Text Simplification to Enable Democratic Participation. The iDEM project

Horacio Saggion

TALN - Natural Language Processing Research Group
Universitat Pompeu Fabra
Barcelona, Spain

SimpleText@CLEF-2025
10th September 2025

Universitat
Pompeu Fabra
Barcelona

Outline

1. Context
   - Why is Simplification important for democratic participation?
2. iDEM project [SOB^{+}24]
   - More inclusive deliberative democracy?
   - iDEM Consortium
   - iDEM Concept
3. Lexical Simplification
   - Definition & Approaches
   - LS Dataset for Catalan & Spanish
   - Human Assessment
   - Lexical Simplification Experiments
4. Segmenting long sentences to facilitate reading
5. iDEM App
6. Current Work

idem

Special Acknowledgements

Thanks to CLEF 2025 SimpleText Organizers, Jaap Kamps, Liana Ermakova.

Information Accessibility

- **Text** is still the most widespread source for transmission of **information** and human knowledge
- People who encounter difficulties making meaning out of **information in natural language** are a diverse group of individuals
- UN Convention demands **access to information as a fundamental human right**
- The Convention for the Rights of Persons with Disabilities includes **accessibility as a fundamental principle**.

Democracy and Accessibility

- Deliberative democracy requires **understandable language** for all stakeholders.
- However, many groups are **excluded from democratic deliberation** due to the **complex language** used by governments, political parties, and democratic institutions.
- People with disabilities, migrants, people with low literacy, the elderly, might be excluded from democratic deliberation

Topic of debate: Climate Change

In common usage, *climate change* describes global warming — the ongoing *increase in global average temperature* — and its effects on *Earth's climate system*. The current rise in global average temperature is primarily caused by *humans burning fossil fuels* since the *Industrial Revolution*.	*Climate change* is the rapid change in the climate of the Earth, caused by human activity.

- Easy to read (E2R) guidelines support readability, especially for intellectual disabilities.
- Limited empirical research on the benefits of E2R, often rules and recommendations are contradictory.
- Text simplification aims at automating the production of E2R.

The iDEM Project

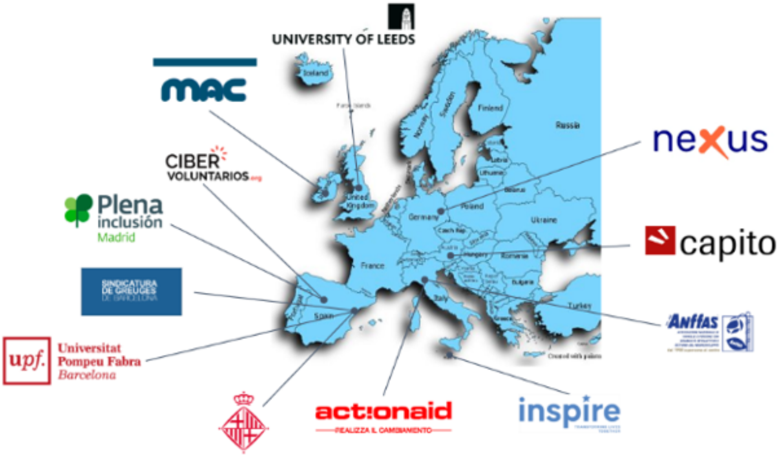
- Call: HORIZON-CL2-2023-DEMOCRACY-01-07 - Intersectionality and equality in deliberative and participatory democratic spaces +WIDERA Call
- iDEM: Innovative and Inclusive Democratic Spaces for Deliberation and articipation
- Start Date: 1/1/2024
- End Date: 31/12/2026
- Runtime: 36M
- Partners: 11 (EU) + 1 (UK)
- Coordination: Universitat Pompeu Fabra

idem

Our Objectives

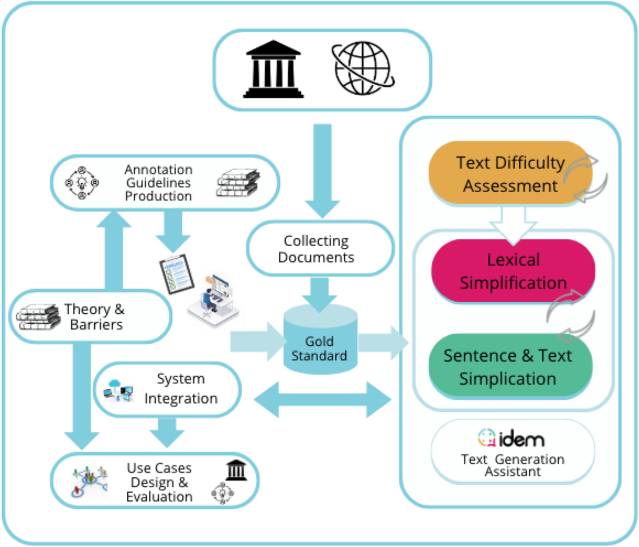
- Address barriers in deliberative and participatory democratic practices that limit the participation of people with limited skills in reading, writing or understanding the complex language of democracy.
- Develop models and tools based on NLP to make democratic spaces more inclusive and accessible.
- Investigate artificial intelligence and natural language processing (NLP) methods as enablers in addressing political inequality.
- Focus on three case studies in Catalan, Italian, and Spanish to evaluate text simplification in democratic deliberation and one use case in English to test text comprehension.

Consortium



idem

iDEM Concept


Annotation Guidelines Production
Collecting Documents
Theory & Barriers
Gold Standard
System Integration
Use Cases Design & Evaluation
Text Difficulty Assessment
Lexical Simplification
Sentence & Text Simplification
idem
Text Generation Assistant

Lexical Simplification [PS17]

Lexical Simplification (LS) is the task of replacing difficult words for easy-to-read equivalents.

John composed these verses in 1995.
John wrote the poem in 1995.

Predicting the lexical complexity of words is also an important task in LS: (i) **discrete** (Complex Word Identification - CWI) or (ii) **continuous** (Lexical Complex. Prediction - LCP).

Lexical Simplification

- Early count-based word-vectors and available dictionaries for modelling word semantics and synonymy [BRDS12]
- Word embedding learned from huge text collection help model semantic similarity [G5]
- Large-scale language models such as BERT and its variations have been applied to predict substitution candidates for complex words (in context) with LS-BERT [QLZ+20] as an example.
- Evaluation frameworks show that LLMs help develop best performing models for the lexical simplification [SŠF+22].
- Techniques such as "prompting" are used to condition the LLMs to produce a simplification or to suggest alternative words.
- However, these models underperform when simplifying low-resourced languages.

idem

Lexical Simplification

- Lexical simplification datasets are relevant for evaluating lexical simplification solutions and/or for training, fine-tuning, or prompting models.
- Lexical simplification datasets are made up of sentences, a target word to simplify, and a list of simpler human-provided simplifications.
- Languages such as Catalan or Spanish lag behind when dataset availability is of concern.

Creating Lexical Simplification Resources for Catalan [SBS+24]

- TeCla corpus of Catalan news (education section)
- Contexts selected heuristically using word frequency threshold
- Two annotators examined content words in context to decide which were simplification candidates
- Three target words selected for each context by agreement
- Lexical complexity values (1-5) and up-to three lexical simplification substitutes were requested
    - Prolific platform
    - 10 subjects per target word
    - knowledge of Catalan

Creating Lexical Simplification Resources for Spanish

- A text simplification corpus in the area of financial education in Spanish (from South America).
- Contexts from simplification alignments $< complex, simple >$ sentence pairs selected heuristically by detecting possible lexical substitutions (*derrotismo* → *pesimismo*).
- Two annotators examined candidate words to decide if they were simplification candidates: one target word selected by agreement and two by random sampling
- Lexical complexity values (1-5) and up-to three lexical simplification substitutes were requested
  - University students
  - 10 students per target word
  - Spanish (Costa Rican variety) as first language

idem

Creating Lexical Simplification Resources for Catalan & Spanish

- Catalan dataset contains 160 contexts with 475 target words.
- Spanish dataset contains 210 contexts with 625 target words.
- Our Catalan & Spanish data is part of MLSP 2024 dataset

Spanish	*Pero uno no puede dejar que el* **derrotismo** *lo detenga e impida que haga un presupuesto*
Complex.	0.7
Subs.	desánimo (4), pesimismo (4), abatimiento (3), derrotismo (2), ... negativismo (1)
Catalan	*No poden tocar-se ni abraçar-se, no hi ha joc col·lectiu, s'ha* **sectoritzat** *el pati i la desinfecció per allà on passen és la nova rutina a l'escola.*
Complex.	0.6
Subs.	dividit (5), segmentat (2), fragmentat (1), ... separat en zones (1)

Data Quality Assessment

- How good are the proposed substitutions?

  Q1 Are they words/expressions in the language? (valid word)

  Q2 Are they equivalent to the target word? (equivalent word)

  Q3 If so, do they fit in the context of the target word? (fit in context)

  Q4 Are they simpler than the target word? (simplicity)

- Answering the questions
  - Manual assessment of a sample of data points
  - Data-set organized in 5 lexical complexity categories [0-0.20]...(0.80-1.00]
  - From each category 10 sentences randomly sampled: 150 Spanish and 120 Catalan
  - Three top-most substitutes selected for answering the above questions by two native speakers of Spanish with Catalan as L2

Data Quality Assessment

Spanish

LC Level	validity		equivalence		in-context fit		simplicity		
	V	NV	E	NE	F	NF	S	EQ	C
1	100%	0%	87%	13%	100%	0%	35%	50%	15%
2	100%	0%	87%	13%	81%	19%	42%	50%	8%
3	100%	0%	63%	37%	79%	21%	42%	58%	0%
4	100%	0%	77%	23%	74%	26%	65%	35%	0%
5	100%	0%	73%	27%	86%	14%	59%	41%	0%
ALL	100%	0%	77%	23%	84%	16%	48%	46%	6%

Catalan

LC Level	validity		equivalence		in-context fit		simplicity		
	V	NV	E	NE	F	NF	S	EQ	C
1	100%	0%	77%	23%	74%	26%	26%	61%	13%
2	97%	3%	93%	7%	70%	30%	44%	56%	0%
3	100%	0%	70%	30%	76%	24%	62%	38%	0%
4	93%	7%	71%	29%	75%	25%	45%	45%	10%
5	100%	0%	67%	33%	100%	0%	50%	50%	0%
ALL	97%	3%	78%	22%	58%	42%	44%	50%	6%

Lexical Simplification Approach [KBE+25]

- **Goal:** Replace complex words with simpler alternatives.
- **Model:** Salamandra-7B (decoder-only LLM for Romance languages).
- Avoids commercial LLM constraints (privacy, cost).
- **Relatively Simple Approach:**
  - Few-shot prompting (0, 2, and 4-shot) using MLSP 2024 trial data.
  - Compared against MLSP baseline and winning systems (e.g., GPT-4).
- **Prompt:**

  "Given the context and the specified target word in {*LANGUAGE*}*, answer 10 simpler alternative words..."*

Lexical Simplification using the MLSP Dataset [SAMBN$^+$24]

- We carried out experiments on the four languages of the project: Catalan, English, Italian, and Spanish
- Catalan dataset contains 160 contexts with 475 target words.
- Spanish dataset contains 210 contexts with 625 target words.
- English dataset contains 200 contexts with 600 target words.
- Italian dataset contains 200 contexts with 600 target words.
- We measure accuracy which expresses the percentage of right solutions produced out of all given solutions (we measure ACC@1@1)

Lexical Simplification: Accuracy Results

Language	0-shot	2-shot	4-shot	Baseline
English	0.1280	0.4017	**0.4035**	0.3877
Spanish	0.0286	**0.3541**	0.3608	0.3254
Catalan	0.0426	**0.2292**	0.2022	0.1977
Italian	0.0350	**0.3596**	0.3315	0.2964

- **Salamandra-7B outperformed** the strong MLSP baseline in all Romance languages (few-shot).
- **Best results** observed with 2-shot prompts.
- Shows **open models** can approach state-of-the-art with simple prompting.

idem

Teacher/Student Approach to Lexical Simplification [HBS25] I

- **Context:** Open-access LLMs require substantial computational resources.
- **Goal:** Propose efficient text simplification with smaller models at comparable performance.
- **Approach:** In-context learning and knowledge distillation.
  - Prompting an LLM (Gemma 2 9B) to provide lexical simplifications for around 30K examples.
  - Fine-tuning smaller LMs (Qwen 1.5B, and Llama 1B) with context, target word, and lexical simplification.
- **Languages:** Multilingual, tested on MLSP 2024.

Teacher/Student Approach to Lexical Simplification [HBS25] II

- Automatic evaluation shows that the teacher model has a very strong performance across languages and at the same time the fine-tuned student models, in general, outperform few-shot approaches.
- Manual evaluation reveals that systems may invent words, change original meaning, or produce more complex output.

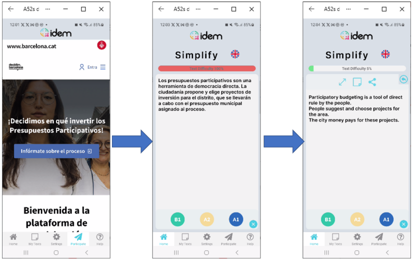
Segmenting long sentences to facilitate reading [KBE^{+}25]

	Sentence	Easy to Read Segmentation
EN	The way this sentence is cut is easy to read.	The way this sentence is cut is easy to read.
ES	Validar es comprobar si un documento es fácil de comprender.	Validar es comprobar si un documento es fácil de comprender.
IT	Nel mese di marzo 2006 è uscito il nuovo disco di CapaRezza ...	Nel mese di marzo 2006 è uscito il nuovo disco di CapaRezza Habemus Capa.

- We acquired a corpus of Spanish E2R segmentation data [CEP24]
- Design a Decision Tree algorithm to decide segmentation points (e.g. scikit-learn)
- Learning instances are tokens and features to represent them are PoS tags, token form, named entity tags, position, etc. with support of spaCy NLP library
- Performance at token level achieves F1 of 0.26 for segmenting at token and 0.91 for non-segmenting
- Experiments on Italian [CS25] using similar approach achieve 0.38 and 0.90.

iDEM App and Deployment

- Complexity detection and simplification
- Customizable
- Varied inputs: text, speech, images, PDFs
- Interaction with deliberative platform



¡Decidimos en qué invertir los Presupuestos Participativos!

Infórmate sobre el proceso

Bienvenida a la plataforma de

Los presupuestos participativos son una herramienta de democracia directa. La ciudadanía propone y elige proyectos de inversión para el distrito, que se llevarán a cabo con el presupuesto municipal asignado al proceso.

Participatory budgeting is a tool of direct rule by the people.
People suggest and choose projects for the area.
The city money pays for these projects.

Current Work

- Our iDEM project aims at enabling more inclusive democratic participation.
- Few-shot lexical simplification with **open models** (Salamandra-7B) exceeded strong baselines.
- Extend simplification pipeline to full sentence transformations.
- Fine-tune Salamandra LLMs on domain-specific E2R data.
- Integrate simplification modules into the iDEM app.

Thank you!

www.idemproject.eu

upf. Universitat Pompeu Fabra Barcelona
UNIVERSITY OF LEEDS
capito
actionaid REALIZZA IL CAMBIAMENTO
Ajuntament de Barcelona
SINDICATURA DE GREUGES DE BARCELONA
mac
Anffas NAZIONALE
CIBER VOLUNTARIOS.org
nexus
Plena inclusión Madrid
inspire

This document is part of a project that has received funding from the European Union´s Horizon Europe research and innovation program under the Grant Agreement No. 101132431 .Views and opinions expressed are however those of the author(s) only and do necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

References I

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

Jesús Calleja, Thierry Etchegoyhen, and David Ponce, *Automating Easy Read Text Segmentation*, Findings of the Association for Computational Linguistics: EMNLP 2024 (Miami, Florida, USA) (Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, eds.), Association for Computational Linguistics, November 2024, pp. 11876–11894.

References II

- Marta Cozzini and Horacio Saggion, *Segmenting Italian Sentences for Easy Reading*, Proceedings of the Eleventh Italian Conference on Computational Linguistics (in press) (Cagliari, Italy), 2025.

- Goran Glavaš and Sanja Štajner, *Simplifying Lexical Simplification: Do We Need Simplified Corpora?*, Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (Beijing, China) (Chengqing Zong and Michael Strube, eds.), Association for Computational Linguistics, July 2015, pp. 63–68.

References III

Akio Hayakawa, Stefan Bott, and Horacio Saggion, *Towards Trustworthy Lexical Simplification: Exploring Safety and Efficiency with Small LLMs*, Proceedings of the International Natural Language Generation Conference (INLG 2025) (in press) (Hanoi, Vietnam), 2025.

Nouran Khallaf, Stefan Bott, Carlo Eugeni, John O'Flaherty, Serge Sharoff, and Horacio Saggion, *Democracy made easy: Simplifying complex topics to enable democratic participation*, Proceedings of the 1st Workshop on Artificial Intelligence and Easy and Plain Language in Institutional Contexts (AI & EL/PL) (Geneva, Switzerland) (María Isabel Rivas Ginel, Patrick Cadwell, Paolo Canavese, Silvia Hansen-Schirra, Martin Kappus, Anna Matamala, and Will Noonan, eds.), European Association for Machine Translation, June 2025, pp. 108–124.

References IV

- Gustavo Paetzold and Lucia Specia, *A Survey on Lexical Simplification*, Journal of Artificial Intelligence Research **60** (2017), 549–593.

- Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu, *LSBert: Lexical Simplification Based on BERT*, IEEE/ACM Transactions on Audio, Speech, and Language Processing (2020), 3064–3076.

- Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar,

References V

Sanja Štajner, Marcos Zampieri, and Horacio Saggion, *The BEA 2024 shared task on the multilingual lexical simplification pipeline*, Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024) (Mexico City, Mexico) (Ekaterina Kochmar, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anaïs Tack, Victoria Yaneva, and Zheng Yuan, eds.), Association for Computational Linguistics, June 2024, pp. 571–589.

Horacio Saggion, Stefan Bott, Sandra Szasz, Nelson Pérez, Saúl Calderón, and Martín Solís, *Lexical complexity prediction and lexical simplification for Catalan and Spanish: Resource creation, quality assessment, and ethical considerations*, Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024) (Miami, Florida,

References VI

USA) (Matthew Shardlow, Horacio Saggion, Fernando Alva-Manchego, Marcos Zampieri, Kai North, Sanja Štajner, and Regina Stodden, eds.), Association for Computational Linguistics, November 2024, pp. 82–94.

Horacio Saggion, John O'Flaherty, Thomas Blanchet, Serge Sharoff, Silvia Sanfilippo, Lian Muñoz, Martin Gollegger, Almudena Rascón, José L. Martí, Sandra Szasz, Stefan Bott, and Volkan Sayman, *Making democratic deliberation and participation more accessible: The idem project.*, SEPLN (Projects and Demonstrations) (Alba Bonet-Jover, Robiert Sepúlveda-Torres, Rafael Muñoz Guillena, Eugenio Martínez-Cámara, Elena Lloret Pastor, Álvaro Rodrigo-Yuste, and Aitziber Atutxa, eds.), CEUR Workshop Proceedings, vol. 3729, CEUR-WS.org, 2024, pp. 71–76.

References VII

Horacio Saggion, Sanja Štajner, Daniel Ferrés, Kim Cheng Sheang, Matthew Shardlow, Kai North, and Marcos Zampieri, *Findings of the TSAR-2022 shared task on multilingual lexical simplification*, Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022) (Abu Dhabi, United Arab Emirates (Virtual)) (Sanja Štajner, Horacio Saggion, Daniel Ferrés, Matthew Shardlow, Kim Cheng Sheang, Kai North, Marcos Zampieri, and Wei Xu, eds.), Association for Computational Linguistics, December 2022, pp. 271–283.