Towards An Artificial Intelligence (AI)-based Framework to Automatically Adapt Short Stories in Spanish into Easier and Accessible Versions

Isam Diab[1]

[1]*Ontology Engineering Group (OEG), Universidad Politécnica de Madrid (UPM), Madrid, Spain*

Abstract

The Easy-to-Read (E2R) Methodology was created with the aim of presenting clear and easily understood documents to improve the daily life of people who present reading comprehension difficulties, such as persons with cognitive disabilities. To do that, the methodology provides a set of guidelines regarding both writing and layout aspects. However, the E2R guidelines are applied manually to create easy-to-read text materials, which demands considerable resources and effort. To help in such a manual process, and considering that cultural materials, in general, and literary materials, in particular, should be accessible for all, our research objective is to develop (a) a technological framework for (semi)-automatically adapting short stories in Spanish into easier and accessible versions close to the E2R principles, and (b) an evaluation framework to ensure that the easier versions of the short stories provided by the technological framework preserve the content and semantics of the original versions. In this regard, the developed technological framework should be useful to help (a) different target groups who present comprehension difficulties (e.g. people with cognitive disabilities, low literacy skills or nonnative speakers) and (b) E2R professionals in terms of streamlining the task of text adaptation.

Keywords

Cognitive Accessibility, Text Adaptation, Easy-to-Read Methodology, Artificial Intelligence

1. Introduction and Motivation

People with cognitive disabilities have some difficulties related to reading comprehension. Over the last decade, the need for comprehensible and accessible materials for people with learning difficulties has received increased attention in the social sphere [1]. In that sense, a methodology called Easy-to-Read (E2R) [2, 3, 4, 5] was created to present clear and easily understood contents to different sectors of the population that include people with cognitive disabilities and people with limited reading proficiency, among others. The methodology provides a collection of guidelines and recommendations related to both writing and layout aspects, such as avoiding the use of complex syntax, long words, or metaphorical and figurative language, for instance. However, this methodology presents three main limitations: (a) it lacks clear information on how to apply the guidelines and recommendations; (b) the guidelines it proposes are general, i.e. there is no specific information for adapting different types of texts; and (c) it relies heavily on manual adaptation, a highly time-consuming and subjective task.

CEUR Workshop Proceedings (CEUR-WS.org)

To cover these identified gaps, our research is focused on applying different Artificial Intelligence (AI) methods and techniques to (semi)-automatically[1] perform, in particular, both the analysis and adaptation based on the E2R Methodology[2] of short stories[3] written in Spanish. The motivation for leaning towards this specific type of text for our research is based on the following three reasons:

1. **Variety of linguistic aspects.** Short stories harbour wide vocabulary registers, different syntactic structures, pragmatic implications and heterogeneous semantics.
2. **Fixed structure.** This type of texts generally present a tripartite organisation: introduction, development and denouement. In addition, they are of moderate length.
3. **Social impact.** The processes involved in reading comprehension are essential for people's cognitive development [6]. Furthermore, the act of reading literary texts is seen as part of a wider process of human development and growth based on understanding both one's own experience and the social world [7]. On this account, cultural materials, in general, and literary materials such as short stories, in particular, should be accessible for all.

The rest of the paper is organised as follows: Section 2 is devoted to the state-of-the-art on the automatic approaches for both checking the E2R guidelines in texts and for adapting texts in Spanish into simpler versions. In Section 3 we pose the Open Research Problem as well as the Hypothesis and Research Questions in which we base this PhD thesis. For its part, Section 4 explains the main Objective and the different Sub-Objectives of this research. We proceed with the Research Methodology in Section 5, and Section 6 shows some Ongoing Work carried out to date. Finally, we present conclusions and further research.

2. State of the Art

The already introduced Easy-to-Read (E2R) Methodology plays a decisive role in the process of adapting materials. However, as mentioned, its main drawback lies in the manual adaptation method, which demands considerable resources and effort. To address this limitation and make the process more efficient, there are semi-automatic tools both for checking the E2R guidelines in texts and for providing simpler adaptations in Spanish. On the E2R assessment, we can mention Easy-to-Read Advisor [8], an E2R conformance checker for assessing a particular document with respect to the E2R guidelines, and FACILE [9], an extension and improvement of Easy-to-Read Advisor. On the other hand, considering the automatic adaptation of texts into simpler versions, we can point out the following approaches based on simplification techniques: Simplext [10], LexSIS [11], DysWebxia or [12] and EASIER [13]. These existing works are based on simplification techniques at both lexical and syntactic levels. Regarding lexical simplification,

---

[1]We refer to a semi-automatic adaptation in the sense that we cannot assume that such an adaptation is completely accurate, so in many cases it will be the adapter who tailors the text to the appropriate context.

[2]The automatic adaptation can be considered an intralinguistic automatic translation. However, we decided to use the word *adaptation* along the research since this is the most appropriate terminology in the E2R area.

[3]In Literature, a short story is a piece of prose fiction that typically can be read in one sitting and focuses on a self-contained incident or series of linked incidents.

the tool presented in LexSIS [11] employs three techniques to find a simpler lexical substitute: a word-based vector model, word frequency, and word length. To make such a substitution for a simpler candidate, it relies on semantic resources available online such as OpenThesaurus[4] and the Corpus de Referencia del Español Actual (CREA[5]). The work presented in DysWebxia [12] makes use of LexSIS [11] to replace words that have been identified as complex for people with dyslexia by synonyms that fit their comprehension needs. For its part, the system proposed in EASIER [13] identifies complex words using Support-Vector Machines (SVM) algorithms and replaces them with easier synonyms by consulting external open-access resources such as BabelNet[6] and OpenThesaurus. In the framework of the Simplext project [10], we find a double automatic simplification: on the one hand, lexical, based on the implementation of LexSIS and a rule-based simplification; and on the other hand, syntactic. On such a syntactic simplification, they make use of a hand-written computational grammar and focus on reducing sentence complexity.

However, to the best of our knowledge, there is no research work oriented towards the automatic adaptation of a specific literary text type such as short stories or short narratives into easy-to-read versions in Spanish (in languages other than Spanish it is worth mentioning ERNESTA tool [14], in which the authors have worked on the syntactic simplification of short stories in Italian developing a system which focuses on (a) resolving anaphoric references and (b) rewriting sentences in a simpler form using the present tense of verbs). In this regard, it is important to note that the typology of a text is crucial when considering its adaptation, since not all text types share similar linguistic properties and, therefore, a general adaptation of texts is not valid for all typologies. On the other hand, as we mentioned, all existing work focuses on automatically simplifying texts, but we must cautiously discern that simplifying and adapting do not point to similar processes, since adapting implies taking into account the needs of target users, while simplifying focuses on reducing the complexity of formal aspects of a text without necessarily taking into account the target user. Thus, an adaptation must be based on user research in order to be useful in terms of functionality.

3. Open Research Problem, Hypothesis and Research Questions

Based on the foregoing, the **Open Research Problem (ORP)** in which our research objectives are based is "The number of adapted short stories in Spanish into easy-to-read versions is quite scarce. Thus, literature is limited for people with reading comprehension difficulties".

In this line, our research work raises the following **Research Hypothesis (RH)**: "The (semi)-automatic adaptation of short stories written in Spanish into easier versions close to the E2R Methodology is possible and the resulting adapted versions will maintain the content and semantic information for people with cognitive disabilities". Having considered both the Open Research Problem (ORP) and the Research Hypothesis (RH), the following **Research Questions (RQ)** can be proposed:

---

[4]https://www.openthesaurus.de/
[5]https://www.rae.es/banco-de-datos/crea
[6]https://babelnet.org/

- **RQ1.** Which AI-based techniques and methods can be used in the automatic adaptation of short stories in Spanish?
- **RQ2.** Which metrics (indices or formulas) can be useful for assessing the reading comprehension of automatically adapted short stories by people with cognitive disabilities?
- **RQ3.** Are the adapted versions by the technological framework faithful to the content and semantics of the original versions?

4. Objective and Sub-Objectives

As a first attempt, the **Research Objective (RO)** of this work in progress is to develop (a) a technological framework for adapting (semi)-automatically short stories in Spanish into easier versions close to the E2R Methodology, and (b) an evaluation framework to ensure that the easier versions of the short stories provided by the technological framework preserve the content and semantics of the original versions.

Hence, to achieve our propose, the **sub-objectives (SOs)** that could be proposed are the following:

- **SO1.** To identify the main cognitive difficulties in short stories regarding reading comprehension by people with cognitive disabilities. This identification will be experiment-based involving user studies.
- **SO2.** To determine and classify the different linguistic criteria (in Spanish) related to the identified cognitive problems, in order to select (a) which E2R guidelines should be included in our framework, and (b) to establish a priority order for the guidelines to be included in the framework.
- **SO3.** To determine which AI-based methods and techniques should be considered in the technological framework development; and thus used them in the automation of the guidelines.
- **SO4.** To evaluate whether the adapted versions of short stories provided by the technological framework maintain both the content and semantics of the original versions.

5. Research Methodology

In order to achieve the aforementioned objectives and bear the hypothesis out, the current research work is structured into three stages:

1. **User-based Studies.** In this first stage we conducted an exhaustive analysis of the cognitive and linguistic issues that affect the readability of short stories, in order to complete the information provided by the E2R Methodology on how to apply it. To achieve this goal, user-based studies are needed. Thus, we designed and set in some studies to find the linguistic aspects involving reading comprehension difficulties. This stage is directly related to the SO1.
2. **Technological Framework Development.** In this second stage we will (a) select which guidelines should be included in the framework and (b) establish a priority order based on

the results obtained in the user studies, with the aim of automating the selected guidelines by means of Artificial Intelligence (AI) methods and techniques such as Natural Language Processing (NLP), Pattern Matching, or Machine Learning.

3. **Evaluation Framework Development.** The latter stage will cover the evaluation of the previous processes. Since it will involve persons with cognitive disabilities, the evaluation will also include user studies. In this stage we will implement readability and understandability values to analyse whether the (semi)-automatically adapted versions of short stories maintain both the content and semantics of the original versions.

Stages 2 and 3 will be iterative and they are related to SO2, SO3 and SO4.
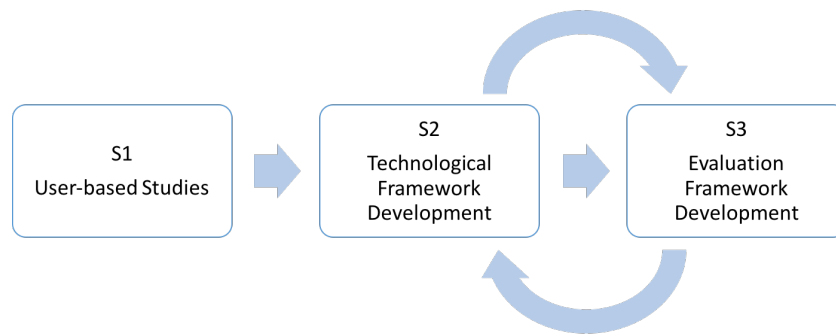
Figure 1: Iterative process of the work plan stages.

In more detail, considering both the objective and sub-objectives and the stages in which the research is structured, the distribution schedule of the different tasks is presented in Figure 2, organised over the four years foreseen.

	Years 0-1	Years 1-2	Years 2-3	Years 3-4
Study of the state-of-the-art related to the (semi)-automation of short stories in Spanish and any other languages.	▮							
Study of the cognitive issues concerning language comprehension by people with cognitive disabilities.		▮						
Design and set in of the user-based studies to find the linguistic aspects involving reading comprehension difficulties.			▮	▮				
Technological framework development based on these linguistic criteria and the results extracted from the user studies.			▮	▮	▮	▮		
Evaluation framework development in terms of readability and understandability values.					▮	▮		
Writing of the thesis.						▮	▮	▮

Figure 2: Gantt chart of the research plan.

6. Ongoing Work

We are currently in the first stage of the PhD thesis, in which, thanks to the collaboration with several institutions in Spain, we have launched a user study to gather data on the problems that affect reading comprehension of persons with cognitive disabilities. More than 70 users

with cognitive or intellectual disabilities participated in this study, including a small group of validators (i.e. people with disabilities who are familiar with the E2R methodology guidelines). In order to extract the cognitive-linguistic problems in the participants' reading comprehension of short stories, we based the user study on questionnaires[7] in which we asked participants about different short stories (we collected a corpus of texts ranging from 100 to 350 lines in length each), by proposing questions based on free recall methods, induced recall methods, and metacognitive questions.

After extracting the data, we observed that the problems or difficulties in terms of reading comprehension of short stories are grouped in the following different linguistic dimensions:

- **Vocabulary.** Participants present difficulties regarding infrequent words, or adverbs ending in -mente (-ly in English).
- **Syntax.** Structures as clause chains (complex subordinate and coordinated constructions), juxtaposition clauses, or tempo-causal appositions, raise a problem for people with reading comprehension difficulties.
- **Figurative Language.** The use of figurative language such as metaphors in morals are very common in this type of text, which poses a major problem in the literal comprehension of the narratives.
- **Cohesion and Coherence.** Likewise, this type of text, as it is narrative, presents challenges when there appear, for example, anaphoric references to characters (it is preferable to always use proper names instead of pronouns or other reference marks), spatio-temporal inferences of events (discourse markers should be used to organise the narrative clearly), or ambiguity in references to different characters (the use of pronouns pointing to more than one character can cause confusion).
- **Layout.** In this dimension the focus is mainly on the format of the dialogues, as they should appear in a theatrical style (i.e. giving prior notice of who the speaker is before the introductory hyphen).

Considering these dimensions, we now have a roadmap on which aspects should be addressed in the second stage regarding the technical framework development.
Furthermore, alongside the launch of the user study, we have been analysing which AI techniques may be the most suitable for our development by implementing several proofs-of-concept based on E2R guidelines, such as to avoid the use of verbal periphrases, to avoid the use of adverbs ending in -mente and superlative forms, to avoid the use of abbreviations and acronyms, to avoid the use of passive voice, or to avoid the use of lexical repetitions. To implement such proofs-of-concept we have mainly relied on Natural Language Processing (NLP)-based techniques. Some of this work is already described in published papers [9, 15, 16, 17], and other work is accepted for presentation at conferences in the coming months. In these papers we address some of the difficulties in the different dimensions outlined above. We have also carried out proofs-of-concept in final theses, trying to solve syntactic[8], metaphorical[9], semantic[10] or dialogue

---

[7]A sample questionnaire (in Spanish) in relation to short story 1 is available here: https://short.upm.es/argi7
[8]https://oa.upm.es/72751/
[9]https://oa.upm.es/71157/
[10]https://oa.upm.es/75098/

formatting[11] issues. For the time being, based on the work already carried out, we have observed that techniques based on declarative methods are proving to be the most effective. The results, which can be found in these published works, are apparently acceptable.

7. Conclusions and Further Research

After laying all this out, and provided the results are favourable, one of the expected conclusions of this PhD thesis work would be considering it as an original contribution, since the task of adaptation of literary texts into easy-to-read versions has not yet been addressed in Spanish.

For this reason this work plan can be ambitious. Thus, there will be a broad path to continue in this research. As a further step, we would like to study particular impairments (not only cognitive disabilities in general), in order to analyse whether a particular adaptation for each disability is needed or not.

Additionally, it would be worthwhile to compile a corpus to apply other AI techniques such as Machine Learning or Deep Learning in terms of, for instance, developing a language model for text adaptation into easy-to-read versions.

On the other hand, the developed framework should be useful to help (a) different target groups who present comprehension difficulties (e.g. people with cognitive disabilities, low literacy skills or nonnative speakers) and (b) E2R professionals in terms of streamlining the task of text adaptation.

In addition, this research work will help on (a) the development and fulfilment of the Sustainable Development Goals[12] (SDGs) 4, 5 and 10 proposed by the UN, and (b) the recognition and achievement of cognitive accessibility.

Acknowledgments

References

[1] K. Matausch, A. Nietzio, Easy-to-read and plain language: Defining criteria and refining rules, 2012. URL: http://www.w3.org/WAI/RD/2012/easy-to-read/paper11.

[2] M. Nomura, G. S. Nielsen, International Federation of Library Associations and Institutions, Library Services to People with Special Needs Section, Guidelines for easy-to-read materials, IFLA Headquarters, The Hague, 2010.

---

[11]https://oa.upm.es/75513/

[12]https://sdgs.un.org/goals

[3] AENOR, Lectura Fácil. Pautas y recomendaciones para la elaboración de documentos (UNE 153101:2018 EX), Asociación Española de Normalización, 2018.

[4] Inclusion Europe, Information for All. European standards for making information easy to read and understand, Inclusion Europe, Brüssel, 2009. OCLC: 838005460.

[5] E. B. Union, Making information accessible for all, 2017. URL: https://www.euroblind.org/publications-and-resources/making-information-accessible-all.

[6] M. Nikolajeva, Leer ficción es bueno para el desarrollo cognitivo, emocional y social, Alabe Revista de Investigación sobre Lectura y Escritura (2019). doi:10.15645/Alabe2019.20.12.

[7] P. Freire, L. Slover, The importance of the act of reading, The Journal of Education 165 (1983) 5–11.

[8] M. C. Suárez-Figueroa, E. Ruckhaus, J. López-Guerrero, I. Cano, Álvaro Cervera, Towards the Assessment of Easy-to-Read Guidelines Using Artificial Intelligence Techniques, in: K. Miesenberger, R. Manduchi, M. C. Rodriguez, P. Peňáz (Eds.), Computers Helping People with Special Needs. ICCHP 2020, volume 12376 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 74–82.

[9] M. C. Suárez-Figueroa, I. Diab, E. Ruckhaus, I. Cano, First steps in the development of a support application for easy-to-read adaptation, Universal Access in the Information Society (2022). doi:10.1007/s10209-022-00946-z.

[10] H. Saggion, S. Stajner, S. Bott, S. Mille, L. Rello, B. Drndarevic, Making It Simplext: Implementation and Evaluation of a Text Simplification System for Spanish, ACM Transactions on Accessible Computing 6 (2015).

[11] S. Bott, L. Rello, B. Drndarevic, H. Saggion, Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish, in: Proceedings of COLING 2012, The COLING 2012 Organizing Committee, Mumbai, India, 2012, pp. 357–374.

[12] L. Rello, R. Baeza-Yates, H. Saggion, DysWebxia: Textos más Accesibles para Personas con Dislexia, Procesamiento del Lenguaje Natural 51 (2013) 205–208.

[13] L. Moreno, R. Alarcón, P. Martínez, EASIER system. Language resources for cognitive accessibility., in: The 22nd International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '20, Association for Computing Machinery, Virtual Event, Greece, 2020, pp. 1–3.

[14] G. Barlacchi, S. Tonelli, Ernesta: A sentence simplification tool for children's stories in italian, 2013, pp. 476–487. doi:10.1007/978-3-642-37256-8_39.

[15] M. C. Suárez-Figueroa, I. Diab, A. González, J. Rivero-Espinosa, First attempt to an easy-to-read adaptation of repetitions in captions, in: Computers Helping People with Special Needs: 18th International Conference, ICCHP-AAATE 2022, Lecco, Italy, July 11–15, 2022, Proceedings, Part I, Springer-Verlag, Berlin, Heidelberg, 2022, p. 417–424. doi:10.1007/978-3-031-08648-9_48.

[16] M. C. Suárez-Figueroa, I. Diab, Á. González Sanz, J. Rivero-Espinosa, Automatic easy-to-read translation of morphological structures in spanish texts, Procesamiento del Lenguaje Natural 71 (2023) 191–203. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6553.

[17] M. C. Suárez-Figueroa, I. Diab, Á. González, J. Rivero-Espinosa, Towards an automatic easy-to-read adaptation of morphological features in spanish texts, in: J. Abdelnour Nocera,