

CRITERIA: Evaluation Criteria for Quality Corpora in Artificial Intelligence

Lucia Sevilla-Requena^{1,2,*,†}, Jairo Madrigal-Cutillas^{1,†}, José Abreu Salas^{1,†}, Elena Lloret^{1,†} and Isabel Espinosa-Zaragoza^{1,2,†}

¹GPLSI Research Group, University of Alicante, Ctra. San Vicente s/n, 03690, San Vicente del Raspeig, Alicante, Spain

²Instituto Interuniversitario de Lenguas Modernas Aplicadas (IULMA), University of Alicante, Ctra. San Vicente s/n, 03690, San Vicente del Raspeig, Alicante, Spain

Abstract

Criterios de Evaluación para Corpus de Calidad en Inteligencia Artificial (CRITERIA) is a three-year research project dedicated to establishing rigorous standards for AI data resources, more specifically, corpora. Funded by the Ramón Areces Foundation, the project is structured into three distinct modules: Project Coordination (M1), Scientific Development (M2), and Dissemination (M3). The core scientific work (M2) is divided into four strategic tasks: T2.1 The systematisation of quality criteria for multilingual corpora (natural and synthetic data); T2.2 The design and development of automated methods for quality assessment; T2.3 The evaluation of existing benchmark corpora.; T2.4 A practical validation use case focusing on prompt engineering and evaluation. By bringing together a multidisciplinary team of experts from the fields of linguistics and computer engineering at the University of Alicante, CRITERIA aims to deepen the theoretical understanding of corpus compilation while fostering the development of more reliable, high-quality datasets for the global AI community.

Keywords

Natural Language Processing (NLP), Computational Linguistics, Corpus Linguistics.

1. Introduction and Motivation

Criterios de Evaluación para Corpus de Calidad en Inteligencia Artificial (CRITERIA)¹ is a three-year research project focused on developing a systematic, interpretable, and scalable methodological framework to evaluate the quality of linguistic corpora used in the training of LLMs. The project is funded by the Fundación Ramón Areces (Ramón Areces Foundation)², having been awarded a grant in the II Concurso Nacional para la Adjudicación de Ayudas a la Investigación en Humanidades. It is conducted by a multidisciplinary team of five researchers from the GPLSI group³ at the University of Alicante.

1.1. The Evolution of Quality Criteria for Corpora

In recent years, the development of Artificial Intelligence (AI) systems based on Large Language Models (LLMs) has transformed the way we interact with textual data, as well as the possibilities for generating, understanding, and analysing human language on a large scale [1]. In the field of Digital Humanities and Natural Language Processing (NLP), these models have demonstrated unprecedented performance in tasks such as text generation, machine translation, summarisation, semantic classification, and automatic question answering. Nevertheless, this success has been accompanied by a fundamental challenge that still lacks a standardised and systematic solution: the quality of the data used to train these systems [2].

SEPLN 2026: 42nd International Conference of the Spanish Society for Natural Language Processing, León, Spain, 22-25 September 2026.

*Corresponding author.

†These authors contributed equally.

✉ lucia.sevilla@ua.es (L. Sevilla-Requena); Jjmc39@gcloud.ua.es (J. Madrigal-Cutillas); ji.abreu@ua.es (J. A. Salas); elloret@dlis.ua.es (E. Lloret); isabel.espinosa@ua.es (I. Espinosa-Zaragoza)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://criteria.gplsi.es/>

²<https://www.fundacionareces.es/>

³<https://gplsi.dlsi.ua.es/>

In the current context of exponential growth in language models and AI systems, the quality of the corpora used for their training and evaluation has become a critical factor [3]. Despite the increasing use of large volumes of linguistic data, the criteria for defining, constructing, and evaluating high-quality corpora remain inconsistent, implicit, or outdated, especially in light of the challenges of the digital age [4].

Traditionally, the construction of linguistic corpora has been guided by a set of well-established criteria designed to ensure data validity and reliability [5], [6], [7], [8], [9]. These include representativeness, which ensures the corpus reflects the diversity of the language variety being studied; balance, the proportional distribution of text types, genres or registers and modalities; and size, often determined by the statistical requirements of the intended task. Additionally, authorship, thematic control, and chronological setting have served as essential benchmarks for situating the linguistic evidence within a specific social or historical context. Thus, these criteria have long formed the foundation of corpus linguistics, providing a structured methodology for data selection.

Nevertheless, some of these criteria, while adequate in previous contexts, are now clearly insufficient in the face of the challenges posed by LLMs. For example, parameters such as the authority and reputation of sources [10] lose some of their applicability when much of the data comes from open digital spaces, where the boundaries between human-generated content and automated content are becoming increasingly blurred. This is particularly problematic for the development of specialised corpora, where it is increasingly challenging to verify whether the content was produced by a professional in the field, thus undermining the traditional pillars of domain-specific expertise.

Similarly, criteria such as presentation, accessibility, and ergonomics [11] —which focus on the visual organisation of resources and human interaction with data— are sidelined in a scenario where corpora are managed and processed primarily by automated systems and large-scale pipelines. Added to this is the fact that aspects such as updating or maintenance are no longer sufficient to guarantee the quality of corpora that are fed by continuous and changing data streams. This transformation forces us to rethink what a quality corpus means today and which criteria should be prioritised. Issues such as multilingualism, data traceability, the detection and mitigation of bias, representativeness, and noise control emerge as fundamental dimensions that traditional frameworks did not explicitly address, but which are essential in the current data ecosystem for AI.

1.2. The Shift from Quantity to Quality

Traditionally, the predominant approach to language model training has focused on scale: access to large volumes of data has been considered for years as a competitive advantage, and even today it remains a key part of the design of LLMs [12]. However, recent studies [13] have shown that the performance of these models does not depend solely on the amount of data for training, but is closely related to its quality, diversity, and contextual relevance. In this sense, the quality of the corpora is emerging as a critical factor, both from a technical and ethical point of view, which can directly affect the behaviour, robustness, and bias of the models [14].

To address this issue, the research community has begun to develop refined corpora that apply strict cleaning filters such as C4 [15], RedPajama [16], SlimPajama [17], and DCLM-baseline [18]. These filters remove duplicate content, non-textual noise, spam, offensive texts, low-quality linguistic data, and irrelevant documents to transform raw data into clean, useful, and ethically responsible corpora for model training. More recently, datasets such as RefinedWeb [19], FineWeb [20], and FineWeb-2 [21] have set new milestones in the construction of optimised corpora, using modular pipelines that seek to guarantee the final quality of the corpus. That said, there is no clear consensus or standard methodology that systematically defines and measures the quality of a corpus, making it difficult to objectively and reproducibly evaluate these resources.

Furthermore, it is important to emphasise that most of these developments are designed almost exclusively for English, both in terms of language coverage and the quality criteria used. As a result, the ecosystem of training resources for other languages, including Spanish, is considerably more limited, less systematised, and, in many cases, lacking in automated quality assessment tools. This inequality

of resources has significant implications for the training of multilingual models, as well as for the development of linguistic applications in Spanish-speaking contexts. The scarcity of curated corpora (i.e., those that have undergone cleaning, standardisation, and even error correction processes, specifically in Spanish, both in the general domain and in specialised domains) limits the performance of models, increases biases, and reduces their generalisation capacity.

Given this scenario, it is essential to develop a robust, transparent, and reproducible methodological framework that allows for the evaluation of the quality of linguistic corpora from a computational perspective. This framework will include clear criteria and specific metrics that facilitate the analysis of the impact of corpus quality on the behaviour and performance of trained models. To this end, a mixed approach will be adopted that combines the theoretical formulation of quality criteria with their empirical validation through experiments with models trained on corpora with different levels of quality.

1.3. The CRITERIA Project

This project aims to address this gap by creating and validating an updated framework of quality criteria for linguistic corpora, specifically designed for use in training language models and AI systems. The objective is not merely theoretical: the aim is to develop an applicable and reproducible metric that allows for the evaluation of the actual quality of existing or developing corpora. What this project proposes is to create a cross-cutting, methodological tool that serves as a reference framework for evaluating and improving these corpora, thereby strengthening their results, reusability, replicability, and scientific robustness.

In short, the CRITERIA project addresses a critical gap in the current state of NLP and computational linguistics: the lack of rigorous mechanisms for measuring, comparing, and improving the quality of corpora. Its impact will be two-fold: (1) it will provide a solid methodological basis for evaluating corpora in future research; and (2) it will contribute to the development of more accurate, inclusive, sustainable, and ethically responsible language models. In a scenario where LLMs are beginning to be integrated into sensitive systems (such as education, justice, or healthcare), the quality of the data with which they are trained is not only a requirement for ethical responsibility but also a prerequisite for their effective and sustainable use.

2. Team

The research team behind this project has a strong track record in the field of NLP and the Digital Humanities (see Figure 1). All members belong to the GPLSI group and have actively participated in numerous national and international projects focused on the creation, analysis, and transformation of linguistic corpora, as well as the design of AI-based tools. This research, which serves as the foundation for the present project, has highlighted the need for high-quality corpora, not only in terms of cleanliness or the absence of noise, but also in terms of their representativeness, linguistic coverage, textual diversity, and suitability for the specific objectives of each task. The group is inherently interdisciplinary, designed to bridge the gap between theoretical linguistics and computational engineering.

More specifically, the Principal Investigator (PI), a Doctor in Linguistic Studies, and the Working Group provide the necessary theoretical foundation for corpus quality and representativeness. Complementing this, the Research Team and collaborator—computer engineers specialising in NLP and AI-based tool design—ensure technical scalability and implementation, facilitating a continuous feedback loop between humanistic inquiry and technological infrastructure.

This research, which serves as the foundation for this project, has highlighted the need for high-quality corpora, not only for cleanliness and the absence of noise but also for representativeness, linguistic coverage, textual diversity, and suitability for the specific objectives of each task.

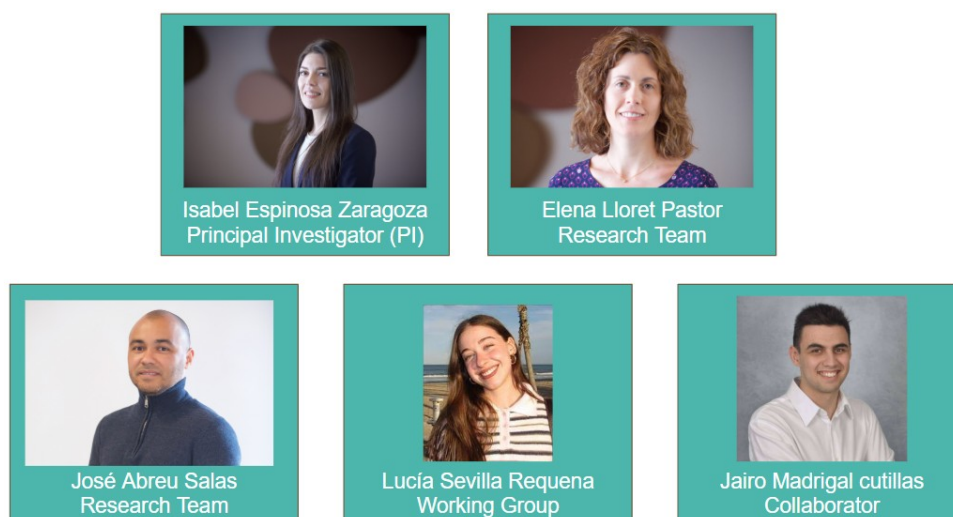


Figure 1: CRITERIA Team

		Year 1 – 2025-26												Year 2 – 2026-27												Year 3 – 2027-28											
		1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
M1	T1.1																																				
	T1.2																																				
M2	T2.1																																				
	T2.2																																				
	T2.3																																				
	T2.4																																				
M3	T3.1																																				

Figure 2: Timeline within the CRITERIA Project.

3. Project Objectives

The CRITERIA project has as its main objective the development of a systematic, interpretable, and scalable methodological framework to evaluate the quality of linguistic corpora used in LLMs. To this end, the following specific objectives are proposed:

- **O1.** To define and systematise a complete set of quality criteria for the linguistic corpora used in NLP research.
- **O2.** To propose a classification methodology that categorises corpora into quality levels (e.g. raw, bronze, silver, golden), according to the defined criteria.
- **O3.** To design a corpus quality framework that is interpretable and, as far as possible, automatable, to operationalise these criteria in quantifiable terms.
- **O4.** To apply the proposed framework to evaluate existing corpora, thus demonstrating its applicability.
- **O5.** To validate the usefulness of the framework through empirical experiments, training language models with corpora of different quality levels and evaluating the resulting performance, robustness, and fairness.

4. Duration and Work Plan

The CRITERIA project has a three-year duration, spanning from November 3rd 2025, to November 3rd, 2028 (see Figure 2). The research is structured into three distinct modules: two transversal modules dedicated to coordination (Module 1) and dissemination (Module 3), and a core scientific module (Module 2), which is further subdivided into four specific tasks (see Figure 3):

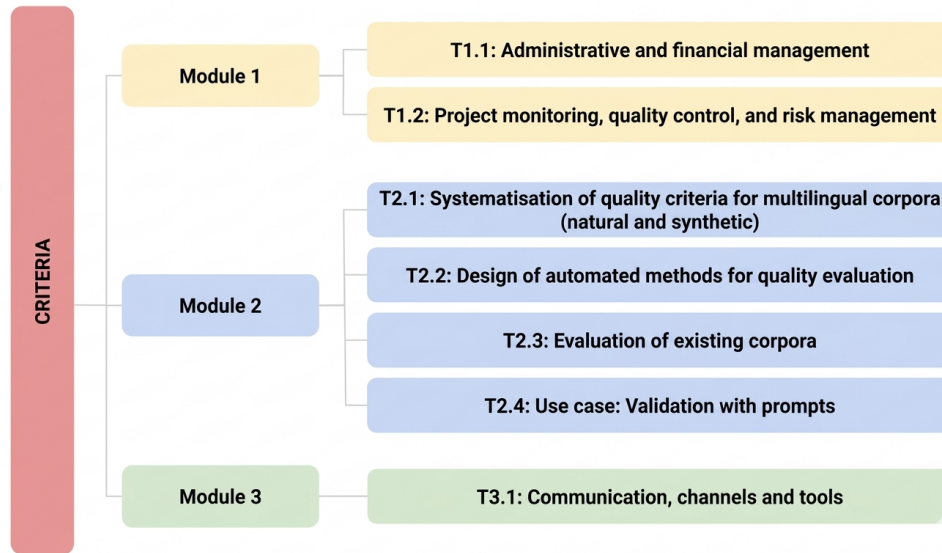


Figure 3: Modules and Tasks within the CRITERIA Project

- **T2.1. Systematisation of quality criteria for multilingual corpora (natural and synthetic)**

The first task entails the conceptual basis of the project and will be carried out mainly during the first year. It will begin with a comprehensive and systematic review of the scientific and technical literature, identifying previous approaches, best practices, and gaps in the definition of quality criteria for linguistic corpora, including natural and synthetic corpora. Seminal contributions on corpus linguistics will be reviewed, such as [5], [6], [7], [8], [9], among others, to define core criteria such as representativeness, size, balance, etc. This review will not only cover research in Corpus Linguistics and NLP, but also emerging regulations and quality standards applicable to textual data in the era of AI. Based on this review, we will proceed to define and structure an integrated set of linguistic, structural, and semantic criteria that will allow us to systematically evaluate the quality of multilingual corpora. An in-depth analysis of possible biases present in the corpora (e.g., gender, language, origin, etc.) will be considered, evaluating their impact on the fairness and reliability of the AI models that use them. Likewise, a layer of pragmatic-discursive analysis will be incorporated, focused on identifying and classifying the predominant communicative intentions and discursive functions in the texts, in order to enrich the quality assessment from a contextual and interpretative perspective. This framework will serve as the basis for subsequent tasks to develop the automation component of these criteria (e.g., metrics, tools, frameworks, or others), as well as for their implementation in the evaluation of the quality of multilingual corpora, ensuring that all results are built on a rigorous and well-founded definition.

- **T2.2. Design of automated methods for quality assessment**

The second task involves the design and development of automated methods, such as metrics, tools, or frameworks, that enable a quantifiable, objective, and reproducible assessment of quality based on the criteria defined in the previous task.

- **T2.3. Evaluation of existing corpora**

Task three involves the application of the criteria, metrics, and other methods developed to analyse already available corpora, in order to validate their suitability and detect possible limitations. For example, the ClearSim corpus [22], INTENT-ES [23], RUN [24], COCOTEROS [25] or VILLANOS [26], among others.

- **T2.4. Use case: Validation with prompts**

Lastly, task four encompasses the practical application of the criteria, metrics, and other methods developed on selected corpora, using existing (pre-trained) language models through prompting techniques. Experiments will be designed to observe the behaviour and sensitivity of these models when using corpora labelled as high or low quality. This validation will allow us to verify the extent to which the proposed criteria are operational and recognisable for current language models, and will serve as empirical evidence of the impact of textual quality on typical NLP tasks.

5. Expected Impacts

The expected impacts of this work are two-fold: (1) at the scientific level, by providing a conceptual and practical framework for the evaluation of corpora in NLP; and (2) at the technological and social level, by promoting the creation of more responsible, inclusive, and reliable language models. Furthermore, it is expected that the results will serve as a basis for the design of automatic corpus evaluation tools and for the development of guidelines to assist researchers, institutions, and developers in the construction of high-quality linguistic resources. In response to the objectives presented in Section 3, we establish the following expected results:

- **R1. Systematisation of a set of quality criteria for linguistic corpora used in NLP:** the development of a solid theoretical framework based on a critical review of the state of the art, which will serve as a conceptual basis for the subsequent stages of the study.
- **R2. Proposal for a typology for classifying linguistic corpora into quality levels:** a precise definition of each level and its potential implications for training language models will be provided.
- **R3. Design and initial validation of a corpus quality framework:** it will combine automated and supervised evaluation. Development of an interpretable, reproducible, and potentially automatable system that operationalises the defined criteria.
- **R4. Evaluation of existing corpora by applying the proposed framework:** this result aims to validate the discriminatory capacity of the metric and establish appropriate threshold values for each quality level.
- **R5. Conducting empirical experiments on training LLMs with corpora of varying quality:** extracting results on the performance, robustness, and fairness of the models, depending on the quality of the training corpus.
- **R6. Critical analysis of experimental results and formulation of recommendations for data curation and selection in future NLP research and development:** identification of weaknesses in current corpora and guidance toward more responsible and efficient practices.

6. Preliminary results

As a preliminary result, we conducted a systematic and critical review of corpus design and compilation practices from 2000 to 2025. The objective was to identify the evolution in the conceptualisation of quality between the pre- and post-LLM eras. The findings indicate a clear paradigm shift in representativeness: a transition from population-based models toward technical and functional implementations. In the post-LLM landscape, the legitimacy of a dataset is increasingly derived from its utility for specific computational tasks rather than its alignment with broad demographic frameworks. Thus, the linguistic corpus has evolved from a descriptive object of study into a fundamental component of technological infrastructure. Consequently, traditional criteria have been redefined: balance now functions as a strategic mechanism for bias mitigation, while size is calibrated against the requirements of algorithmic performance and annotation feasibility.

Declaration on Generative AI

The author(s) have employed Generative AI tools (i.e. Gemini) for proofreading and spelling checking.

Acknowledgments

This research work has been funded by the Ramón Areces Foundation through the project Criteria for Evaluation of Quality Corpora in Artificial Intelligence (CRITERIA), developed within the framework of the II National Call for Research Grants in the Humanities 2025, under the theme “Digital Humanities” (Reference: FRAHUMANIDADES25-01).

References

- [1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. URL: <https://arxiv.org/abs/2005.14165>. arXiv: 2005.14165.
- [2] M. Matoni, A. Kesper, G. Taentzer, How to define the quality of data? a feature-based literature survey, 2025. URL: <https://arxiv.org/abs/2504.01491>. arXiv: 2504.01491.
- [3] K. Zhang, M. Hu, H. A. D. Le, F. K. Torsha, Z. Jiang, M. K. Bui, C.-Y. Chang, Y.-N. Chuang, Z. Xiong, Y. Lin, G. Wang, N. Zou, The llm data auditor: A metric-oriented survey on quality and trustworthiness in evaluating synthetic data, 2026. URL: <https://arxiv.org/abs/2601.17717>. arXiv: 2601.17717.
- [4] X. Yu, Z. Zhang, F. Niu, X. Hu, X. Xia, J. Grundy, What makes a high-quality training dataset for large language models: A practitioners’ perspective, in: Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering, ASE ’24, Association for Computing Machinery, New York, NY, USA, 2024, p. 656–668. URL: <https://doi.org/10.1145/3691620.3695061>. doi:10.1145/3691620.3695061.
- [5] J. Sinclair, Corpus, concordance, collocation (1991).
- [6] J. Sinclair, Corpus and text-basic principles, Developing linguistic corpora: A guide to good practice 92 (2005) 1–16.
- [7] D. Biber, Representativeness in corpus design, Literary and linguistic computing 8 (1993) 243–257.
- [8] T. McEnery, A. Hardie, Corpus linguistics: Method, theory and practice, Cambridge University Press, 2011.
- [9] R. Xiao, Corpus creation, Handbook of natural language processing 2 (2010) 147–165.
- [10] A. Cooke, A guide to finding quality information on the Internet: selection and evaluation strategies, Facet Publishing, 2001.
- [11] M. J. Piano, V. O.-R. Jiménez, Evaluación y calidad de sedes web, Ediciones Trea, 2007.
- [12] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models, 2020. URL: <https://arxiv.org/abs/2001.08361>. arXiv: 2001.08361.
- [13] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, et al., Lima: Less is more for alignment, Advances in Neural Information Processing Systems 36 (2023) 55006–55021.
- [14] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big? , in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’21, Association for Computing Machinery, New York, NY, USA, 2021, p. 610–623. URL: <https://doi.org/10.1145/3442188.3445922>. doi:10.1145/3442188.3445922.
- [15] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring

the limits of transfer learning with a unified text-to-text transformer, *Journal of machine learning research* 21 (2020) 1–67.

- [16] T. Computer, Redpajama: An open source recipe to reproduce llama training dataset, Retrieved from (2023).
- [17] D. Soboleva, F. Al-Khateeb, R. Myers, J. R. Steeves, J. Hestness, N. Dey, Slimpajama: A 627b token cleaned and deduplicated version of redpajama, 2023.
- [18] J. Li, A. Fang, G. Smyrnis, M. Ivgi, M. Jordan, S. Gadre, H. Bansal, E. Guha, S. Keh, K. Arora, et al., Datacomp-lm: In search of the next generation of training sets for language models, *Advances in Neural Information Processing Systems* 37 (2024) 14200–14282.
- [19] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, J. Launay, The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only, *arXiv preprint arXiv:2306.01116* (2023).
- [20] G. Penedo, H. Kydlíček, A. Lozhkov, M. Mitchell, C. Raffel, L. Von Werra, T. Wolf, et al., The fineweb datasets: Decanting the web for the finest text data at scale, *Advances in Neural Information Processing Systems* 37 (2024) 30811–30849.
- [21] G. Penedo, H. Kydlíček, V. Sabolcec, B. Messmer, N. Foroutan, M. Jaggi, L. von Werra, T. Wolf, Fineweb2: A sparkling update with 1000s of languages, Available at:[link] (2024).
- [22] B. Botella-Gil, I. Espinosa-Zaragoza, P. Moreda, M. Palomar, et al., Gplsi: Corpus clearsim (2024).
- [23] M. M. Maestre, E. L. Estevanell-Valladares, R. Sepúlveda-Torres, A. S. Cueto, Enhancing pragmatic processing: A two-dimension approach to detecting intentions in spanish, *Procesamiento del lenguaje natural* 74 (2025) 263–276.
- [24] A. Bonet-Jover, R. Sepúlveda-Torres, E. Saquete, P. Martínez-Barco, M. Nieto-Pérez, Run-as: a novel approach to annotate news reliability for disinformation detection, *Language Resources and Evaluation* 58 (2024) 609–639.
- [25] M. M. Maestre, I. Martínez-Murillo, E. Lloret, P. Moreda, A. S. Cueto, Cocoterros: A spanish corpus with contextual knowledge for natural language generation., in: *SEPLN Posters, 2024*, pp. 36–47.
- [26] B. Botella-Gil, R. Sepúlveda-Torres, A. Bonet-Jover, P. Martínez-Barco, E. Saquete, Semi-automatic dataset annotation applied to automatic violent message detection, *IEEE Access* 12 (2024) 19651–19664.