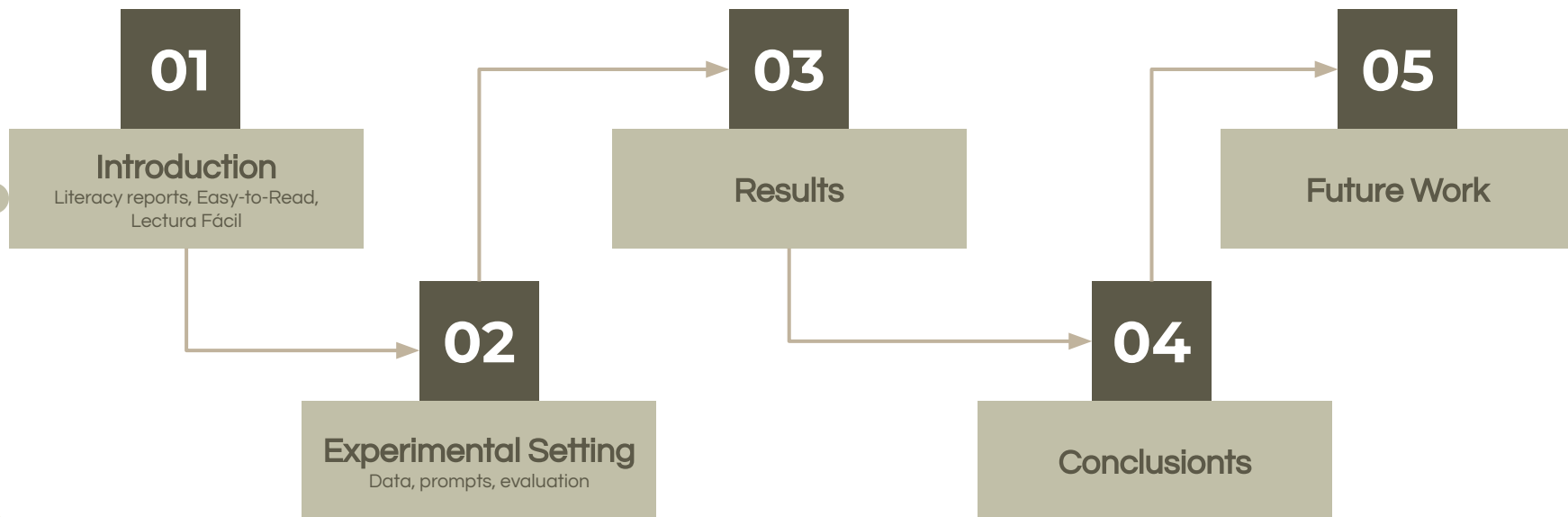


# A Preliminary Study of ChatGPT for Spanish E2R Text Adaptation

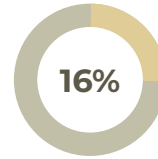
Margot Madina, Itziar Gonzalez-Dios & Melanie Siegel

# Agenda

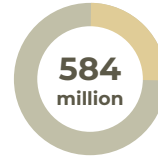


# Introduction

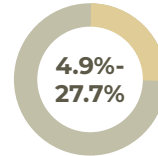
# Literacy



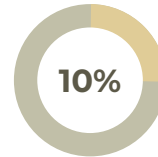
16% of the world's population (759 million adults), do not possess the basic literacy skills



584 million children struggle with reading skills



In OECD countries, adult literacy proficiency at the most basic levels varies between 4.9% and 27.7%

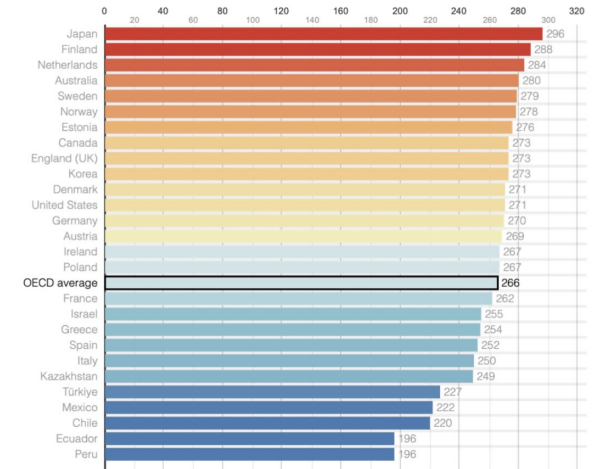


10% of graduates of OECD countries exhibit weak literacy abilities

# Spanish

Results obtained by the Programme for the International Assessment of Adult Competencies (PIAAC) showed that **Spanish-speaking countries scored very low for literacy skills**

According to the **Imserso** (Institute for the Elderly and Social Services in Spain) data, in Spain there are **283,256 people with a recognised intellectual disability of 33% or more**



Mean literacy score in the Survey of Adult Skills (PIAAC) (aged 16-65)

# Easy-to-Read (E2R) and Lectura Fácil (LF)



© European Easy-to-Read logo:  
Inclusion Europe

Easy-to-Read Language (E2R) is a controlled language variant of a standard language,

- with **reduced complexity**
- with the aim to **improve the readability and comprehensibility** of texts

One of its functions is to make content **accessible**, and to ensure participation for people with communication impairments.

Spanish E2R is called **Lectura Fácil (LF)**

# E2R target groups

People with **intellectual** or **developmental** disabilities

Groups with **cognitive difficulties**

People with **auditory** disabilities

People with **low literacy**

Migrants who **do not speak the language** of the country they live in

Children in need of **reading reinforcement**

# E2R document adaptation

Norma *UNE 153101:2018 EX de Lectura Fácil. Pautas y recomendaciones para la elaboración de documentos*

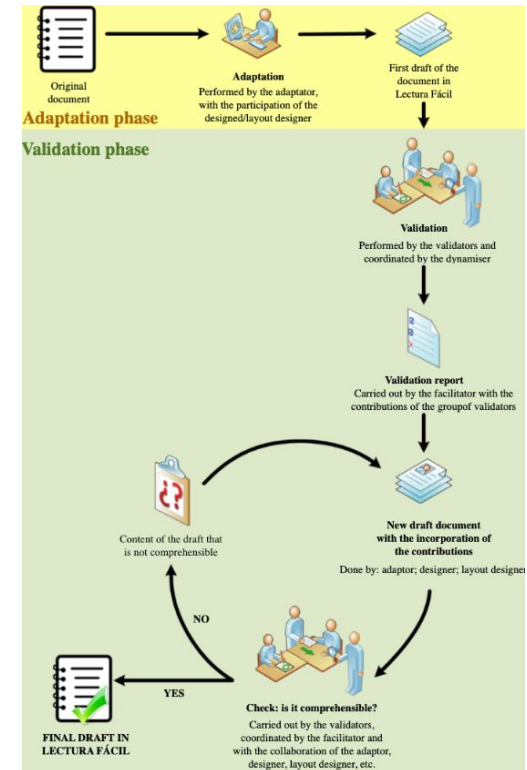
4 roles:

- Adaptor
- Designer and/or layout designer
- Facilitator
- Validators

Two-phase process:

**adaptation phase and validation phase**

They iteratively repeat until the text is correct





# Motivation and Aim

The creation of E2R texts is very costly, both in terms of time and money

Natural Language Processing (NLP) tools can help to some extent

As conversations around ChatGPT continue to permeate different spheres, it is natural to consider its potential as a solution to the problem at hand.

**Can ChatGPT automatically adapt standard Spanish texts into LF?**

Emphasis on the user perspective, underlining the importance of making technologies both accessible and reliable for the general public

# Experimental Setting

# Experimental Setting - data



10 texts from *Irekia* (the open-government communication channel of the Basque Government)

original version + LF version

Each of them deals with different topics, but all of them belong to the journalism genre and offer an account of events

# Experimental Setting - prompts

Design several and try three texts on ChatGPT-4

## PROMPT

Use 85 of the 130 rules in the UNE standard (exclude those related to images)

Prompt: Las normas de redacción de Lectura Fácil Según la Norma UNE 153101:2018 EX  
Lectura Fácil son las siguientes: (lista de normas)

Output: ...

Prompt: Adapta esta noticia a Lectura Fácil siguiendo las normas de redacción de Lectura  
Fácil Según la Norma UNE 153101:2018 EX Lectura Fácil: (noticia)

Output:

# Experimental Setting - evaluation



No standard to assess the adequacy of E2R or LF texts and Automatic Text Simplification (ATS) metrics are not entirely reliable

New three-folded evaluation method:

1. Automatic evaluation based on linguistic features with *MultiAzterTest*
2. Manual evaluation based on checklist
3. Evaluation with target group

# Results

# Results - automatic evaluation

*MultiAzterTest* examines 125 metrics of different levels of language

number of words, number of sentences, number of rare lexical words, number of subordinate clauses, and number of propositions

Compared the values of the original text, LF text and the text created by ChatGPT

## Results

The **ChatGPT** generated texts are generally **the shortest ones**

**LF** versions contain the **highest number of sentences** and also the **shortest ones**

**ChatGPT** generated texts contain the **lowest mean distinct rare lexical words**

The **subordinate clauses** per 1000 words is always the **lowest in the LF** versions

The mean of the number of **propositions per sentence** is always the **lowest in LF** versions

# Results - manual evaluation

List of certain characteristics (capabilities) to be met by simplified texts

Two linguists with experience in LF evaluated texts created by ChatGPT

Three new capabilities were proposed:

- (1) Format is easier to read
- (2) Line spacing is bigger
- (3) Phrases are not split into lines

## Results

- ✓ Grammar
- ✓ Simplicity
- ✗ Important information was lost
- ✗ Syntagmas divided into different lines

	Average of 10 texts
Evaluator 1	% 68,67
Evaluator 2	% 65,00
Average	% 66,83
Agreement (Cohen's kappa)	0,85



# Results - user evaluation

10 people with a disability percentage ranging between 33% and 79% read the text in LF and the texts created by ChatGPT

They were asked to retell what they read and answer some questions about the content

Finally, they were asked to pick the version that they preferred

## Results: They all preferred the LF version

Understand more words

More separation between lines

Shorter sentences

Better explained

More content and more explanation

Participant	1	2	3	4	5	6	7	8	9	10
Able to retell ChatGPT-4	yes	no	no	no	no	yes	no	yes	yes	no
Able to retell LF	yes	no	no	yes	no	yes	no	yes	yes	no
Able to answer questions ChatGPT-4	yes	yes	no	no	no	yes	no	yes	yes	no
Able to answer questions LF	yes	yes	no	no	no	yes	no	yes	yes	no
Preferred version	LF	LF	LF	LF	LF	LF	LF	LF	LF	LF

# Conclusions

# Conclusions - ChatGPT performance

- ✓ Lexical simplification (low presence of rare words)
- ✗ Creating examples or explanations
- ✗ Important information was lost
- ✗ Syntagmas divided into different lines

## Difficulties

Always getting the desired output

Difficulties in replication

Design the right prompt

## Output must be revised:

Check that all the information in the original version is in the E2R/LF version

Check that all E2R/LF rules are followed

**The output looks good at first glance, but if you analyze the texts well, there are gaps and errors**

# Conclusions - evaluation

Given the limitations of automatic metrics such as BLUE, ROUGE and SARI to evaluate E2R texts, we propose a new method

*MultiAztertTest* useful for analyzing specific features

Lack of capabilities in the checklist method → 3 proposed new ones

Fundamental importance of evaluation with target group

This method provides a lot of information, but it is expensive

# Future Work

# Future Work



Similar Experiments in other languages



Try other language models (Bloom, LLaMa...)



Propose a standard evaluation method to check compliance with E2R/LF standards

# References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification. *Computational Linguistics*, 47(4):861–889.

Kepa Bengoetxea and Itziar Gonzalez-Dios. 2021. Multiaztertest: A multilingual analyzer on multiple levels of language for readability assessment. arXiv preprint arXiv:2109.04870.

Stefan Bott, Horacio Saggion, and Simon Mille. 2012. Text Simplification Tools for Spanish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1665–1671.

Ursula Bredel and Christiane Maaß. 2016. *Leichte Sprache: Theoretische Grundlagen? Orientierung für die Praxis*. Bibliographisches Institut GmbH.

Oscar M Cumbicus-Pineda, Itziar Gonzalez-Dios, and Aitor Soroa. 2021. Linguistic Capabilities for a Checklist-based evaluation in Automatic Text Simplification. In *CTTS@SEPLN*.

Silvana Deilen, Sergio Hernández Garrido, Ekaterina Lapshinova-Koltunski, and Christiane Maaß. 2023. Using ChatGPT as a CAT tool in Easy Language translation. *TSAR 2023*.

Biljana Drndarević, Sanja Štajner, Stefan Bott, Susana Bautista, and Horacio Saggion. 2013. Automatic text simplification in Spanish: A comparative evaluation of complementing modules. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 488–500. Springer.

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. *TSAR 2023*.

Rudolf Flesch. 1943. *Marks of readable style; a study in adult education*. Teachers College Contributions to Education.

Rudolf Flesch. 1948. A readability formula in practice. *Elementary English*, 25(6):344–351.

Natalia Grabar and Horacio Saggion. 2022. Evaluation of Automatic Text Simplification: Where are we now, where should we go from here. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles*. Volume 1: conférence principale, pages 453–463.

Silvia Hansen-Schirra and Christiane Maaß. 2020. Easy language, plain language, easy language plus: perspectives on comprehensibility and stigmatisation. *Easy Language Research: Text and User Perspectives*, 2:17.

Imsero. 2022. Base Estatal de Datos de Personas con Valoración del Grado de Discapacidad.

Instituto Cervantes. 2022. El español: una lengua viva. Informe 2022.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, pages 150–157.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Christiane Maaß. 2019. Easy Language and Beyond: How to Maximize the Accessibility of Communication. Invited Plenary Speech at the Klaara 2019 Conference on Easy-to-Read Language Research (Helsinki, Finland. 19-20 September 2019).

Christiane Maaß and Sergio Hernández Garrido. 2020. Easy and Plain Language in Audiovisual Translation. *Easy language research: Text and user perspectives*, 2:131.

Jean Nitzke, Silvia Hansen-Schirra, Ann-Kathrin Habig, and Silke Gutermuth. 2022. Translating Subtitles into Easy Language: First Considerations and Empirical Investigations. *Translation, Mediation and Accessibility for Linguistic Minorities*, 128:127.

OECD. 2013. *OECD Skills Outlook 2013*

OECD. 2015. *OECD Skills Outlook 2015*



OECD. 2015. OECD Skills Outlook 2015

OpenAI. 2023. GPT-4 Technical Report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing 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.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 11054–11070. Curran Associates, Inc.

Enid Reichrath and Xavier Moonen. 2022. Assessing the effects of Language for all. *Nordic Journal of Linguistics*, 45(2):232–248.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François

Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Bleu is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

UNE. 2018. UNE 153101:2018 EX Lectura fácil.

*Pautas y Recomendaciones para la Elaboración de Documentos*. Madrid: Asociación Española de Normalización.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.