Simple.Text: A Tool for the Automatic Transformation of Spanish Texts into Easy-to-Read

Beatriz Botella-Gil CENID University of Alicante
Isabel Espinosa-Zaragoza Department of English Philology University of Alicante
Paloma Moreda DLSI University of Alicante
Manuel Palomar CENID University of Alicante
{bea.botella, isabel.espinosa, moreda, mpalomar } @ua.es

Abstract

Automatic Text Simplification (ATS) has emerged as a key area of research within the field of Natural Language Processing, aiming to improve access to information by reducing the linguistic complexity of texts. Simplification can be applied at various levels—lexical, syntactic, semantic, and stylistic—and must be tailored to meet the needs of different target audiences, such as individuals with cognitive disabilities, low-literacy readers, or non-native speakers. This work introduces a tool that automatically adapts Spanish texts into Easy-to-Read format, enhancing comprehension for people with cognitive or reading difficulties. The proposal addresses the need for accessible, well-documented solutions aligned with official guidelines, reinforcing the potential of text simplification as a strategy for inclusion.

1 Introduction

Individuals with cognitive disabilities may experience significant limitations in intellectual functioning and often face challenges in adapting to everyday situations. Among the most common difficulties are deficits in understanding oral and written language, which can manifest in the misinterpretation of literal meanings, difficulties in following complex instructions, or confusion caused by idiomatic expressions, abstract concepts, rare vocabulary, and complex syntax.

In today's digital society—where information flows rapidly and online interaction has become a central component of daily life—these limitations translate into barriers that deepen social exclusion. The so-called "digital divide" particularly affects those who, due to limited digital skills or inaccessible textual content, are unable to fully participate in public life, access essential services, or exercise their rights on equal terms.

In this context, ensuring the accessibility of digital texts becomes a pressing priority. Just as physical barriers have been removed to facilitate mobility, it is essential to eliminate linguistic and cognitive barriers that hinder written comprehension. Natural Language Processing (NLP) technologies have reached a sufficient level of maturity to automate complex textual simplification processes, thereby opening new possibilities for enhancing the inclusion of individuals with cognitive disabilities.

Despite these advances, there remains a significant lack of adapted resources for the Spanish language. While English benefits from numerous corpora, tools, and well-established standards for easy-to-read content, Spanish lacks comparable resources, which restricts the development of technological solutions in this field. Moreover, current machine translation tools fail to meet the simplification standards required to ensure effective comprehension.

In response to this gap, the ClearText project - funded by the Government of Spain and the European Union (grant reference TED2021-130707B-I00) and carried out by the GPLSI research group at the University of Alicante — aims to research, design, and implement language technologies that facilitate the drafting of accessible content in Spanish, particularly for public sector institutions. With the goal of promoting the inclusion and empowerment of individuals with cognitive disabilities, the project encompasses multiple stages, from data collection and annotation to the development of automatic simplification tools.

One of the main outcomes of this initiative is the creation of a Spanish-language corpus adapted to various levels of simplification, including examples of Easy-to-Read texts and intermediate levels. This corpus enables the training, evaluation, and enhancement of automatic text simplification systems in Spanish.

As a result of this work, we developed Sim-

Proceedings of Recent Advances in Natural Language Processing,pages 204–209
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-024

ple.Text, a language technology tool that reformulates complex texts into accessible versions adapted to diverse cognitive needs. Beyond improving equity in information access, this research contributes to a more inclusive digital ecosystem, where adapted content benefits everyone and fosters a fairer and more respectful society.

2 State of the Art

Automatic Text Simplification (ATS) transforms complex texts into simpler versions while preserving meaning (Al-Thanyyan and Azmi, 2021). Approaches range from rule-based to data-driven and hybrid methods, targeting lexical, syntactic, semantic, and stylistic levels. Effectiveness depends on the linguistic phenomena addressed. Since users include children, language learners, or people with reading or cognitive difficulties, recent research emphasizes customization to meet diverse needs (Alva-Manchego et al., 2020; Scarton and Specia, 2018).

2.1 Available Corpora for Spanish

Martin et al. identified ten corpora developed for text simplification in Spanish, including well-known resources such as FIRST (Štajner and Saggion, 2013), IrekiaLF (Gonzalez-Dios and Alkorta, 2020), and CLARA-MED (Campillos-Llanos et al., 2022). Other resources are mentioned in the works by Bott and Saggion; Saggion et al.; Štajner et al. and Stajner et al., though they remain unnamed. In addition, two bilingual corpora—Newsela (Xu et al., 2015) and SIMPLETICO (Shardlow and Alva-Manchego, 2022)—offer aligned English-Spanish texts.

The review by Martin et al. highlights several limitations in current resources: most corpora are in English, and only a subset of EU languages are represented; there is a notable lack of domain-specific corpora, especially in critical areas like healthcare and public services; few resources are designed for individuals with cognitive disabilities; user involvement in corpus creation is rare; and, finally, over half of the corpora do not provide clear documentation on the simplification strategies employed.

Most available corpora focus on general content (e.g., news articles or Wikipedia entries). Notable exceptions include CLARA-MED and SIMPLETICO, both of which are focused on healthcare, and IrekiaLF, which targets administrative language. For comprehensive comparisons across dimensions such as domain, audience, linguistic alignment, size, and metadata (Martin et al., 2023).

2.2 Tools for Simplification in Spanish

Espinosa-Zaragoza et al. conducted an in-depth analysis of tools available for Spanish-language ATS. Their findings reveal five main challenges: limited language coverage, insufficient attention to multiple linguistic levels, lack of diverse NLP-based simplification options, a need for more audience-specific customization, and restricted public accessibility of existing tools.

Seven tools were identified for Spanish ATS: arText (da Cunha and Núñez, 2017), Simplext (Saggion et al., 2015a), DysWebxia (Rello et al., 2013), EASIER (Alarcón et al., 2021), LexSIS (Bott et al., 2012), NavegaFácil (Bautista et al., 2018), and Open Book (Barbu et al., 2015). As of the latest evaluation, only three tools—arText, EASIER, and Simplext—remain functional and accessible. arText assists users in identifying complex linguistic features; EASIER focuses on replacing difficult vocabulary with simpler alternatives; and Simplext enables sentence-level simplification, particularly useful for managing texts constrained by character limits.

3 Simple.Text tool

Simple.Text[1] is a NLP tool that automatically adapts Spanish texts into an Easy-to-Read format, facilitating comprehension for individuals with cognitive or reading difficulties.

This tool, developed as part of the Clear text research project, covers lexical, orthotypographical, syntactic, and semantic transformations. The Simple.Text interface features two text boxes: the first for inputting the original text and the second for displaying the transformed text, whether summarised or simplified, see Figure 1.

Two buttons allow users to choose between summarisation or Easy-to-Read (E2R) adaptation, with the selected option highlighted in a more intense color. Additionally, users can attach a file instead of copying the text directly. The transformed text and any generated glossary are displayed in an additional box, allowing for download in .txt format. The API is responsive, meaning it adapts to mobile phones, tablets, and laptops, and it supports various input formats, including PDF, TXT, and

[1] https://simpletext.demos.gplsi.es/

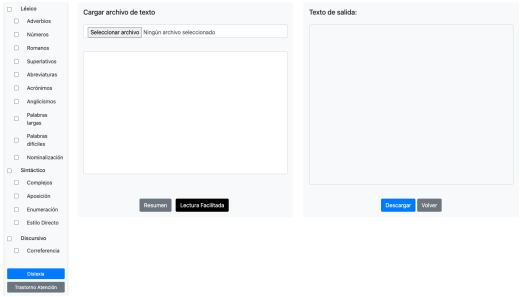

Figure 1: "Simple-Tool"

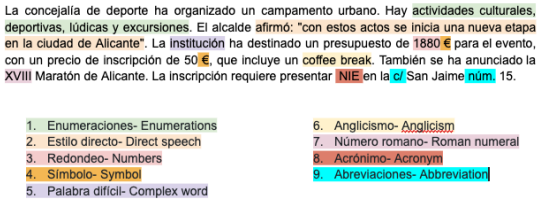

Figure 2: Example Transformation

Word. Implemented in Python 3 under a client-server architecture, Simple.Text detects specific linguistic phenomena using advanced algorithms and applies the defined transformations. It is the first tool designed to adapt texts according to user needs, addressing linguistic phenomena and considering cognitive disabilities, thereby improving accessibility and text comprehension in Spanish.

3.1 Functionality and Transformation Processes

The system applies three groups of transformations—lexical, syntactic, and discursive—following the official Easy-to-Read guidelines, see example in Appendix **??**, Figure 2.

3.1.1 Lexical Transformations

The system processes and transforms:

1. Adverbs ending in *-mente*
2. Numbers (including conversion from digits to words and vice versa, rounding, ordinal numbers, percentages, dates, times, and telephone numbers)
3. Roman numerals
4. Superlatives
5. Abbreviations and acronyms
6. Anglicisms
7. Long and complex words
8. Verbal nominalizations

3.1.2 Syntactic Transformations

The system processes and transforms:

- Complex connectors
- Appositions
- Enumerations
- Direct speech

Figure 2 presents an example illustrating the lexical, syntactic, and discursive transformations applied by the tool to generate the easy-to-read version of the text.

3.1.3 Discourse Transformations

The system processes and transforms:

- Coreference, to ensure textual coherence

Most of these transformations are implemented through predefined rules supported by ad hoc dictionaries specifically developed for the tool. For those transformations that require additional linguistic information (e.g., synonyms) necessary for simplification, a BERT-based model has been integrated using the Torch framework. In addition, common Natural Language Processing libraries have been used, such as spaCy [2], nltk [3], and Python modules including re [4], string [5], num2words [6], and roman [7].

3.1.4 Input and Output

- Input: The system accepts free text input through a text box on the web interface. Alternatively, users may upload documents in Word, PDF, or TXT format for content extraction, using the docx and pdfplumber libraries, respectively.
- Output: The transformed text is displayed in a dedicated output box on the web interface. Users also have the option to download the result in TXT format.

[2] https://spacy.io/
[3] https://www.nltk.org/
[4] https://docs.python.org/es/3/library/re.html
[5] https://docs.python.org/es/3/library/string.html
[6] https://pypi.org/project/num2words/
[7] https://pypi.org/project/roman/

3.1.5 User Interface

The web interface has been designed with a clean and functional layout, ensuring a clear and user-friendly experience. Key features include:

- Responsive Design: The interface adapts to various screen sizes and devices, offering a smooth user experience on both desktop and mobile platforms.
- Transformation Selection: On the left-hand side, users can select the specific transformations to be applied, see Figure 1. Customization can be carried out based on three criteria: (1) Linguistic level (lexical, syntactic, or discursive); (2) Obstacles associated with specific conditions, such as dyslexia or attention disorders; (3) Specific linguistic difficulties identified by the user.
- Text Input and File Upload: Users can either manually input text or upload files in Word, PDF, or TXT format.
- Text Processing and Output: The tool offers two main buttons: one for Easy-to-Read transformation and one for summarization.
  - E2R Button: Instantly applies the selected transformations. The result is displayed in a side-by-side output box, allowing users to compare the original and simplified versions. The simplified output is presented in a poem-like format and highlights words for which definitions or synonyms are provided using color coding. A Download button is also available to save the modified text in TXT format.
  - Summarization Button: Generates an abstractive summary of the main ideas contained in the text.
- Accessibility and Clarity: Visual elements have been incorporated to promote intuitive use of the tool. The visual and functional structure of the interface has been designed to ensure usability by individuals without technical expertise. Font sizes and color schemes comply with accessibility requirements. See Figure 3 illustrates a real example of the transformation from the original text to its easy-to-read version.

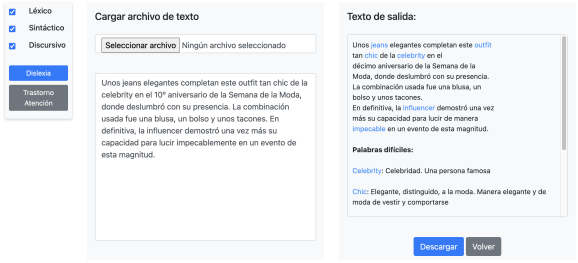

Figure 3: "Simple-Tool"-example

Despite the recent rise of large language models (LLMs), pretrained neural architectures, and sequence-to-sequence (seq2seq) models for text generation and simplification, rule-based systems remain a valid and valuable approach for Easy-to-Read (E2R) text adaptation—particularly in contexts where guidelines are explicitly defined. The subset of E2R rules addressed in this tool (e.g., numerical expressions, superlatives, among others) is sufficiently concrete and structured to be implemented through deterministic techniques such as regular expressions and controlled vocabularies. Although the outputs of rule-based systems may lack the fluency, flexibility, and human-like variation typical of LLM-generated text, they offer transparent, repeatable transformations that align well with strict compliance requirements. As such, this approach serves as a solid baseline and a first step toward automation in a domain traditionally reliant on manual, multi-stage human processes. Future work could explore hybrid strategies that combine the reliability of rule-based methods with the generative capabilities of LLMs, aiming to achieve both formal compliance and naturalness in E2R text adaptation.

Acknowledgments

This research work is part of the R&D&I projects: NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation with grant reference (CIPROM/2021/021) funded by the Generalitat Valenciana; 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.

References

Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Rocío Alarcón, Antonio Moreno, Jesús Vilares, and Manuel Vilares. 2021. Easier corpus: A lexical simplification resource for people with cognitive disabilities. *PLOS ONE*, 16(4):e0250416.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. Asset: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. *arXiv preprint arXiv:2005.00481*.

Catalina Barbu, Horacio Saggion, Radu Ion, and Dan Tufiș. 2015. Open book: A multilingual tool for text simplification. In *Proceedings of the 6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT 2015)*, pages 42–50. Association for Computational Linguistics.

Susana Bautista, Antonio Moreno, Jesús Vilares, and Manuel Vilares. 2018. Navegafácil: A web application for text simplification. In *Proceedings of the 8th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion (DSAI 2018)*, pages 123–130. ACM.

Stefan Bott, Luz Rello, Biljana Drndarevic, and Horacio Saggion. 2012. Can spanish be simpler? lexsis: Lexical simplification for spanish. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 357–374. The COLING 2012 Organizing Committee.

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

Leonardo Campillos-Llanos, Ana Rosa Terroba Reinares, Sofía Zakhir Puig, Ana Valverde Mateos, and Adrián Capllonch Carrión. 2022. Clara-med corpus.

Iria da Cunha and Juan Antonio Núñez. 2017. artext: A tool for assisting the writing of specialized texts. In *Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM 2017)*, pages 1–6. ACM.

Isabel Espinosa-Zaragoza, José Abreu-Salas, Paloma Moreda, and Manuel Palomar. 2023. Automatic text simplification for people with cognitive disabilities: Resource creation within the ClearText project. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 68–77, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Itziar Gonzalez-Dios and Jon Alkorta. 2020. Exploring the enrichment of Basque WordNet with a sentiment lexicon. In *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)*, pages 20–24, Marseille, France. The European Language Resources Association (ELRA).

Tania Josephine Martin, José Ignacio Abreu Salas, and Paloma Moreda Pozo. 2023. A review of parallel corpora for automatic text simplification. key challenges moving forward. In *Natural Language Processing and Information Systems*, pages 62–78. Springer Nature Switzerland.

Luz Rello, Ricardo Baeza-Yates, and Horacio Saggion. 2013. Dyswebxia 2.0!: More accessible text for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility (W4A 2013)*, pages 1–4. ACM.

Horacio Saggion, Estela Saquete, Paloma Moreda, Antonio Ferrández, and Manuel Palomar. 2015a. Automatic text simplification for spanish: A comparative evaluation of complementary modules. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2015)*, pages 360–372. Springer.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015b. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):1–36.

Carolina Scarton and Lucia Specia. 2018. Learning simplifications for specific target audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718.

Matthew Shardlow and Fernando Alva-Manchego. 2022. Simple tico-19: A dataset for joint translation and simplification of covid-19 texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3093–3102.

Sanja Štajner, Ruslan Mitkov, and Horacio Saggion. 2014. One step closer to automatic evaluation of text simplification systems. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 1–10.

Sanja Stajner, Sergiu Nisioi, and Ioana Hulpuș. 2020. CoCo: A tool for automatically assessing conceptual complexity of texts. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7179–7186, Marseille, France. European Language Resources Association.

Sanja Štajner and Horacio Saggion. 2013. Adapting text simplification decisions to different text genres and target users. *Sociedad Española para el Procesamiento del Lenguaje Natural*, 51:135–142.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.