Procesamiento del Lenguaje Natural, Revista nº 75, septiembre de 2025, pp. 393-400 recibido 03-07-2025 revisado 18-07-2025 aceptado 01-08-2025

Overview of CLEARS at IberLEF 2025: Challenge for Plain Language and Easy-to-Read Adaptation for Spanish texts

Resumen de CLEARS en IberLEF 2025: Reto para la adaptación de textos en español a lenguaje claro y lectura fácil

Beatriz Botella-Gil,[1] **Isabel Espinosa-Zaragoza,**[2] **Alba Bonet-Jover,**[1] **Margot Madina,**[3] **Lucas Molino Piñar,**[4] **Paloma Moreda,**[1] **Itziar Gonzalez-Dios,**[3] **M.Teresa Martín-Valdivia,**[4] **L.Alfonso Ureña-López**[4]

[1]Department of Software and Computing Systems, University of Alicante, Spain

[2]Department of English Philology, University of Alicante, Spain

[3]HiTZ Basque Center for Language Technology - Ixa NLP Group, University of the Basque Country UPV/EHU

[4]Computer Science Department, University of Jaén, Spain

{beatriz.botella, isabel.espinosa, alba.bonet, moreda, raul.gc}@ua.es, itziar.gonzalezd@ehu.eus, margot.madina-gonzalez@h-da.de, {laurena, maite, lmolino}@ujaen.es

Abstract: This paper presents CLEARS, a shared task organized within the framework of the evaluation campaign for Natural Language Processing systems in Spanish and other Iberian languages, IberLEF 2025. The main objective of CLEARS is to investigate and develop automatic techniques for adapting Spanish texts to plain language and easy-to-read formats. Two subtasks are proposed: the first focuses on adapting texts to plain language format, while the second aims to produce easy-to-read texts. CLEARS attracted participation from four teams in Subtask 1 and five teams in Subtask 2. Each team submitted their results and approaches, which are presented below.

Keywords: Plain Language, Easy-to-read, Automatic Text Simplification, Natural Language Processing.

Resumen: Este artículo presenta CLEARS, una tarea compartida organizada en el marco de la campaña de evaluación de sistemas de Procesamiento del Lenguaje Natural en español y otras lenguas ibéricas, IberLEF 2025. CLEARS tiene como objetivo investigar y desarrollar técnicas automáticas para adaptar textos en español a lenguaje claro y lectura fácil. Se proponen dos subtareas, la primera centrada en adaptar textos a lenguaje claro y la segunda en obtener textos en lectura fácil. En CLEARS participaron cuatro grupos en la Subtarea 1 y cinco grupos en la Subtarea 2. Cada equipo presentó sus resultados y enfoques, los cuales se muestran a continuación.

Palabras clave: Lenguaje Claro, Lectura Fácil, Simplificación Automática de Textos, Procesamiento del Lenguaje Natural.

1 Introduction

The inherent difficulty in certain written texts has caused society to demand more transparent and accessible texts. This has resulted in several movements, like the Plain Language (PL) movement,[1] and the Easy-to-Read (E2R) movement.[2]

The aim of PL is to clarify complex texts by formulating and structuring them in such a way that they are understood and reach the whole of society. It seeks to transform the specialized language used in some documents, such as administrative documents,

[1]https://www.plainlanguage.gov/resources/articles/beyond-a-movement/

[2]https://www.inclusion-europe.eu/easy-to-read/

ISSN 1135-5948 DOI 10.26342/2025-75-28

into clear language, which is that which uses a correct, succinct and clear syntax, and an understandable and not complicated lexicon, but without ever renouncing precision and rigour (Tascón and Montolío, 2020).

On the other hand, E2R is concerned with increasing both the reading and comprehension of texts for those with cognitive disabilities such as aphasia, dementia, autism, attention deficit hyperactivity disorder, deafblindness, or dyslexia, to name a few. In this case, there is a set of guidelines published by UNE (AENOR, 2018), that outlines the principles and rules for making a text easy to read in Spanish. These guidelines serve as a reference for adapting texts from the standard language variety to a simpler version that fosters accessibility and ensures the right to understand to every member in society. In this way, people are able to make real informed decisions which promote real inclusion in society.

Currently, the process of adapting a text for E2R or PL is a manual process, making it labor-intensive, time-consuming and costly. Even more, given the enormous amounts of information generated every day, the task of adapting texts in real time has become practically unattainable. This is why Artificial Intelligence (AI), through advancements in Natural Language Processing (NLP), presents a promising solution by automating text adaptation processes.

Most efforts in NLP thus far have focused on Automatic Text Simplification (ATS) processes, either at the lexical, syntactic or discourse level.[3] Specifically, this involves replacing complex vocabulary with simpler alternatives and transforming long, complex sentences into shorter, clearer ones (Bott and Saggion, 2012). The adaptation of a text to E2R or PL goes beyond these processes of lexical and syntactic simplification, that is, it demands an exploration of advanced processes that support and enhance text adaptation to meet the requirements of both PL and E2R.

Previous shared tasks on text simplification include:

1. BEA 2024 Shared Task on the Multilingual Lexical Simplification Pipeline (Yaneva et al., 2024).
2. TSAR-2022 Shared Task on Multilingual Lexical Simplification (Saggion et al., 2022).
3. LREC 2016 Workshop & Shared Task on Quality Assessment for Text Simplification (QATS)(Calzolari et al., 2016).
4. SemEval-2012 Task 1: English Lexical Simplification (Specia, Jauhar, and Mihalcea, 2012).

These initiatives highlight that there is still room for improvement in the two most challenging subtasks of the lexical simplification process and demonstrates the growing emphasis on text simplification as a research area. Nonetheless, they also underscore the need for further efforts to enhance performance and advance the field. Finally, it is important to highlight that while Romance languages such as Spanish, French, Italian, and Portuguese are represented in these tools, most advancements in E2R, PL, and ATS have been made predominantly for English (Espinosa-Zaragoza et al., 2023). In this context, we propose to investigate automatic processes that deal with the adaptation of texts to E2R and PL, especially for Spanish.

2 Task description

With the aim of investigating and developing automatic techniques to adapt Spanish texts into accessible formats —specifically plain language and easy-to-read versions— the CLEARS task (Challenge for plain Language and easy-to-read adaptation for Spanish texts) has been proposed for the shared evaluation campaign IberLEF 2025 (González-Barba, Chiruzzo, and Jiménez-Zafra, 2025). This task is structured into two complementary subtasks, which are described in detail below.

2.1 Subtask 1: Adaptation of texts to PL

The main goal was for participants to adapt these texts to the PL format, following specific simplification criteria to enhance content comprehension (Prodigioso Volcán, 2020). This is not a general simplification but rather an adaptation aligned with the UNE standard, but without the need for all standards to be met in their entirety. The objective is to simplify the text, approaching the standards

[3]For a comprehensive review of ATS tools, see (Espinosa-Zaragoza et al., 2023)

established by law. Participants were able to apply NLP techniques, including the design of algorithms or Machine Learning (ML) models, to transform the original texts into more accessible versions. The following examples illustrate this:

- Original: *Del 17 al 23 de febrero, Madrid se convertirá en la capital del atletismo con tres citas destacadas* (From February 17 to 23, Madrid will become the capital of athletics with three major events).
- PL: *Madrid será la capital del atletismo con 3 eventos importantes del 17 al 23 de febrero* (Madrid will be the capital of athletics with three major events from February 17 to 23).

2.2 Subtask 2: Adaptation of texts to E2R

The goal was for participants to adapt these texts to the E2R format, adhering to the corresponding UNE guidelines (AENOR, 2018) and following accessibility criteria to facilitate content comprehension for specific audiences. To achieve this, participants may employ NLP techniques, such as the use of algorithms or ML models, to transform the original texts into versions that meet the E2R standards. Although in this case it is important to follow each of the rules of the UNE guidelines, it is not necessary to comply with all of them. The number of adapted rules as established in the UNE will be evaluated.This can be exemplified by the following examples:

- Original: *Del 17 al 23 de febrero, Madrid se convertirá en la capital del atletismo con tres citas destacadas* (From February 17 to 23, Madrid will become the capital of athletics with three major events).
- E2R:

  Madrid será la ciudad del Atletismo
  en los próximos días
  3 pruebas importantes de atletismo
  se harán en Madrid del 17 al 23 de febrero.
  (Madrid will be the city of athletics
  in the coming days
  Three important athletics events
  will take place in Madrid from February 17 to 23).

3 Evaluation and metrics

Automatically assessing the quality of E2R texts remains a challenge (Alva-Manchego, Scarton, and Specia, 2021; Al Ajlouni, Li, and Mo'ataz, 2023). Since ATS can be considered a form of intralingual translation, evaluation metrics originally designed for Machine Translation (MT), such as BLEU (Papineni et al., 2002) and BERTscore (Zhang et al., 2020), have been adapted for this purpose. However, these metrics have limitations. For instance, BLEU relies on n-gram overlap and does not explicitly evaluate semantic meaning. Alternatively, SARI (Xu et al., 2016) measures lexical paraphrasing by analyzing which n-grams are inserted, deleted, or retained by the system output compared to human references. Notwithstanding that, SARI penalizes valid simplifications that use synonyms instead of exact matches. SAMSA (Sulem, Abend, and Rappoport, 2018), on the other hand, focuses specifically on sentence splitting quality. In the context of E2R texts, comprehensive lexical and syntactic guidelines must be followed for a text to be considered properly adapted.

Given the availability of a validated corpus created by E2R experts, we proposed using two evaluation metrics to assess the generated texts:

- Cosine Similarity: This metric will be used to measure the textual similarity between the participants' generated texts and the reference texts created by APSA, an NGO specialized in text adaptation for people with disabilities, which collaborated in the creation of the corpus. Higher similarity scores indicate that the generated text aligns more closely with the human expert-adapted versions. To obtain the final similarity score, we calculated the average of two cosine similarity measures:
  1. Syntactic similarity, based on representations that capture surface-level or structural aspects of the text.
  2. Semantic similarity, derived from embeddings that reflect deeper meaning and conceptual alignment between texts.
- Fernández Huerta Readability Index (Fernández-Huerta, 1959): This readability metric, designed specifically for

Spanish texts, is based on the Flesch-Kincaid readability formula for English. It evaluates texts by considering average sentence length and average syllable length, assigning a readability score. A higher readability score suggests a text that is easier to understand.

Additionally, we have attached a help guide for participants regarding the use of the metrics.[4]

The winning submissions were those that achieved, on the one hand, high similarity to the expert-adapted texts — computed as the average of two cosine similarity measures — and, on the other hand, higher readability scores according to the Fernández Huerta Index.

4 Data

The corpus consists of a total of 3,000 news items from various municipalities in the province of Alicante (Spain), covering topics such as sports, culture, leisure, and festivities. It has been developed and validated by a team of expert validators in this field. For each of these articles, two versions have been generated:

- One in PL format, referred to as the "facilitated version" in the corpus. This version adheres to the adaptation criteria but is less strict than the other, particularly in its presentation, though it still aids in text comprehension (related to Subtask 1); and
- Another in E2R format, which strictly complies with the corresponding UNE 153101 EX guidelines (AENOR 2018), including both the language used and its presentation (related to Subtask 2).

The dataset has been divided into two parts: 70% for training and 30% for testing, corresponding to 2,100 and 900 news items, respectively. The news items were selected in Spanish because, as described by (Instituto Cervantes, 2022), although it is the fourth most spoken language in the world and the second most widely spoken mother tongue, it is necessary to develop more corpora to support the adaptation of texts into E2R and PL formats. This will enable NLP tools to more effectively address the challenges associated with accessibility and text comprehension.

In Subtask 1, an additional dataset, LengClaro2023 (Agüera-Marco and Gonzalez-Dios, 2025), was used for testing purposes. This dataset consists of seven texts from the most commonly used procedures on the website of the Spanish *Seguridad Social*, which have been simplified according to PL guidelines. The LengClaro simplifications were specifically employed in this task. The dataset comprises administrative texts, which tend to be more challenging than newspaper articles, providing an opportunity to evaluate system performance across different domains.

5 Systems and results

This section presents the systems and results for each subtask. Specifically, Section 5.1 covers Subtask 1, which focuses on the adaptation of texts to PL, while Section 5.2 addresses Subtask 2, involving the adaptation of texts to E2R.

5.1 Subtask 1: PL adaptation

A total of four different groups participated in Subtask 1. Table 1 presents the official results for Subtask 1 of the CLEARS Challenge 2025. Participating teams were evaluated using two semantic similarity metrics—cosine similarity based on TF-IDF and sentence embeddings—and the Fernández-Huerta readability index. The final ranking was determined by averaging the two cosine similarity scores, while the readability score was reported separately.

According to this composite scoring, the HULAT-UC3M team ranked first with an average cosine similarity of 0.75, followed by NIL_UCM (0.71), and both CardiffNLP and Vicomtech (0.70). These results indicate that although some systems achieved strong semantic alignment, readability scores varied, with Vicomtech obtaining the highest Fernández-Huerta score (82.98), suggesting simpler sentence structures.

It is important to note that the Fernández-Huerta index, though informative, is a surface-level readability metric that may favor syntactically simpler texts regardless of their semantic clarity. Therefore, its inclusion in the final average provides a balance between meaning preservation and textual

[4] https://colab.research.google.com/drive/1WBrl2B_lTABb2JInjCT3L3UH6hJ7PoTo?usp=sharing

Team	Cosine TF-IDF	Cosine Embed.	Fernández-Huerta	Cosine's Avg
HULAT-UC3M	0.71	0.78	69.72	0.75
NIL_UCM	0.67	0.75	70.42	0.71
CardiffNLP	0.63	0.77	78.81	0.70
Vicomtech	0.63	0.76	82.98	0.70

Table 1: Performance comparison of participating teams in Subtask 1 (Plain Language) of the CLEARS challenge at IberLEF 2025. Teams are ranked by average cosine similarity, computed as the mean of TF-IDF and sentence embedding similarities.

Team	Cosine TF-IDF	Cosine Embed.	Fernández-Huerta	Cosine's Avg
NIL_UCM	0.68	0.75	69.40	0.72
CardiffNLP	0.65	0.77	77.85	0.71
UR	0.64	0.76	85.12	0.70
UNED-INEDA	0.60	0.75	72.39	0.68
Vicomtech	0.58	0.74	85.44	0.66

Table 2: Performance comparison of participating teams in Subtask 2 (Easy-to-Read adaptation) of the CLEARS challenge at IberLEF 2025. Teams are ranked by average cosine similarity, computed as the mean of TF-IDF and sentence embedding similarities.

accessibility, but it may not fully reflect actual understandability for end users.

5.2 Subtask 2: E2R adaptation

A total of five different groups participated in Subtask 2. The results obtained from these groups are presented in Table 2. The evaluation metric for this Subtask was based on the average of TF-IDF and sentence embeddings cosine similarity, with the Fernández-Huerta index serving as an additional readability measure.

The NIL_UCM team achieved the highest average cosine similarity score with 0.72. CardiffNLP obtained a score of 0.71, followed by UR (0.70), UNED-INEDA (0.68) and Vicomtech (0.66). NIL_UCM obtained the highest TF-IDF cosine similarity score (0.68). With regard to sentence embeddings cosine similarity, CardiffNLP led the ranking with 0.77. Finally, with respect to the Fernández-Huerta index, Vicomtech achieved the best readability score overall (85.44), closely followed by UR (85.12).

These results indicate that high semantic similarity alone was not enough to secure a top ranking. For instance, VIicomtech had a competitive embedding similarity (0.74) and the highest readability score (85.44), but its lower TF-IDF performance (0.58) significantly reduced its average. In contrast, NIL_UCM achieved the best balance between TF-IDF similarity and embedding similarity, which helped secure the top position despite having the lowest readability score. CardiffNLP's results reflect strong performance in sentence embeddings, which compensated for moderate TF-IDF scores.

6 Submitted approaches

Each participating team was allowed to submit different approaches, although only the last submission would be considered for the competition. All participants, regardless of the task, made use of LLMs in their submissions. Both subtasks 1 and 2 used prompting strategies combined with other automatic processes such as post-editing or refinement.

6.1 Subtask 1

- **HULAT-UC3M**. This team tackled the PL task through prompt engineering over generative models trained in Spanish. They experimented with Salamandra-7B-Instruct and RigoChat-7B-v2, selecting the latter for its balance between semantic fidelity and linguistic clarity. Their final system combined a normalization step, a task-specific prompt for Plain Language, and a LoRA-adapted RigoChat model. It achieved the highest semantic similarity score (SIM = 0.75).
- **NIL-UCM**. This team employed Mistral-7B-Instruct-v0.3, chosen for its balance between coherence, instruction-following ability, and cost-efficiency. They explored two main approaches: only prompting and prompting combined with fine-tuning on CLEARS examples. Prompts were designed

with a mission statement, detailed step-by-step instructions, and good/bad examples for each rule, resulting in a prompt of nearly 20,000 characters. The team also implemented automatic post-processing to remove unintended explanations added by the model, despite explicit instructions to avoid them.

- **CardiffNLP**. This team focused exclusively on prompt-based adaptation using large language models (LLMs), with a final choice of Gemma-3 due to its strong instruction-following capabilities and overall performance. They iterated prompts across several dimensions—including factuality, grammaticality, and format fidelity—and found that operating at the sentence level improved output quality. Prompts were more effective when written in Spanish, and the final outputs were structured as Python dictionaries to facilitate post-processing. While zero-shot prompting was abandoned due to inconsistencies and hallucinations, few-shot prompting with in-domain examples proved most effective.
- **Vicomtech**. This team proposed a text simplification system using LLaMA 3.1 Instruct 8B, evaluated through cosine similarity (at both BoW and embedding levels) and Fernández-Huerta scores. They experimented with multiple initial adaptation strategies, including zero-shot, few-shot (with BM25 and semantic retrieval), supervised fine-tuning, and Direct Preference Optimization (DPO). The most competitive results came from combining BM25-based retrieval or DPO with a post-editing mechanism called APEC (Automatic Post-Editing Cycles), which uses the LLM as both evaluator and editor in iterative refinement loops.

6.2 Subtask 2

In Subtask 2, the teams followed identical approaches to Subtasks 1 and all of them relied on LLMs. UNED-INEDA, CardiffNLP, UR and NIL-UCM used different prompting strategies where they included E2R adaptation guidelines and tested different language models to select their runs. More specifically:

- **NIL-UCM**. This team implemented the Mistral-7B-Instruct-v0.3 model with a prompt that included guidelines and examples.
- **CardiffNLP**. This team employed Gemma-3 and few-shot prompting including the role (persona), adaptation instructions and three examples of the task and formatting instructions.
- **UR**. This team utilized zero-shot prompting with the Gemma model and they included mandatory adaptation instructions from E2R guidelines.
- **UNED-INEDA**. This team used the Dolphin Mistral model with a zero-shot prompt and the specifications of the UNE 153101:2018 EX standard.
- **Vicomtech**. This team also used prompting engineering strategies, but adopted another strategy: the APEC approach, where the outputs of the LLMs are being post-edited iteratively until automatic evaluation metrics indicate that no further improvement was needed. In order to perform the initial adaptations prior to the iterative process, they tested different prompts and strategies incorporating E2R guidelines. The selected approach employed a few-shot method approach combined with lexical RAG, BM25 indexing and querying, and Direct Preference Optimisation.

7 Conclusions and future work

In this work, we have presented the CLEARS 2025 challenge, focused on the automatic adaptation of Spanish texts into accessible formats, specifically Plain Language (PL) and Easy-to-Read (E2R). The results demonstrate that combining large language model-based techniques with automatic post-processing methods can achieve significant adaptations that improve readability while preserving semantic fidelity to expert-adapted original texts.

The corpus and systems developed contribute valuable resources to the relatively underexplored field of NLP for accessibility in Spanish, which has historically been overshadowed by English-centric research. Our findings further underscore the importance of balancing semantic preservation with enhanced readability to produce adaptations

that are both practical and effective for real-world applications, including administrative and informational texts.

Nevertheless, automatic evaluation remains a major challenge. Current metrics, such as readability indices, often fall short in capturing the actual comprehensibility of simplified content, especially in accessibility contexts where understanding by the target audience is paramount. There is thus a pressing need for more descriptive, nuanced, and context-sensitive evaluation frameworks that reflect users' real comprehension and the communicative adequacy of adaptations.

To address this, future work should prioritize the development and validation of such advanced evaluation metrics, along with continued refinement of LLMs through domain-specific training that incorporates expert knowledge and accessibility-oriented linguistic rules. Incorporating human feedback through interactive systems can further personalize and improve outcomes. In parallel, expanding and diversifying annotated corpora across genres and complexity levels will strengthen model robustness. Ultimately, integrating these advancements into user-facing tools will be key to fostering inclusive communication in everyday contexts.

Collectively, these efforts aim to propel automated text simplification towards truly accessible and meaningful communication for all.

Acknowledgments

This research has been supported by multiple funding sources. It has received funding from the Generalitat Valenciana (Conselleria d'Educació, Investigació, Cultura i Esport) through the project NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation (CIPROM/2021/021). CLEAR.TEXT: Enhancing the modernization public sector organizations by deploying Natural Language Processing to make their digital content CLEARER to those with cognitive disabilities (TED2021-130707B-I00), funded by MCIN/AEI/10.13039/501100011033.

DeepKnowledge (PID2021-127777OB-C21) project funded by MCIN/AEI/10.13039/501100011033 and by FEDER Ixa group A type research group (IT1570-22) funded by the Basque Government

In addition, this work is supported by the Ministerio para la Transformación Digital y de la Función Pública and the Spanish Recovery, Transformation and Resilience Plan, funded by the European Union – NextGenerationEU, under the framework of the Desarrollo Modelos ALIA project. Further support has been provided by the projects CONSENSO (PID2021-122263OB-C21), MODERATES (TED2021-130145B-I00), AWARE (TED2021-131617B-I00) and SocialTox (PDC2022-133146-C21), all funded by MCIN/AEI/10.13039/501100011033 and the European Union through NextGenerationEU/PRTR.

References

AENOR, 2018. *Lectura Fácil. Pautas y recomendaciones para la elaboración de documentos.* AENOR Internacional SAU, Madrid.

Agüera-Marco, B. and I. Gonzalez-Dios. 2025. Lengclaro2023: A dataset of administrative texts in spanish with plain language adaptations. *arXiv preprint arXiv:2506.05927*.

Al Ajlouni, A. B., J. Li, and A. A. Mo'ataz. 2023. Towards a comprehensive metric for evaluating text simplification systems. In *2023 14th International Conference on Information and Communication Systems (ICICS)*, pages 1–6. IEEE, November.

Alva-Manchego, F., C. Scarton, and L. Specia. 2021. The (un) suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Bott, S. and H. Saggion. 2012. Automatic simplification of spanish text for e-accessibility. In *Computers Helping People with Special Needs: 13th International Conference, ICCHP 2012, Linz, Austria, July 11–13, 2012, Proceedings, Part I.* Springer Berlin Heidelberg, pages 527–534.

Calzolari, N., K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors. 2016. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, May.

European Language Resources Association (ELRA).

Espinosa-Zaragoza, I., J. Abreu-Salas, E. Lloret, P. M. Pozo, and M. Palomar. 2023. A review of research-based automatic text simplification tools. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 321–330, September.

Fernández-Huerta, J. 1959. Medidas sencillas de lecturabilidad. *Consigna*, (214):29–32.

González-Barba, J. Á., L. Chiruzzo, and S. M. Jiménez-Zafra. 2025. Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org*.

Instituto Cervantes. 2022. El español: una lengua viva. informe 2022. Informe técnico, Instituto Cervantes.

Papineni, K., S. Roukos, T. Ward, and W. J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, July.

Prodigioso Volcán. 2020. Plain language in administrative texts in spanish. ¿habla claro la administración? Technical report, Prodigioso Volcán.

Saggion, H., S. Štajner, D. Ferrés, K. C. Sheang, M. Shardlow, K. North, and M. Zampieri. 2022. Findings of the TSAR-2022 shared task on multilingual lexical simplification. In S. Štajner, H. Saggion, D. Ferrés, M. Shardlow, K. C. Sheang, K. North, M. Zampieri, and W. Xu, editors, *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 271–283, Abu Dhabi, United Arab Emirates (Virtual), December. Association for Computational Linguistics.

Specia, L., S. K. Jauhar, and R. Mihalcea. 2012. Semeval-2012 task 1: English lexical simplification. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 347–355.

Sulem, E., O. Abend, and A. Rappoport. 2018. Semantic structural evaluation for text simplification. *arXiv preprint arXiv:1810.05022*.

Tascón, M. and E. Montolío. 2020. *El derecho a entender: la comunicación clara, la mejor defensa de la ciudadanía*. Los libros de la Catarata.

Xu, W., C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Yaneva, V., K. North, P. Baldwin, L. A. Ha, S. Rezayi, Y. Zhou, S. Ray Choudhury, P. Harik, and B. Clauser. 2024. Findings from the first shared task on automated prediction of difficulty and response time for multiple-choice questions. In E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, and Z. Yuan, editors, *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 470–482, Mexico City, Mexico, June. Association for Computational Linguistics.

Zhang, T., V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. 2020. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.