653-B-C21).

References

Rodrigo Alarcón, Lourdes Moreno, and Paloma Martínez. 2020. Hulat-alexs cwi task-cwi for language and learning disabilities applied to university educational texts. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR-WS, Malaga, Spain*.

Paul van den Broek. 2010. Using texts in science education: Cognitive processes and knowledge representation. *Science (New York, N.Y.)*, 328:453–6.

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

Sam Davidson, Aaron Yamada, Paloma Fernandez Mira, Agustina Carando, Claudia H Sanchez Gutierrez, and Kenji Sagae. 2020. Developing nlp tools with a new corpus of learner spanish. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7238–7243.

Pierre Finnimore, Elisabeth Fritzsch, Daniel King, Alison Sneyd, Aneeq Ur Rehman, Fernando Alva-Manchego, and Andreas Vlachos. 2019. Strong baselines for complex word identification across multiple languages. 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)*, pages 970–977, Minneapolis, Minnesota. Association for Computational Linguistics.

Sian Gooding 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*, pages 184–194.

Tomoyuki Kajiwara and Mamoru Komachi. 2018. Complex word identification based on frequency in a learner corpus. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 195–199, New Orleans, Louisiana. Association for Computational Linguistics.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Alessio Miaschi, Sam Davidson, Dominique Brunato, Felice Dell'Orletta, Kenji Sagae, Claudia Helena Sanchez-Gutierrez, and Giulia Venturi. 2020. Tracking the evolution of written language competence in l2 spanish learners. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 92–101.

Jenny A. Ortiz-Zambranoa and Arturo Montejo-Ráezb. 2020. Overview of alexs 2020: First workshop on lexical analysis at sepln.

Gustavo Paetzold and Lucia Specia. 2016. Semeval 2016 task 11: Complex word identification. pages 560–569.

Giovanni Parodi. 2015. Corpus de aprendices de español (caes). *Journal of Spanish Language Teaching*, 2(2):194–200.

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

Alberto Anula Rebollo. 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*, pages 162–170. Servicio de Publicaciones.

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

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Trans. Access. Comput.*, 6(4).

Erin Schnur and Fernando Rubio. 2021. Lexical complexity, writing proficiency and task effects in spanish dual language immersion.

Isabel Segura-Bedmar and Paloma Martinez. 2017. Simplifying drug package leaflets written in spanish by using word embedding. *Journal of Biomedical Semantics*, 8.

Matthew Shardlow, Richard Evans, Gustavo Paetzold, and Marcos Zampieri. 2021. Semeval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2021)*.

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

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.

Elena Zotova, Montse Cuadros, Naiara Perez, and Aitor García-Pablos. 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*.