

PUCP-Metrix: A Comprehensive Open-Source Repository of Linguistic Metrics for Spanish

Javier Alonso Villegas Luis[†] and Marco Antonio Sobrevilla Cabezudo^{†‡}

[†]Research Group on Artificial Intelligence, Pontificia Universidad Católica del Perú

[‡]Aveni

{alonso.villegas, msobrevilla}@pucp.edu.pe

Abstract

Linguistic features remain essential for interpretability and tasks involving style, structure, and readability, but existing Spanish tools offer limited coverage. We present PUCP-Metrix, an open-source repository of 182 linguistic metrics spanning lexical diversity, syntactic and semantic complexity, cohesion, psycholinguistics, and readability. PUCP-Metrix enables fine-grained, interpretable text analysis. We evaluate its usefulness on Automated Readability Assessment and Machine-Generated Text Detection, showing competitive performance compared to an existing repository and strong neural baselines. PUCP-Metrix offers a comprehensive, extensible resource for Spanish, supporting diverse NLP applications.

1 Introduction

Linguistic features have gained renewed importance in explainable NLP, particularly for tasks requiring interpretability, stylistic sensitivity, or attention to surface-level properties. Despite advances in end-to-end neural models, recent work shows that handcrafted or derived features remain essential in applications such as AI-generated text detection (Kumarage et al., 2023; Ciccarelli et al., 2024; Petukhova et al., 2024), educational NLP (Mizumoto and Eguchi, 2023; Hou et al., 2025; Atkinson and Palma, 2025), and readability assessment (Zeng et al., 2024; Liu et al., 2025). In automated essay scoring, for instance, models incorporating linguistic features offer more transparent and pedagogically meaningful evaluations (Hou et al., 2025). These trends highlight the need for robust, modular repositories of linguistic metrics that can complement deep models.

Beyond NLP applications, these repositories also support linguistic research, offering standardized, quantifiable descriptions of texts across genres, registers, and proficiency levels (Jiang, 2016; Kuiken,

2023). They enable empirical analyses of morphosyntactic variation, cohesion, or lexical sophistication, and facilitate cross-linguistic comparisons.

Existing tools have demonstrated the value of this approach. For instance, Coh-Metrix (McNamara et al., 2010) provides extensive metrics for English across various linguistic levels. Similar resources include NILC-Metrix for Portuguese (Leal et al., 2023), Coh-Metrix-Esp for Spanish (Quispesparavia et al., 2016), and MultiAzterTest for Spanish, English, and Basque (Bengoetxea and Gonzalez-Dios, 2021).

In this work, we introduce PUCP-Metrix, a new open-source toolkit for extracting linguistic metrics from Spanish texts. It expands the range of available metrics across lexical, syntactic, discourse, psycholinguistic, and readability dimensions. In addition, we demonstrate its utility in two downstream tasks: Automated Readability Assessment and Machine-Generated Text Detection.

Our main contributions are:

- PUCP-Metrix, a comprehensive and extensible open-source repository of linguistic metrics for Spanish, featuring metrics not available in existing resources.¹
- An empirical study evaluating its usefulness in Automated Readability Assessment and Machine-Generated Text Detection.

2 Related Work

Linguistic analysis tools have played a key role in understanding and quantifying text complexity. Coh-Metrix (McNamara et al., 2010), a widely used tool for English, provides metrics capturing lexical, syntactic, semantic, and discourse characteristics of texts. These metrics support applications from educational assessment to psycholinguistics.

¹The code is available at <https://github.com/iapucp/pucp-metrix>.

tic research, offering a detailed view of text complexity. Inspired by this framework, similar tools have been developed for other languages, adapting metrics to reflect language-specific features.

For Portuguese, NILC-Metrix (Leal et al., 2023) provides a comprehensive set of over 200 metrics covering lexical, syntactic, semantic, discourse, and psycholinguistic dimensions, enabling detailed text analysis for educational and linguistic research. In Spanish, both Coh-Metrix-Esp (Quispesaravia et al., 2016) and MultiAzterTest (Bengoetxea and Gonzalez-Dios, 2021) offer comparable capabilities. Coh-Metrix-Esp adapts the original Coh-Metrix to Spanish, implementing 45 linguistic features and is applied in a Automated Readability Assessment task. MultiAzterTest combines over 125 linguistic and stylistic features with machine learning classifiers to evaluate text complexity in Spanish, English, and Basque.

3 PUCP-Metrix

3.1 System Design

We used an open-source implementation of *Coh-Metrix* for the Spanish language (Quispesaravia et al., 2016) as a starting point for our work. To implement new metrics, we analyzed the metrics in the tools described in Section 2 and consulted their implementation details when available.

3.2 Linguistic Categories and Metrics

For our processing needs, including tokenization, dependency parsing, and POS tagging, we utilized the NLP library spaCy. We developed custom pipelines to extract linguistic metrics from the texts efficiently. Following these steps, we compiled a collection of 182 linguistic metrics for Spanish texts. The complete list is available at Appendix A.

- **Descriptives:** 27 indicators that capture general statistics of the text, such as *number of words*, *number of sentences*, *number of paragraphs*, *minimum and maximum length of sentences*, *average word length*.
- **Lexical Diversity:** 22 indicators measure the diversity of the text’s vocabulary, including the *type-token ratio for various word categories* (*nouns*, *verbs*, etc.), *noun density*, *verb density*, *adverb density*, *adjective density*, the *Maas Index* (Mass, 1972), *MLTD*, and *vocd* (McCarthy Philip M, 2010). Our implementation extends these measures with type-token

ratios for additional word categories and their lemmatized forms. Key indicators include the following:

- **MTLD (Measure of Textual Lexical Diversity):** Addresses TTR’s length sensitivity by calculating the average length of sequential word segments that maintain a certain TTR threshold, providing more stable measures across varying text lengths (McCarthy Philip M, 2010).
- **VOCd (Vocabulary Complexity Diversity):** Estimates vocabulary richness through curve-fitting techniques on random samples, offering insights into the probability of encountering new word types (McCarthy Philip M, 2010).
- **Maas Index:** A logarithmic transformation that provides an alternative measure of lexical diversity, particularly useful for comparing texts of different lengths (Mass, 1972).

- **Readability:** 7 indicators that represent how difficult to understand the text is, such as *Flesch Grade Level*, *Brunet Index*, *Gunning Fog Index*, *Honore’s Statistic*, *SMOG Grade*, *The Szigriszt-Pazos Perspicuity Index* and *Readability μ* . Among the important measures are:

- **Flesch Grade Level (Fernández-Huertas adaptation):** Adapted for Spanish texts, this measure estimates the grade level required for comprehension.
- **Szigriszt-Pazos Perspicuity Index:** A Spanish-specific readability measure that evaluates text clarity, offering insights into Spanish text comprehensibility.
- **SMOG Grade:** Estimates the years of education required to understand a text by analyzing polysyllabic words (3+ syllables).
- **Gunning Fog Index:** Calculates readability by considering both sentence length and complex word percentage, estimating the education level needed for comprehension.
- **Honore’s Statistic:** Measures vocabulary richness by analyzing hapax legomena (words appearing only once).

- *Readability μ* : A statistical measure that evaluates text complexity through letter distribution patterns.
- **Syntactic Complexity**: 12 indicators, reflecting the structural intricacy of text, such as *proportion of sentences with 1-7 clauses, minimal edit distances of words, POS tags and lemmas*. Following *Coh-Metrix*, our implementation extends minimal edit distance measures to POS tags and lemmatized forms, providing comprehensive syntactic variation analysis.
- **Psycholinguistics**: 30 indicators, these reflect psycholinguistic properties of words, specifically how they are understood by humans: *concreteness, imageability, familiarity, age of acquisition, valence and arousal*. These psycholinguistic properties were collected from the EsPal database ([Duchon et al., 2013](#)) and works from [Stadthagen-Gonzalez et al. \(2017\)](#):
 - *Concreteness*: Measures the degree to which words refer to tangible, physical objects versus abstract concepts. Higher concreteness values indicate words that are easier to visualize and process cognitively.
 - *Imageability*: Assesses how easily words can evoke mental images. Words with higher imageability are processed more quickly and remembered more easily.
 - *Familiarity*: Evaluates how well-known words are to speakers. Familiar words are processed faster and require less cognitive effort.
 - *Age of Acquisition*: Measures the age at which words are typically learned. Earlier acquired words are processed more automatically and efficiently.
 - *Valence*: Assesses the emotional positivity or negativity of words. Valence influences emotional processing and memory formation.
 - *Arousal*: Measures the emotional intensity or activation level of words. Arousal affects attention and memory consolidation.
- **Word Information**: 24 indicators, more detailed word-level statistics, such as: *number of nouns, number of verbs, number of adverbs, number of adjectives and number of content words*.
- **Referential Cohesion**: 12 indicators, serve to measure the interconnections within a text: *noun overlap, argument overlap, stem overlap, content word overlap and anaphor overlap*.
- **Textual Simplicity**: 4 indicators, measure the simplicity of the text using the ratio of short or large sentences, such as: *proportion of short sentences, proportion of medium sentences, proportion of long sentences, proportion of very long sentences*.
- **Semantic Cohesion**: 8 indicators, assessing the degree of semantic relatedness between different parts of the text, such as: *LSA overlap of adjacent sentences, LSA overlap of all sentences, LSA overlap of adjacent paragraphs*.
- **Word Frequency**: 16 indicators, various measurements involving the Zipf's frequency for different kinds of words, such as *rare nouns count, rare verbs count, rare adverbs count, rare content words count and mean word frequency*.
- **Syntactic Pattern Density**: 14 indicators, reflecting the density of various syntactic elements, such as: *noun phrase density, verb phrase density, negative expressions density, coordinating conjunctions density and subordinating conjunction density*.
- **Connectives**: 6 indicators, measuring the use of words or phrases that establish logical, temporal, or other relationships between different parts of the text, such as: *casual connectives incidence, logical connectives incidence, adversative connectives incidence, temporal connectives incidence, additive connectives incidence, all connectives incidence*.

3.3 Comparison with Existing Tools

Table 1 shows the number of linguistic metrics implemented in *Coh-Metrix-Esp*, *MultiAzterTest* and *PUCP-Metrix* (ours). *PUCP-Metrix* provides a broader coverage of linguistic metrics compared to *CohMetrix-Esp* and *MultiAzterTest*, comprising a total of 182 metrics across 13 categories. Notably, *PUCP-Metrix* includes metrics in categories

that are entirely missing or underrepresented in the other tools, such as Semantic Cohesion, Textual Simplicity, and Psycholinguistics, with 8, 4, and 30 metrics, respectively. This way, PUCP-Metrix can capture higher-level discourse, cognitive readability, and psycholinguistic properties.

Furthermore, PUCP-Metrix distributes its metrics more evenly across lexical, syntactic, semantic, and psycholinguistic dimensions. This comprehensive and balanced coverage allows for a more detailed and nuanced characterization of texts, making PUCP-Metrix better suited for in-depth linguistic analysis and a wide range of NLP applications.

4 Applications

In order to verify the usefulness of PUCP-Metrix, we use it for two tasks where linguistic metrics have proven to be helpful in past work. In particular, we select Automated Readability Assessment (ARA) and Machine-Generated Text Detection.

4.1 Automated Readability Assessment (ARA)

We follow an approach similar to that of Vásquez-Rodríguez et al. (2022). The original work introduced a benchmark for ARA of Spanish texts. The authors combined multiple corpora labeled by language proficiency levels and proposed 2-label and 3-label classification schemes.

In contrast, our study comprises four publicly available datasets —CAES, Coh-Metrix-Esp, Kwiziq, and HablaCultura— to ensure reproducibility and open accessibility. We adopt the same label mappings described in the paper, adapting all texts to two readability classification schemas: 2-label (simple, complex) and 3-label (basic, intermediate, advanced). The dataset’s descriptions and the labeling strategy can be found in Appendix B.

Overall, the dataset contains 32,167 instances, distributed across the four sources as follows: 31,149 from CAES, 100 from Coh-Metrix-Esp, 206 from Kwiziq, and 713 from HablaCultura.

We experiment with two readability classification schemas mentioned before. All experiments are performed at the document level². The corpus is divided into 80% training, 10% validation, and 10% test sets, stratified by label. We evaluate models using Precision, Recall, and F1-score.

²We use the same texts that come from the available resources

We implement a baseline model based on the Spanish variant of RoBERTa (named RoBERTa-BNE) (Fandiño et al., 2022)³. We fine-tune RoBERTa-BNE for both the 2-label and 3-label classification tasks using the aforementioned splits.

4.2 Machine-Generated Text Detection

We adopt the AuTeXTification 2023 shared task dataset (Sarvazyan et al., 2023), which comprises over 160,000 texts in English and Spanish across five domains: tweets, reviews, news, legal, and how-to articles and generated by both human and large language models (machine).

For our experiments, we focus on the Machine-generated Text Detection task, which consists of identifying if a text has been created by a human or a machine. The task includes 26,996 human-generated instances and 25,195 machine-generated instances, totaling 52,191 instances. More details about the dataset can be found in Appendix B.

Following the shared task settings of AuTeXTification and in line with the ARA task, we adopt RoBERTa-BNE as our baseline. It is fine-tuned on the official training splits and evaluated on the corresponding test splits to ensure comparability.

For both tasks, we trained various machine learning models: Logistic Regression (LR), XGBoost (XGB), Support Vector Machines (SVM) and Random Forest (RF) on the metrics extracted with both MultiAzter and PUCP-Metrix.

5 Results and Discussion

5.1 Automated Readability Assesment

Table 2 compares PUCP-Metrix, MultiAzter, and RoBERTa-BNE on two-label ARA/complexity classification. PUCP-Metrix slightly outperforms MultiAzter across simple and complex texts, achieving an overall F1 of 97.46 with XGBoost versus 97.24 for MultiAzter. In addition, XGBoost consistently yields the highest F1 scores, with Random Forest also performing competitively, while Logistic Regression and SVM score slightly lower.

RoBERTa-BNE achieves the best overall F1 of 98.30, indicating that although PUCP-Metrix captures rich linguistic cues, deep contextual models excel at detecting subtle semantic patterns.

Table 3 compares PUCP-Metrix, MultiAzter, and RoBERTa-BNE on the 3-label ARA task. PUCP-Metrix again performs slightly better than

³The model was available at <https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne>

Category	CohMetrix-Esp	MultiAzterTest	PUCP-Metrix (ours)
Descriptive	11	22	27
Referential Cohesion	12	10	12
Lexical Diversity	2	20	22
Readability	1	1	7
Connectives	6	12	6
Syntactic Complexity	2	19	12
Pattern Density	3	0	14
Semantic Cohesion	0	0	8
Word Information	11	32	24
Word Frequency	0	15	16
Textual Simplicity	0	0	4
Psycholinguistics	0	0	30
Word Semantic Information	0	4	0
Semantic Overlap	0	6	0
Total	48	141	182

Table 1: Number of linguistic metrics per Category for each tool.

Model	Simple			Complex			F1
	P	R	F1	P	R	F1	
Multiazter							
LR	96.42	97.27	96.85	91.75	89.37	90.54	93.70
XGB	98.05	99.20	98.62	97.57	94.19	95.85	97.24
SVM	96.51	97.32	96.91	91.89	89.62	90.74	93.82
RF	97.25	99.29	98.26	97.76	91.72	94.64	96.45
PUCP-Metrix							
LR	96.68	97.65	97.16	92.87	90.11	91.47	94.31
XGB	98.38	99.08	98.73	97.22	95.18	96.19	97.46
SVM	96.60	97.69	97.14	92.97	89.86	91.39	94.27
RF	97.45	99.20	98.32	97.52	92.34	94.86	96.59
RoBERTa-BNE	99.04	99.24	99.14	97.76	97.16	97.46	98.30

Table 2: Results on 2-label ARA/Complexity Classification task

Multiazter, reaching an overall F1 of 96.72 with XGBoost versus 96.56 for Multiazter, being XGBoost the model that achieves the highest scores.

Similarly to previous results, PUCP-Metrix does not surpass RoBERTa-BNE; that achieves the best overall F1 of 98.13, with near-perfect performance on Basic and Intermediate texts and strong results on Advanced ones.

5.2 Machine-Generated Text Detection

Table 4 shows the performance of various machine learning models using the metrics provided by PUCP-Metrix, and Multiazter. Also, it shows the performance of a RoBERTa-BNE model fine-tuned on the AuTexTification dataset and the overall performance of RoBERTa-BNE and the best model reported at the shared task.

In general, PUCP-Metrix consistently outperforms MultiAzter across classifiers. For human texts, PUCP-Metrix increases F1 scores from 42–51 (Multiazter) to 60–66, and for machine texts from 70–73 to 71–76, showing its ability to capture

linguistic and structural cues critical for detecting human-written content. Tree-based models, particularly XGBoost and Random Forest, leverage PUCP-Metrix most effectively, achieving the highest overall F1 scores.

Compared to RoBERTa-BNE, PUCP-Metrix achieves more balanced performance across classes. While RoBERTa-BNE attains very high precision for human texts (93.96), its recall is low (37.86), yielding an F1 below PUCP-Metrix’s best result. This indicates that contextual embeddings may miss the diversity of human writing, whereas interpretable linguistic metrics maintain robust detection across both classes.

Furthermore, PUCP-Metrix slightly surpasses the best model reported in the shared task (F1 70.77), suggesting that integrating linguistic features with neural models could further improve classification performance.

Finally, we conduct an analysis about what are the linguistic metrics that are more important for classification. In general, Machine-generated text detection relies on features related to frequency, readability, and cohesion, while ARA tasks prioritize descriptive, syntactic, and simplicity features. Details are provided in Appendix C.

6 Tool Usage

PUCP-Metrix can be installed via pip:

```
pip install iapucp-metrix
```

To use the library, we need to import the Analyzer class and call compute_metrics to compute all metrics. The function supports multipro-

Model	Basic			Intermediate			Advanced			F1
	P	R	F1	P	R	F1	P	R	F1	
Multiazter										
LR	91.43	92.56	91.99	85.66	86.30	85.97	83.00	74.20	78.36	85.44
XGB	97.62	98.59	98.10	96.43	96.59	96.51	98.48	91.87	95.06	96.56
SVM	90.54	93.08	91.79	85.29	85.71	85.50	84.72	68.55	75.78	84.36
RF	96.32	98.07	97.18	94.38	94.93	94.66	98.37	85.16	91.29	94.38
PUCP-Metrix										
LR	92.25	92.85	92.55	86.35	86.71	86.53	82.02	77.39	79.64	86.24
XGB	97.68	98.59	98.13	97.16	96.59	96.88	96.72	93.64	95.15	96.72
SVM	91.10	93.55	92.31	86.06	85.63	85.85	82.72	71.02	76.43	84.86
RF	95.55	98.18	96.85	95.11	93.77	94.44	97.63	87.28	92.16	94.48
RoBERTa-BNE	99.30	99.24	99.27	98.83	98.42	98.63	95.50	97.53	96.50	98.13

Table 3: Results on 3-label ARA/Complexity Classification task

Model	Human			Machine			F1
	P	R	F1	P	R	F1	
Multiazter							
LR	70.52	30.28	42.37	61.84	89.93	73.29	57.83
XGB	68.10	39.73	50.18	63.98	85.19	73.08	61.63
SVM	70.43	30.74	42.80	61.95	89.73	73.30	58.05
RF	62.08	43.98	51.49	63.82	78.62	70.45	60.97
PUCP-Metrix							
LR	71.09	55.93	62.61	70.02	81.90	75.49	69.05
XGB	71.34	61.36	65.97	72.33	80.38	76.14	71.06
SVM	71.04	56.05	62.66	70.06	81.82	75.48	69.07
RF	63.57	58.24	60.79	68.85	73.44	71.07	65.93
RoBERTa-BNE	93.96	37.86	53.97	66.48	98.06	79.24	66.61
RoBERTa-Autex*	-	-	-	-	-	-	68.52
Best model*	-	-	-	-	-	-	70.77

Table 4: Results on AuTexTification. *The authors of the shared task only provide F1 in the report.

cessing through spaCy, allowing us to specify the number of workers and the batch size.

```
from iapucp_metrix.analyzer import Analyzer
analyzer = Analyzer()

texts = ["Este es mi ejemplo."]

metrics_list = analyzer.compute_metrics(
    texts,
    workers=4,
    batch_size=2
)

for i, metrics in enumerate(metrics_list):
    print(Readability (Fernández-Huertas):)
    print(f"{metrics['RDFHGL']:.2f}")
```

The output of the code described above is:

```
Readability (Fernández-Huertas):
201.86
```

In addition, PUCP-Metrix supports computing metrics grouped by linguistic categories (via `compute_grouped_metrics`), enabling users to analyze model behavior across dimensions such as lexical, syntactic, and semantic features.

7 Conclusion and Future Work

PUCP-Metrix provides a linguistically rich set of 182 metrics for Spanish, offering broader coverage and a larger metric set than previous resources. Empirical evaluations demonstrate its effectiveness in ARA and Machine-generated text detection tasks. Models trained on these metrics match or outperform baseline neural models, underscoring their ability to capture nuanced linguistic information.

Future work includes expanding the metric set to incorporate more discourse and pragmatic metrics, adapting PUCP-Metrix to other Spanish varieties, and integrating these metrics into pre-trained language models or NLP pipelines. Benchmarking on larger and more diverse tasks/datasets will further validate its robustness and support the development of specialized metric sets.

Limitations

The current evaluation has several limitations. Although PUCP-Metrix has been tested on multiple datasets, the experiments primarily focus on learner essays, children’s texts, and selected AuTexTification domains, leaving its performance on other genres and domains uncertain. Additionally, PUCP-Metrix depends heavily on spaCy-based linguistic processing and external lexicons (e.g., psycholinguistic norms), so parsing errors and coverage gaps in these resources can directly affect the reliability of the computed metrics.

References

- John Atkinson and Diego Palma. 2025. An llm-based hybrid approach for enhanced automated essay scoring. *Nature: Scientific Reports*, 15.

- Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.
- Kepa Bengoetxea and Itziar Gonzalez-Dios. 2021. Multiaztertest: A multilingual analyzer on multiple levels of language for readability assessment. *arXiv preprint arXiv:2109.04870*.
- Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Vittorio Ciccarelli, Cornelia Genz, Nele Mastracchio, Wiebke Petersen, Anna Stein, and Hanxin Xia. 2024. Team art-nat-HHU at SemEval-2024 task 8: Stylistically informed fusion model for MGT-detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1690–1697, Mexico City, Mexico. Association for Computational Linguistics.
- Andrew Duchon, Manuel Perea, Nuria Sebastián-Gallés, Antonia Martí, and Manuel Carreiras. 2013. Espal: One-stop shopping for spanish word properties. *Behavior Research Methods*, 45(4):1246–1258.
- Asier Gutiérrez Fandiño, Jordi Armengol Estapé, Marc Pàmies, Joan Llop Palao, Joaquin Silveira Ocampo, Casimiro Pio Carrino, Carme Armentano Oller, Carlos Rodriguez Penagos, Aitor Gonzalez Agirre, and Marta Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.
- Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.
- Zhaoyi Hou, Alejandro Ciuba, and Xiang Li. 2025. Improving llm-based automatic essay scoring with linguistic features. In *Proceedings of the Innovation and Responsibility in AI-Supported Education Workshop*, volume 273 of *Proceedings of Machine Learning Research*, pages 41–65. PMLR.
- Kevin Jiang. 2016. Douglas biber and bethany gray: Grammatical complexity in academic english: Linguistic change in writing. *Applied Linguistics*, 37.
- Folkert Kuiken. 2023. Linguistic complexity in second language acquisition. *Linguistics Vanguard*, 9(s1):83–93.
- Tharindu Kumarage, Joshua Garland, Amrita Bhattacharjee, Kirill Trapeznikov, Scott Ruston, and Huan Liu. 2023. Stylometric detection of ai-generated text in twitter timelines. *arXiv preprint arXiv:2303.03697*.
- Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.
- Sidney Evaldo Leal, Magali Sanches Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann, and Sandra Maria Aluísio. 2023. Nilc-metrix: assessing the complexity of written and spoken language in brazilian portuguese. *Language, Resources and Evaluation*, 58(1):73–110.
- Angela Leis, Francesco Ronzano, Miguel A Mayer, Laura I Furlong, and Ferran Sanz. 2019. Detecting signs of depression in tweets in spanish: Behavioral and linguistic analysis. *J Med Internet Res*, 21(6).
- Fengkai Liu, Tan Jin, and John S. Y. Lee. 2025. Automatic readability assessment for sentences: neural, hybrid and large language models. *Language Resources and Evaluation*.
- Heinz-Dieter Mass. 1972. Über den Zusammenhang zwischen Wortschatzumfang und Länge eines Textes. *Zeitschrift für Literaturwissenschaft und Linguistik*, 2(8):73.
- Jarvis Scott McCarthy Philip M. 2010. Mtld, vodc-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. In *Behavior Research Methods*, pages 381–392.
- Danielle S. McNamara, Max M. Louwerse, Philip M. McCarthy, and Arthur C. Graesser. 2010. Cohmetrix: Capturing linguistic features of cohesion. *Discourse Processes*, 47(4):292–330.
- Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.
- M. Dolores Molina-González, Eugenio Martínez-Cámarra, M. Teresa Martín-Valdivia, and L. Alfonso Ureña-López. 2014. Cross-domain sentiment analysis using spanish opinionated words. In *Natural Language Processing and Information Systems*, pages 214–219, Cham. Springer International Publishing.
- Giovanni Parodi. 2015. Corpus de aprendices de español (caes). *Journal of Spanish Language Teaching*, 2(2):194–200.
- Kseniia Petukhova, Roman Kazakov, and Ekaterina Kochmar. 2024. PetKaz at SemEval-2024 task 8: Can linguistics capture the specifics of LLM-generated text? In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1140–1147, Mexico City, Mexico. Association for Computational Linguistics.

Andre Quispesaravia, Walter Perez, Marco Sobrevilla Cabezudo, and Fernando Alva-Manchego. 2016. Coh-Metrix-Esp: A complexity analysis tool for documents written in Spanish. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4694–4698, Portorož, Slovenia. European Language Resources Association (ELRA).

Areg Mikael Sarvazyan, José Ángel González, Marc Franco-Salvador, Francisco Rangel, Berta Chulvi, and Paolo Rosso. 2023. Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains. *arXiv preprint arXiv:2309.11285*.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Hans Stadthagen-Gonzalez, Constance Imbault, Miguel A. Pérez Sánchez, and Marc Brysbaert. 2017. Norms of valence and arousal for 14,031 spanish words. *Behavior Research Methods*, 49(1):111–123.

Laura Vásquez-Rodríguez, Pedro-Manuel Cuenca-Jiménez, Sergio Morales-Esquível, and Fernando Alva-Manchego. 2022. A benchmark for neural readability assessment of texts in Spanish. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 188–198, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Jinshan Zeng, Xianchao Tong, Xianglong Yu, Wenyan Xiao, and Qing Huang. 2024. Interpretara: Enhancing hybrid automatic readability assessment with linguistic feature interpreter and contrastive learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19497–19505.

A List of metrics in PUCP-Metrix

B Datasets for Automated Readability Assessment and Machine-generated Text Detection

B.1 Automated Readability Assessment

- CAES (*Corpus de Aprendices del Español*)⁴ (Parodi, 2015). This corpus consists of essays written by learners of Spanish as a foreign language. Each document is annotated with a CEFR level (A1–C2). Following Vásquez-Rodríguez et al. (2022), we map A1–B1 to "simple" and B2–C2 to "complex" for the 2-label schema, and A1–A2 to "basic", B1–B2 to

⁴Available at <https://galvan.usc.es/caes/>

"intermediate" and C1–C2 to "advanced" for the 3-label schema.

- Coh-Metrix-Esp (Quispesaravia et al., 2016). This dataset is a collection of short Spanish stories that includes children’s tales and texts intended for adults. It provides explicit simple and complex labels, directly aligned to our 2-label schema and to "basic" vs "advanced" in the 3-label schema.
- Kwiziq⁵. Kwiziq is an online language-learning platform that offers graded Spanish readings labeled with CEFR levels. We use the available data proposed by Vásquez-Rodríguez et al. (2022) and map the CEFR annotations to our 2- and 3-label classification schemes using the same criteria.
- HablaCultura. This dataset comprises educational readings sourced from the HablaCultura platform⁶, where each text is labeled by instructors with CEFR levels. We use the same level mappings used by Vásquez-Rodríguez et al. (2022).

B.2 Machine-generated Text Detection

Human-generated texts in AuTexTification were sourced from publicly available datasets, including MultiEURLEX (Chalkidis et al., 2021) (legal), XLSUM/MLSUM (Hasan et al., 2021; Scialom et al., 2020) (news), COAR/COAH (Molina-González et al., 2014) (reviews), XLM-Tweets (Barbieri et al., 2022) and TSD (Leis et al., 2019) (tweets), and WikiLingua (Ladhak et al., 2020) (how-to articles). Machine-generated texts were produced using six large language models: three from the BLOOM family (BLOOM-1B⁷, BLOOM-3B⁸, BLOOM-7B⁹) and three from the GPT-3 family (babbage, curie, text-davinci-003).

C Feature Analysis

We applied Anova over our dataset using all the metrics. We set a p-value of 0.05 and remove the features that do not make contribution for our analysis.

⁵The platform is available at <https://www.kwiziq.com/>

⁶Available at <https://hablacultura.com/>

⁷Available at <https://huggingface.co/bigscience/bloom-1b7>.

⁸Available at <https://huggingface.co/bigscience/bloom-3b>.

⁹Available at <https://huggingface.co/bigscience/bloom-7b1>.

Category	Metric Description	Category	Metric Description
Descriptive Indices	DESPC: Paragraph count	Syntactic Complexity Indices	SYNNP: Mean number of modifiers per noun phrase
	DESPCI: Paragraph count incidence per 1000 words		SYNLE: Mean number of words before main verb
	DESSC: Sentence count		SYNMEDwd: Minimal edit distance of words between adjacent sentences
	DESSCI: Sentence count incidence per 1000 words		SYNMEDlem: Minimal edit distance of lemmas between adjacent sentences
	DESWC: Word count (alphanumeric words)		SYNMEDpos: Minimal edit distance of POS tags between adjacent sentences
	DESWCU: Unique word count		SYNCLS1: Ratio of sentences with 1 clause
	DESWCUi: Unique word count incidence per 1000 words		SYNCLS2: Ratio of sentences with 2 clauses
	DESPL: Average paragraph length (sentences per paragraph)		SYNCLS3: Ratio of sentences with 3 clauses
	DESPLd: Standard deviation of paragraph length		SYNCLS4: Ratio of sentences with 4 clauses
	DESSL: Average sentence length (words per sentence)		SYNCLS5: Ratio of sentences with 5 clauses
Readability Indices	DESSLd: Standard deviation of sentence length		SYNCLS6: Ratio of sentences with 6 clauses
	DESSNL: Average sentence length excluding stopwords		SYNCLS7: Ratio of sentences with 7 clauses
	DESSNSLd: Standard deviation of sentence length excluding stopwords	Syntactic Pattern Density Indices	DRNP: Noun phrase density per 1000 words
	DESSLmax: Maximum sentence length		DRNPc: Noun phrase count
	DESSLmin: Minimum sentence length		DRVp: Verb phrase density per 1000 words
	DESWLsy: Average syllables per word		DRVpc: Verb phrase count
	DESWLsyd: Standard deviation of syllables per word		DRNEG: Negation expression density per 1000 words
	DESCWLsy: Average syllables per content word		DRNEGc: Negation expression count
	DESCWLsyd: Standard deviation of syllables per content word		DRGER: Gerund form density per 1000 words
	DESCWLlt: Average letters per content word		DRGERc: Gerund count
Referential Cohesion Indices	DESCWLlt: Standard deviation of letters per content word		DRINF: Infinitive form density per 1000 words
	DESWLlt: Average letters per word		DRINFc: Infinitive count
	DESWLlt: Standard deviation of letters per word		DRCCONJ: Coordinating conjunction density per 1000 words
	DESWNSLlt: Average letters per word (excluding stopwords)		DRCCONJc: Coordinating conjunction count
	DESWNSLlt: Standard deviation of letters per word (excluding stopwords)		DRSConj: Subordinating conjunction density per 1000 words
	DESLLlt: Average letters per lemma		DRSConjc: Subordinating conjunction count
	DESLLlt: Standard deviation of letters per lemma		CNCALL: All connectives incidence per 1000 words
	RDFHG: Fernández-Huertas Grade Level		CNCaus: Causal connectives incidence per 1000 words
	RDSP: Szigriszt-Pazos Perspicuity		CNCLogic: Logical connectives incidence per 1000 words
	RDMU: Readability μ index		CNCADC: Adversative connectives incidence per 1000 words
Lexical Diversity Indices	RDMSOG: SMOG index		CNCTemp: Temporal connectives incidence per 1000 words
	RDFOG: Gunning Fog index		CNCAdd: Additive connectives incidence per 1000 words
	RDHS: Honré Statistic		WRDCONT: Content word incidence per 1000 words
	RDBR: Brunet index		WRDCONTc: Content word count
	CRFNO1: Noun overlap between adjacent sentences		WRDOUN: Noun incidence per 1000 words
	CRFAO1: Argument overlap between adjacent sentences		WRDOUNC: Noun count
	CRFSO1: Stem overlap between adjacent sentences		WRDVERB: Verb incidence per 1000 words
	CRFCW01: Content word overlap between adjacent sentences (mean)		WRDVERBc: Verb count
	CRFCW01d: Content word overlap between adjacent sentences (std dev)		WRADAd: Adjective incidence per 1000 words
	CRFANP1: Anaphore overlap between adjacent sentences		WRADAdc: Adjective count
Word Frequency Indices	CRFNOa: Noun overlap between all sentences		WRADAdV: Adverb incidence per 1000 words
	CRFAOa: Argument overlap between all sentences		WRADAdVc: Adverb count
	CRFSOa: Stem overlap between all sentences		WRDPRO: Personal pronoun incidence per 1000 words
	CRFCWOa: Content word overlap between all sentences (mean)		WRDPROc: Personal pronoun count
	CRFCWOad: Content word overlap between all sentences (std dev)		WRDPRP1s: First person singular pronoun incidence per 1000 words
	CRFANPa: Anaphore overlap between all sentences		WRDPRP1sc: First person singular pronoun count
	LDTTRA: Type-token ratio for all words		WRDPRP1pc: First person plural pronoun incidence per 1000 words
	LDTTRkw: Type-token ratio for content words		WRDPRP2s: Second person singular pronoun incidence per 1000 words
	LDTTRno: Type-token ratio for nouns		WRDPRP2sc: Second person singular pronoun count
	LDTTRvb: Type-token ratio for verbs		WRDPRP2p: Second person plural pronoun incidence per 1000 words
Semantic Cohesion Indices	LDTTRAdv: Type-token ratio for adverbs		WRDPRP2pc: Second person plural pronoun count
	LDTTRAdj: Type-token ratio for adjectives		WRDPRP3s: Third person singular pronoun incidence per 1000 words
	LDTTRLa: Type-token ratio for all lemmas		WRDPRP3sc: Third person singular pronoun count
	LDTTRLno: Type-token ratio for noun lemmas		WRDPRP3p: Third person plural pronoun incidence per 1000 words
	LDTTRLvb: Type-token ratio for verb lemmas		WRDPRP3pc: Third person plural pronoun count
	LDTTRLAdv: Type-token ratio for adverb lemmas		PSYC: Overall concreteness ratio
	LDTTRLAdj: Type-token ratio for adjective lemmas		PSYC0: Very low concreteness ratio (1-2.5)
	LDTTRLpron: Type-token ratio for pronouns		PSYC1: Low concreteness ratio (2.5-4)
	LDTTRLpron: Type-token ratio for relative pronouns		PSYC2: Medium concreteness ratio (4-5.5)
	LDTTRLpron: Type-token ratio for indefinite pronouns		PSYC3: High concreteness ratio (5.5-7)
Textual Simplicity Indices	LDTTRLfn: Type-token ratio for functional words		PSYIM: Overall imageability ratio
	LDMLTd: Measure of Textual Lexical Diversity (MTLD)		PSYIM0: Very low imageability ratio (1-2.5)
	LDVOCd: Vocabulary Complexity Diversity (VoCD)		PSYIM1: Low imageability ratio (2.5-4)
	LDMAas: Maas index		PSYIM2: Medium imageability ratio (4-5.5)
	LDDbn: Noun density		PSYIM3: High imageability ratio (5.5-7)
	LDDbvb: Verb density		PSYFM: Overall familiarity ratio
	LD Dadv: Adverb density		PSYFM0: Very low familiarity ratio (1-2.5)
	LD Adadj: Adjective density		PSYFM1: Low familiarity ratio (2.5-4)
	WFRCno: Rare noun count		PSYFM2: Medium familiarity ratio (4-5.5)
	WFRCnot: Rare noun incidence per 1000 words		PSYFM3: High familiarity ratio (5.5-7)
Psycholinguistic Indices	WFRCvb: Rare verb count		PSYAoA: Overall age of acquisition ratio
	WFRCvbi: Rare verb incidence per 1000 words		PSYAoA0: Very early acquisition ratio (1-2.5)
	WFRCadj: Rare adjective count		PSYAoA1: Early acquisition ratio (2.5-4)
	WFRCadj: Rare adjective incidence per 1000 words		PSYAoA2: Medium acquisition ratio (4-5.5)
	WFRCadv: Rare adverb count		PSYAoA3: Late acquisition ratio (5.5-7)
	WFRCadv: Rare adverb incidence per 1000 words		PSYARO: Overall arousal ratio
	WFRCcw: Rare content word count		PSYARO0: Very low arousal ratio (1-3)
	WFRCcwid: Distinct rare content word count		PSYARO1: Low arousal ratio (3-5)
	WFRCcwid: Distinct rare content word incidence per 1000 words		PSYARO2: Medium arousal ratio (5-7)
	WFMcw: Mean frequency of content words		PSYARO3: High arousal ratio (7-9)
TSSh	WFMw: Mean frequency of all words		PSYVAL: Overall valence ratio
	WFMrv: Mean frequency of rarest words per sentence		PSYVAL0: Very negative valence ratio (1-4)
	WFMrcw: Mean frequency of rarest content words per sentence		PSYVAL1: Negative valence ratio (3-5)
	SECOLSadj: LSA overlap between adjacent sentences (mean)		PSYVAL2: Positive valence ratio (5-7)
	SECOLSadjd: LSA overlap between adjacent sentences (std dev)		PSYVAL3: Very positive valence ratio (7-9)
	SECOLSal: LSA overlap between all sentences (std dev)		TSSh: Ratio of short sentences (<11 words)
	SECOLPadj: LSA overlap between adjacent paragraphs (mean)		TSSRmd: Ratio of medium sentences (11-12 words)
	SECOLPadjd: LSA overlap between adjacent paragraphs (std dev)		TSSRlg: Ratio of long sentences (13-14 words)
	SECOLSgiv: LSA overlap between given and new sentences (mean)		TSSRxl: Ratio of very long sentences (\geq 15 words)
	SECOLSgivd: LSA overlap between given and new sentences (std dev)		

Table 5: List of linguistic metrics implemented in PUCP-Metrix

Figure 1 shows a heatmap representing the coverage of linguistic categories along the ranking, i.e., the distribution of linguistic features as more signals are included. Overall, the contribution of linguistic features varies across tasks. For machine-generated content detection, top-ranked signals are dominated by word frequency, readability, and semantic cohesion metrics. In contrast, descriptive and connective metrics play a more limited role and appear only at later ranks.

For ARA tasks, the importance shifts toward descriptive features, syntactic pattern density, readability, syntactic complexity, and textual simplicity metrics. Conversely, semantic cohesion and connective metrics are comparatively less important.

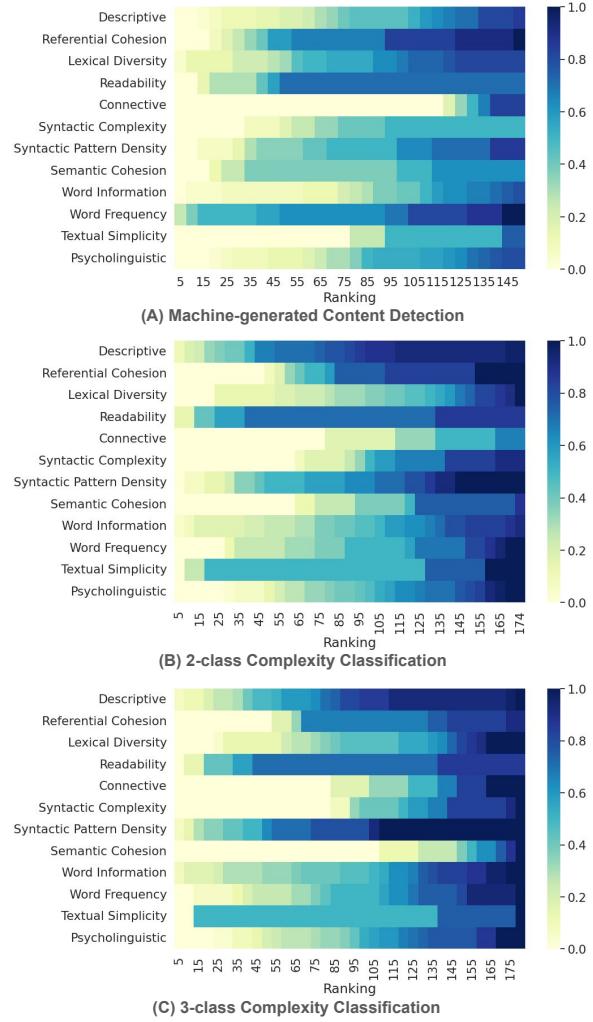


Figure 1: Category Coverage along the ranking for PUCP-Metrix