The iDEM Project: Addressing Linguistic Barriers in Deliberative Processes

H. Saggion, S. Bott, J. Martí, S. Szasz, S. Sharoff, J. O'Flaherty, T. Blanchet V. Sayman, M. Gollegger, A. Rascón, S. Sanfilippo, L. Muñoz
UPF, UoL, MAC
NEXUS, CAPITO, PMI, ANFFAS, CIBER

Abstract

Democratic processes should be fully inclusive. However, it is well documented that people with limited language skills – such as people with cognitive disabilities struggle with democratic deliberations, and this is despite the advocacy work of organizations that promote human rights. The iDEM project aims to remove the barriers that limit the participation of citizens who are marginalized from democratic processes because of the inherent difficulties associated with documents, debates, and discourses used and produced in these settings. In addition to a theoretical investigation of the limitations of current marginalisation from deliberative processes due to a lack of language skills, the project adopts a user-centred approach for designing more accessible and inclusive deliberative spaces thanks to the use of natural language processing technology to make information in democracy easier to produce, read, and understand, thus allowing a fairer participation. Our project will develop use cases in Catalan, Spanish, and Italian languages with a diverse group of citizens and deliberative processes.

Keywords: Democracy, Deliberation, Natural Language Processing, Text Simplification, Text Generation, Readability

1. Introduction: Language and Democracy

Deliberative and participatory processes (Beauvais, 2018) currently lack full legitimacy due to the exclusion and marginalisation of several vulnerable communities from democratic spaces (Gherghina et al., 2021). Literacy is fundamental to human development as it enables people to contribute to their communities and to society. People who encounter difficulties making meaning out of content are a diverse group of individuals with varying ranges of reading, writing, and understanding abilities. This part of the population includes, for instance, those with low levels of literacy, intellectual disabilities, dyslexia, aphasia, temporary impairments, or limited language skills (e.g. second-language learners, immigrants, and displaced populations).

The challenges of including people with disabilities in deliberative or participatory processes are even bigger than those of including them in representative processes. In the 2019 European Parliament elections many people were excluded due to barriers such as insufficiently accessible information about candidates and debates. But in the case of deliberative processes, the number of people excluded due to linguistic barriers even exceed this, if one considers that 1% of the human population is affected by an intellectual disability [1].

[1] https://www.psychiatry.org/patients-families/intellectual-disability/what-is-intellectual-disability

2. Overview of the iDEM Project

The iDEM Project, which started in January 2024, has received funding from the Horizon Europe under call *HORIZON-CL2-2023-DEMOCRACY-01-07* in the area of intersectionality and equality in deliberative and participatory democratic spaces. iDEM aims at making information more accessible and inclusive in the context of democratic discourse and in particular in deliberative and participatory processes. In order to address this challenge, we will investigate using a theoretical approach current marginalisation from deliberative processes of diverse under-represented groups due to language skills in order to understand what are the linguistic barriers which hamper their participation. We will then create, following well-defined guidelines and schemata (e.g. (García Muñoz, 2012)), parallel domain-specific annotated datasets for simplification in Catalan, Spanish, and Italian. Data will be gathered from past participatory processes, democratic institutions, and the Web. The annotated corpus will be used to fine-tune natural language processing models to automatically identify and classify the source of text complexity and simplify them accordingly. We will also assist our users with tools for the generation of coherent discourses, adopting carefully fine-tuned Large Language Models.

By working with associations for people with intellectual disabilities, iDEM will adopt a user-centred approach in use case design and corpus creation to ensure maximum impact in the community thus contributing to making democracy more accessible and inclusive. An iDEM innovative service will be

created to deploy the developed language technologies: it will be open-source, and well-documented. It will provide an architecture for communication, as well as the open iDEM API.

3. Making Information More Accessible

In order to face the barriers imposed by the way information is written, and from a methodological viewpoint, easy-to-read guidelines (Orero and Matamala, 2018; Matamala and García Muñoz, 2021) can be followed when preparing documentation for specific targets. Easy-to-read is an inclusive (i.e. taking a wide range of people into account), user-centred method: that is, it considers the inclusion and involvement of end-users throughout the content creation process and validation. Examples of easy-to-read texts in a democracy context and their non-adapted counterparts are shown in Table 1. Example (a)[2] [3] shows information about the political program of the party *Junts* for the 2023 Barcelona City Council elections, where the easy-to-read version shows how information is summarized in one short and simple sentence. In example (b)[4] [5], information from the Labour Party for the 2019 UK General Elections is presented in non-adapted and adapted form; the adapted form is heavily compressed containing just key information. The profile of an Italian political candidate is presented in example (c)[6][7], where the easy-to-read version preserves most information but presenting it in two separate short sentences. Instead in example (d)[8] [9], a short original statement from the PSOE political party is enriched in the easy-to-read version adding a definition of a complex term (i.e. Aulas Matinales)[10]. Thus, as shown in these examples easy-to-read texts can be quite complicated to produce since they require high levels of paraphrasing, summarizing, lexical substitution, splitting, syntactic transformations, and proper layout for effective communication. From a computational viewpoint, transforming complex texts into easy-to-read and understand ones, has been addressed in the field of *automatic text simplification* (Saggion, 2017) which has usually concentrated on two different tasks: *lexical simplification* and *syntactic simplification*, each addressing different sub-problems. *Lexical simplification* will attempt to modify the vocabulary (e.g. target complex words) by choosing substitutes which are more appropriate for the reader (Shardlow, 2014). Changing words in context is not an easy task because it may alter the meaning of the original text. *Syntactic simplification* will transform complex sentences into more readable or understandable equivalents. For example, relative or subordinate clauses or passive constructions, which may be more difficult to read, could be transformed into simpler sentences or into active form. Although some non-English research has been produced in this area, it is fair to say most research and resources have been generated for the English language. Our project will contribute with solutions also for Catalan, Italian, and Spanish. Needless to say that no language resources exist in the field of democracy and deliberation to study how to adapt simplification systems, our project aims to make a contribution in that direction.

4. Text Simplification and Generation

Early research on text simplification applied rule-based methods for syntactic simplification and corpus-based unsupervised techniques for lexical simplifications (Saggion et al., 2015). Where parallel complex-simple sentences are available (e.g. English Wikipedia and Simple English Wikipedia pairs) text simplification can be addressed as monolingual Statistical Machine Translation (SMT) (Coster and Kauchak, 2011). Recent research has shown that adding control tokens during training (Alva-Manchego et al., 2017) improves the performance of sentence simplification models, achieving state-of-the-art in English and Spanish (Sheang and Saggion, 2021) using large pre-trained language models (Raffel et al., 2020). Concerning lexical simplification, several past approaches used traditional raw count word-vectors and available

[2] https://www.elnacional.cat/es/elecciones/municipales-2023/programa-electoral-junts-barcelona-2023-xavier-trias-elecciones-municipales_1026717_102.html

[3] https://repositori.lecturafacil.net/ca/node/909

[4] https://labour.org.uk/wp-content/uploads/2019/11/Real-Change-Labour-Manifesto-2019.pdf

[5] https://labour.org.uk/wp-content/uploads/2019/12/12981_19-Easy-Read-manifesto.pdf

[6] https://it.wikipedia.org/wiki/Renato_Soru

[7] https://informazionefacile.it/elezioni-regionali-sardegna-in-breve/

[8] https://www.psoevaldepenas.es/programa-electoral-2019.php

[9] https://www.psoevaldepenas.es/ARCHIVO/documentos/documentos/programa-lectura-facil-2023.pdf

[10] In this case we have also noticed that the easy-to-read version contains a further explanation of tha acronym *AMPAS* which is defined as: Asociación de madres y padres de los alumnos de un centro educativo.

Ex.	Context	Original version	Easy-to-read version
(a)	Political Program (Catalan)	Augmentarem la ràtio d'agents de la Guàrdia Urbana fins a 2,5 agents per cada mil habitants, davant dels 2 agents per cada mil habitants, arribant als 4.000 agents el 2027.	Ampliar la Guàrdia Urbana amb 4.000 agents.
(b)	Political Program (English)	Invest an additional £400 million in our diplomatic capacity to secure Britain's role as a country that promotes peace, delivers ambitious global climate agreements and works through international organisations to secure political settlements to critical issues	Spend a lot of money (£400 billion) to make the economy work better for every part of Britain and the environment
(c)	Political Candidates (Italian)	Renato Soru (Sanluri, 6 agosto 1957) è un imprenditore, politico e dirigente d'azienda italiano, fondatore di Tiscali e della disciolta Andala UMTS, oltreché presidente della Regione Sardegna dal 2004 al 2009.	Renato Soru è un politico e imprenditore noto per aver fondato la società di telecomunicazioni Tiscali. Nel 2004 è stato eletto presidente della Regione Sardegna con la coalizione di centrosinistra.
(d)	Political Program (Spanish)	Apoyaremos a las AMPAS en el mantenimiento de las Aulas Matinales	Ayudaremos a las AMPAS a mantener las aulas matinales. Las aulas matinales son actividades que se hacen antes de que los niños y niñas entren al colegio para que sus padres puedan ir a trabajar.

Table 1: Example of complex and simple sentences in the context of democracy in four different languages (Catalan, English, Italian, and Spanish).

dictionaries for modelling word semantics and to select simple word replacement for complex words (Bott et al., 2012); nowadays, large-scale language models such as BERT and its variations have been applied to predict substitution candidates for complex words(Qiang et al., 2020; Sheang and Saggion, 2023). In this context, a Masked Language Model predicts substitute words which are ranked for simplicity. These models, however, have clearly neglected bias-related aspects when proposing simplifications. Generative AI, such as ChatGPT – based on Generative Pre-trained Transformers (GPT) (Brown, 2020), shows the possibilities of In-Context Learning. However, at the moment they have been outperformed by specialised training (Caines, 2023).

5. Expected Results

To be able to make informed decisions and actively get involved in society, people need to understand written information, especially information to participate in democratic processes which affect their lives. Unfortunately, information used by policymakers and democratic institutions requires high literacy levels. Although several organizations offer accessible information in many countries, they depend on well-trained human editors who can only produce a handful of documents at a time and this at a high cost. The iDEM project will help improve the accessibility to information in democracy by creating novel participatory spaces in which text simplification technology a natural language generation will be used to adapt texts or generate new ones thus facilitating participation in democratic processes. Our project will create curated datasets for text simplification in three languages and develop models for text simplification and generation in the context of democratic discourse.

6. Acknowledgments

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 (iDEM Project). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. UOL was funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (grant number 10103529)

7. Bibliographical References

Fernando Alva-Manchego, Joachim Bingel, Gustavo Paetzold, Carolina Scarton, and Lucia Specia. 2017. Learning how to simplify from explicit labeling of complex-simplified text pairs. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 295–305, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

Edana Beauvais. 2018. 144Deliberation and Equal-

ity. In *The Oxford Handbook of Deliberative Democracy*. Oxford University Press.

S. Bott, L. Rello, B. Drndarevic, and H. Saggion. 2012. Can spanish be simpler? LexSiS: Lexical simplification for spanish. In *24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers*, pages 357–374.

Tom Brown. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Andrew Caines. 2023. On the application of large language models for language teaching and assessment technology. In *AIED2023 workshop: Empowering Education with LLMs – the Next-Gen Interface and Content Generation*.

Will Coster and David Kauchak. 2011. Learning to simplify sentences using Wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 1–9, Portland, Oregon. Association for Computational Linguistics.

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019. EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3393–3402, Florence, Italy. Association for Computational Linguistics.

Óscar García Muñoz. 2012. *Léctura fácil - Métodos de redacción y evaluación*. Real Patronato sobre Discapacidad.

Sergiu Gherghina, Monika Mokre, and Sergiu Miscoiu. 2021. Deliberative democracy, underrepresented groups and inclusiveness in europe. *Innovation: The European Journal of Social Science Research*, 34:1–5.

Anna Matamala and Óscar García Muñoz. 2021. *Easy Language in Spain*, pages 493–526.

Pilar Orero and Anna Matamala. 2018. Standardising accessibility: transferring knowledge to society. *Journal of Audiovisual Translation*, 1:139–154.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Horacio Saggion. 2017. *Automatic Text Simplification*, volume 10 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015. Making It Simplext: Implementation and Evaluation of a Text Simplification System for Spanish. *ACM Transactions on Accessible Computing*, 6(4):1–36.

Matthew Shardlow. 2014. A Survey of Automated Text Simplification. *International Journal of Advanced Computer Science and Applications*, 4(1).

Kim Cheng Sheang and Horacio Saggion. 2021. Controllable sentence simplification with a unified text-to-text transfer transformer. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 341–352, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Kim Cheng Sheang and Horacio Saggion. 2023. Multilingual controllable transformer-based lexical simplification. *Proces. del Leng. Natural*, 71:109–123.

S. Štajner, H. Béchara, and H. Saggion. 2015. A deeper exploration of the standard PB-SMT approach to text simplification and its evaluation. In *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*, volume 2, pages 823–828.

Sanja Stajner. 2021. Automatic text simplification for social good: Progress and challenges. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.