Procesamiento del Lenguaje Natural, Revista nº 72, marzo de 2024, pp. 99-108 recibido 15-12-2023 revisado 27-01-2024 aceptado 03-02-2024

# Enhancing Clarity: An Evaluation of the Simple.Text Tool for Numerical Expression Simplification

## *Mejorando la claridad: Una evaluación del sistema Simple.Text para la simplificación de expresiones numéricas*

**Isabel Espinosa-Zaragoza,[1] Paloma Moreda[2] and Manuel Palomar[2]**
[1]Centre of Digital Intelligence, University of Alicante
[2]Department of Computing and Information Systems, University of Alicante
isabel.espinosa@ua.es, {paloma,mpalomar}@dlsi.ua.es

**Abstract:** Numerical information in written texts impacts their readability and is considered complex for people with cognitive disabilities by the Easy-to-Read guidelines. This paper presents Simple.Text, a rule-based system designed to automatically simplify all numerical expressions deemed complex, with a focus on Rules 19-25 from Section 6.2 of the Easy-to-Read guidelines. The results from the evaluation indicate a high precision and accuracy in numerical phenomena detection and transformation, although with some limitations. This system proves to be an efficient and cost-effective tool for the simplification of numerical expressions.
**Keywords:** Numerical expressions, Easy-to-Read (E2R), rule-based system, cognitive disabilities.

**Resumen:** La información numérica en los textos escritos afecta a su legibilidad y las pautas de Lectura Fácil las consideran complejas para las personas con discapacidad cognitiva. Este artículo presenta el sistema Simple.Text, un sistema basado en reglas diseñado para simplificar automáticamente todas las expresiones numéricas consideradas complejas, que aborda específicamente las reglas 19-25 de la Sección 6.2 de Lectura Fácil. Los resultados de la evaluación indican una alta precisión y exactitud en la detección de fenómenos numéricos y en su transformación, aunque con algunas limitaciones. Este sistema demuestra ser una herramienta eficiente y rentable para la simplificación de expresiones numéricas.
**Palabras clave:** Expresiones numéricas, Lectura Fácil, sistema de reglas, discapacidad cognitiva.

## *1 Introduction*

Numerical information in texts impacts their readability (Rello et al., 2013). The simplification of this information is required to guarantee an egalitarian access to information. Facilitating the understanding of language helps citizens to properly exercise their rights and obligations.

The Easy-to-Read guidelines (AENOR, 2018) include several rules for the simplification of numerical expressions, namely Section 6.2, Rules 19-25. These basically entail the clarification of percentages and fractions, dates and hours, and ordinal numbers, amongst others.

As a way of example, percentages like "20%" are transformed into descriptive explanatory clauses in order to avoid using the symbol "%" and providing a more comprehensible quantity. The target audience of these recommendations is people with cognitive disabilities, especially dyslexia and dyscalculia, which are particularly affected by numerical expressions, but not limited to those people.

Previous works emphasise how difficult some numerical expressions are to process and suggest that numbers are more readable in figures than in letters for people with dyslexia (Rello et al., 2013). Additionally, numerical expressions also pose comprehen-

ISSN 1135-5948 DOI 10.26342/2024-72-7 

sion problems for people with limited education (Bautista et al., 2011). As can be seen, this issue affects a wide sector of the population that could benefit from more accessible texts. Therefore, Automatic Text Simplification (ATS), "a technology for producing adaptive texts by reducing their syntactic and lexical complexity to make them readable for a user group of users" (Bott and Saggion, 2012), can be of assistance in the simplification of numerical phenomena for any target audience.

The purpose of this paper is to present a rule-based system, Simple.Text, to simplify all of the numerical expressions considered complex in Section 6.2, Rules 19-25, from the Easy-to-Read guidelines. This tool is developed within the ClearText project[1], funded by the MCIN/AEI/10.13039/501100011033 Government and the European Union NextGenerationEU/PRTR (grant reference TED2021- 130707B-I00) and developed by the GPLSI research group[2] of the University of Alicante.

This paper is structured as follows: Section 2 includes a literature review covering ATS tools for numerical expressions; Section 3 presents the papers' objectives and methodology; Section 4 delves into the rule implementation in the system, by describing every category identified and transformed by the system; Section 5 presents the Simple.Text system; Section 6 describes the system evaluation while Section 7 details its findings. Lastly, Section 8 concludes with the future work ahead.

## 2 Related Work

Previous works in the ATS of numerical expressions are scarce. We depart from the findings from an empirical study in Bautista et al. (2012) on a parallel corpus of original and manually simplified Spanish texts, along with a survey. This study focuses on the simplification of numerical expressions with the intention of implementing the rules computationally, but no actual simplification system is presented.

Similarly, in Drndarević and Saggion (2012), we also encounter the findings of an analysis of a parallel corpus in Spanish (original and simplified) where numerical expressions are taken into account for the development of a simplification system for Spanish. More particularly, (1) the replacing of a word with a figure ("cinco" turns into "5"); (2) the rounding of big numbers ("más de 540.000 personas" turns into "medio millón de personas); (3) the rounding by elimination of decimal points ("1,9 millones" turns into "2 millones"); (4) the simplification of noun phrases containing two numerals in plural and the preposition of by eliminating the first numeral ("cientos de miles de personas" turns into "miles de personas"); (5) the substitution of words denoting a certain number of years (decade, centenary) by the corresponding number; and (6) the representation of thousands and millions in big numbers expressed by means of a word ("17.000" becomes "17 mil").

To our knowledge, the earliest rule-based system that addresses such issues with a lexical transformation component and a syntactic simplification module is present in Bautista et al. (2013). There we can find a first approximation to the task of simplifying numerical expressions automatically in a text and to varying degrees of difficulty. More specifically, the following replacements are considered: (1) replacing decimal percentages with percentages without decimals; (2) replacing decimal percentages with ratios; (3) replacing percentages with ratios; (4) replacing decimal percentages with fractions; (5) replacing percentages with fractions; (6) replacing ratios with fractions; (7) replacing numerical expressions in words with numerical expressions in digits. This proposal is for English, but with the intention of developing a version for Spanish.

In Bautista and Saggion (2014), the researchers present a rule-based lexical component for the simplification of numerical expressions in Spanish texts based on survey choices for simplification. The system is composed of: (1) text processing using FreeLing; (2) the transformation of the FreeLing output into XML representation; (3) the application of grammars for numerical expression recognition; (4) the simplification of target numerical expression; and lastly, (5) a sentence rewriting stage. Among the numerical expressions tackled in this work, there is the rounding of percentages ("18,55%" is transformed into "19%") but not the simplification of the percentage in itself, as recommended by the European Easy-to-Read

[1] https://cleartext.gplsi.es/

[2] https://gplsi.dlsi.ua.es/

guidelines (AENOR, 2018).

Lastly, an ATS system for Spanish is presented in Bautista et al. (2017), where the following phenomena are considered: (1) partitive numerals like for example, "un millón" (a million) or "una centena" ("a hundred"); (2) monetary expressions consisting of quantity and the monetary unit, as in " 2.000 dólares" ("2,000 dollars"); (3) fractions and percentages, like "34%", are substituted with the lemma "34/100"; and (4) physical measures, for example, "30 km/h".

As can be observed, apart from the scarcity of systems for the simplification of numerical expressions in Spanish, the papers presented offer a partial and not a global solution to the simplification of numerical expressions. That is, these do not encompass the entire range of numbers identified as obstacles in the Easy-to-Read guidelines. Thus, we propose a rule-based system to tackle the entirety of these numerical phenomena described in the Easy-to-Read guidelines (AENOR, 2018).

## 3 Objective and Methodology

The objective of this paper is to identify and resolve the complexity associated with the numeric phenomena deemed as difficult to comprehend by the Easy-to-Read guidelines (AENOR, 2018) in a rule-based system by transforming them into simpler expressions.

As the Easy-to-Read guidelines are often general and flexible rules, the collaboration with the non-governmental organisation APSA[3] has enlightened the path by defining the restrictions to such rules. This NGO has a group of expert Easy-to-Read validators with cognitive disabilities. From this collaboration, we have been able to define and restrict several rules that were rather loose. For example, Rule 19 in Section 6.2 suggests using Arabic numerals. However, for the numbers "100" and "1000", APSA recommends using the written version (e.g. "cien" (a hundred) and "mil" (a thousand), respectively). Drawing on the expertise of APSA's specialists in text simplification and validation according to the Easy-to-Read guidelines, we opted to incorporate their insights and rule specifications into our system. This decision was made to leverage their expertise on the matter, contributing to the creation of a more effective system thanks to this synergy.

Our methodology consists of the following steps:

- 1. Developing a system for the identification and transformation of numerical expressions in Spanish texts.
- 2. Building a rule-based system for automatic simplification of numerical expressions.
- 3. Evaluating the automatically simplified output.

## 4 Rule Implementation

All the rules contemplated in this system correspond to Section 6.2, the lexical simplification section, in the Easy-to-Read guidelines (AENOR, 2018). Many rules include more than one transformation or implementation. Table 1 includes a summary of each rule implemented by the system and its corresponding rule number plus an example.

First of all, it is necessary to identify numbers in the text whether they are expressed in letters or figures. For this reason, it is imperative to undertake a preliminary processing in order to identify and resolve numbers in letters. In this preprocess, numbers in letters are identified using SpaCy library.[4] Once a number written in letters is identified, then it is replaced by its corresponding Arabic numeral using a predefined dictionary. Then, the process to identify and resolve numbers is run.

The process is divided into two phases: (i) identification and (ii) resolution. The identification phase is carried out using the SpaCy tool (part-of-speech tagging) in order to determine if a word begins and finishes with numerical characters. In that case, we will regard it as a number to be treated in the resolution phase.

The resolution phase requires different converting rules depending on the type of number identified. Consequently, it is essential to determine to which numerical category each of the numbers identified in the identification phase belongs to. This is accomplished in the following order: dates, times, telephones, percentages, ordinals and Roman numerals and other quantities. In this way, in order to assign a number to a category, we first verify that it has not been identified

[3] https://www.asociacionapsa.com/

[4] https://spacy.io/

| Rule | Numerical expression | Original | Easy-to-Read |
|---|---|---|---|
| Rule 19 | Figures | dos | 2 |
| Rule 19 | Rounding quantities | 139 | más de 100 |
| Rule 19 | Explain big numbers | 60.000 | 60 mil |
| Rule 20 | Phone numbers | 123456789 | 123 45 67 89 |
| Rule 21 | Ordinal numbers | primero, undécimo | primero, 11 |
| Rule 22 | Percentages | 20% | 2 de cada 10 |
| Rule 23 | Dates | 01/01/(20)20, 01-01-(20)20 | 1 de enero de 2020 |
| Rule 24 | Time | 23:30 | 11 y media de la noche |
| Rule 25 | Roman numbers | Jaime I | Jaime primero |

Table 1: Summary of the rules implemented.

in any of the previous categories, and then we check if it complies with the identification rule of that particular category. This sequential order is necessary to prevent incorrect and duplicate substitutions.

The following subsections include the different numerical expressions or categories considered complex for people with cognitive disabilities by the Easy-to-Read guidelines. Each category is defined, followed by an explanation of its detection pattern and resolution or transformation process.

### 4.1 Dates

According to the Easy-to-Read guidelines, compact dates expressed with hyphens or slashes are not recommended. These include orthotypographic symbols that can be complex to process (Section 6.2, Rule 23, (AENOR, 2018)). This recommendation closely aligns with Rule 8 in Section 6.1 (AENOR, 2018), which specifies that these orthotypographic symbols should be avoided.

The identification of dates is achieved by using a regular expression. By utilising this regular expression, dates following the format DD/MM/YY(YY) are identified. It must be pointed out that the year can be two or four digits and expressed by means of a hyphen or a dot as a separator.

To transform the detected dates into the recommended format, a dictionary is utilised to establish the relationship between months in letters and months in numbers. Hence, dates such as "12/04/2020" or "12-04-20" should be displayed fully written, as follows: "12 de abril de 2020".

### 4.2 Times

To identify whether a number represents a time, a regular expression is used to detect the format "hh:mm". When it comes to the automatic simplification of time, time slots are highly cultural. In Spanish these are:

- In the morning (from 06:00 to 12:59), e.g. 6 y media de la mañana.
- In the afternoon (from 1:00 p.m. to 8:59 p.m.), e.g. 1 y 20 de la tarde.
- At night (from 9:00 p.m. to 12:59 a.m.), e.g. 6 y 35 de la tarde.
- In the early morning (from 01:00 to 05:59), e.g. 3 menos cuarto de la madrugada.

Therefore, set hours (e.g. o'clock, quarter past, half past and a quarter to) are always written (e.g. en punto, y cuarto, y media, menos cuarto). Consequently, instead of "23:30 PM", the simpler written version should be "11 y media de la noche". In-between times, like "10:10 AM" or "23:35 PM" should be transformed into "10 y 10 de la mañana," and "11 y 35 de la noche", respectively.

### 4.3 Telephone Numbers

If the identified number in the text consists of nine consecutive digits, and the sentence in which the number appears contains the word "teléfono" (telephone) or "móvil" (mobile) either two words before or after the nine consecutive digits, it will be classified as a telephone number. In such instances, the correct transformation resolution is to separate the nine consecutive digits with spaces following a 3-2-2-2 structure (e.g. 123 45 67 89).

### 4.4 Percentages

Concerning the simplification of percentages, it is recommended to find alternative rephrasing options, as orthotypographic symbols

such as "%" are regarded as complex (i.e. Section 6.1., Rule 8, (AENOR, 2018)). Thus, if the character after the number is the percentage symbol (%), then the number is categorised as a percentage (e.g. 20%, 37%).

Once identified, it is then substituted with an analogous expression using a rule that replaces the number and the percentage symbol with the number and a text. This text varies depending on whether the amount is divisible by 10 or not. On the one hand, when the number is divisible by 10, the text is " x de cada 10" (x out of 10). On the other hand, when the number is not divisible by 10, the text is "x de cada 100" (x out of 100). this is exemplified below:

- "20%" is simplified as "2 de cada 10" (2 out of 10).
- "37%" is simplified as "37 de cada 100" (37 out of 100).

### 4.5 Ordinal Numbers

In line with the Easy-to-Read guidelines, the use of ordinal numbers should be changed to cardinal. Nevertheless, our collaborators affirm that written ordinal numbers from one to ten are understood by people with cognitive disabilities and do not need to be changed, according to their experience. For example, "primer/primero(s)/a(s)" (first), "segundo(s)/a(s)" (second), "tercer, tercero(s)/a(s)" (third), etc. Thus, these remain ordinal and in written form, as the validators understand them. However, from eleven onwards, these are changed to cardinal numbers: "undécimo" (eleventh) is changed to "11". Therefore, an example such as "Juan vive en la planta 18º/decimoctava" (Juan lives on the 18th/eighteenth floor) should be reworded to a simpler version using cardinal numbers, like the following: "Juan vive en la planta número 18" (Juan lives on floor number 18). The detection and substitution are performed by using a dictionary that has been manually created specifically for this purpose.

### 4.6 Roman Numerals

As for the named entities when these are proper names of kings, Roman numerals adapt to letters. Nevertheless, they do not undergo a double adaptation from Roman numerals to letters and from ordinal numbers to cardinals in the first ten cases:

- "Jaime I" (James I) changes to "Jaime primero" (James the first).
- "Siglo XX" (20th century) becomes "Siglo 20" (20 century).

Even though the Easy-to-Read guidelines (AENOR, 2018) indicate including "que se lee" (what reads as) when treating these cases (e.g. Alfonso X que se lee Alfonso décimo), our collaborators indicated that it is much more straightforward to do it in this way. In order to detect and replace Roman numerals a Roman Phyton Library[5] is used.

Within the Roman numerals, it is necessary to distinguish kings' names (e.g. Jaime I) since the resolution process is different. This case is identified when Named Entity Recognition (NER) + Roman numeral appears in the text. Upon detection using SpaCy, the Roman number is replaced by the corresponding ordinal number (e.g "Jaime I" becomes "Jaime primero").

### 4.7 Other Quantities

If the detected number has not been considered in any of the above categories, then it is regarded as a quantity. To identify them, a regular expression is applied to detect numbers with or without (1) a thousands separator or (2) a decimal point for decimals. The substitution process is subdivided into the following subsections:

#### 4.7.1 Figures

It is recommended to write numbers in figures up to 1,000, that is, from 1 to 999. This is already done in the preprocessing phase explained in Section 4.

#### 4.7.2 Explain Big Numbers

From there, big numbers are changed to a hybrid format where part of the number is written with Arabic numbers and the rest is expressed in written format: "2 mil" (2 thousand). This approach replaces the zeros with their textual equivalent, rather than representing them as numerals. This complies with the Easy-to-Read guidelines (AENOR, 2018), which state that numbers with many digits are difficult to read and, thus, writing them in letters can make them easier to understand. To facilitate their understanding, alternative options are contemplated, such as:

[5]https://pypi.org/project/roman/

- Qualitative comparisons (e.g. as many people as those who live in Granada).
- Replacement by terms such as " several", " thousand" and others when the context allows it.

When it comes to numbers, what is considered big is open to interpretation. The Easy-to-Read guidelines are flexible in so much that these do not set a limit in principle: it depends on the validation sessions and whether it is understandable or not there. Most of the time it depends on the context and the relevance that this number in question has in the text. Currently, according to the validation groups working on this project, we established that the number from which we would apply this is "10.000" which is transformed into "10 mil" (10 thousand). That being said, "100" and "1.000" are also transformed into "cien" (a hundred) and "mil" (a thousand), as previously discussed in Section 3.

#### 4.7.3 Rounding Quantities

Rounding numbers is recommended by the Easy-to-Read guidelines (AENOR, 2018) at the expense of losing precision. This is applied to decimal numbers (e.g. "1.3" is rounded to "1") and other quantities. That is, "1.999" is rounded to "casi 2 mil" (almost 2 thousand). Nevertheless, some exceptions are contemplated, like ticket prices, won prizes, and others, although no implementations are applied yet in this regard until we enter the project's meaning and disambiguation module.

## *5 The Simple.Text System*

The current version of the Web App allows for the selection of (1) individual language phenomena simplification, enabling the simplification of specific language phenomena such as superlative forms or *-mente* adverbs, amongst others; (2) language level simplification, which offers the choice of simplifying the entire palette of linguistic phenomena organised by language levels (currently limited to lexical and syntactic); and (3) applying all simplifications at once. Subsequently, users submit the text for simplification on the top box and obtain the output in the box below. Figure 1 illustrates an example of simplification in the current preliminary interface.

## *6 System Evaluation*

The system evaluation is performed by detecting and resolving the numeric phenomena in 5,000 texts from the CLEARSIM corpus, which contains texts from the public administration. This accounts for one third of the total texts in that corpus. These texts were gathered from the official websites of municipalities in the Alicante area, focusing on the domains of culture, sports, and leisure. We utilised the Simple.Text System to identify and transform the numeric expressions deemed complex by the Easy-to-Read guidelines (AENOR, 2018).

Given the impossibility of presuming complete system detection and transformation, we conducted a manual evaluation involving a representative quantity of texts to simplify and, subsequently, we scaled the results. This corpus will be available on the project's website.[6] To do so, we extracted a representative number of texts out of the 5,000 texts by following the Formula 1 presented in (Pita-Fernández, 1996):

$$M = \frac{N * K^2 * P * Q}{E^2 * (N-1) + K^2 * P * Q} \quad (1)$$

The symbols in the equation stand for the following: $N$ for population, $K$ for the confidence interval, $P$ for the success probability, $Q$ for failure probability and $E$ for the error rate. The values given to each of these parameters, more specifically, $K$=0.95, $E$=0.05, $P$=0.5, and $Q$=0.5 were taken from (Vázquez et al., 2010).

After calculating the formula, the resulting number of texts $M$ was 89, which then was rounded up to 90 texts. These texts were manually analysed by a human to check the accuracy of both the linguistic phenomena detection and the linguistic phenomena resolution. The human detection evaluation yielded 1,597 numerical expressions that are categorised as follows:

- Figures: 966
- Written numbers: 178
- Decimal numbers: 51
- Dates: 14
- Hours: 226
- Percentages: 11

[6] `https://cleartext.gplsi.es/`

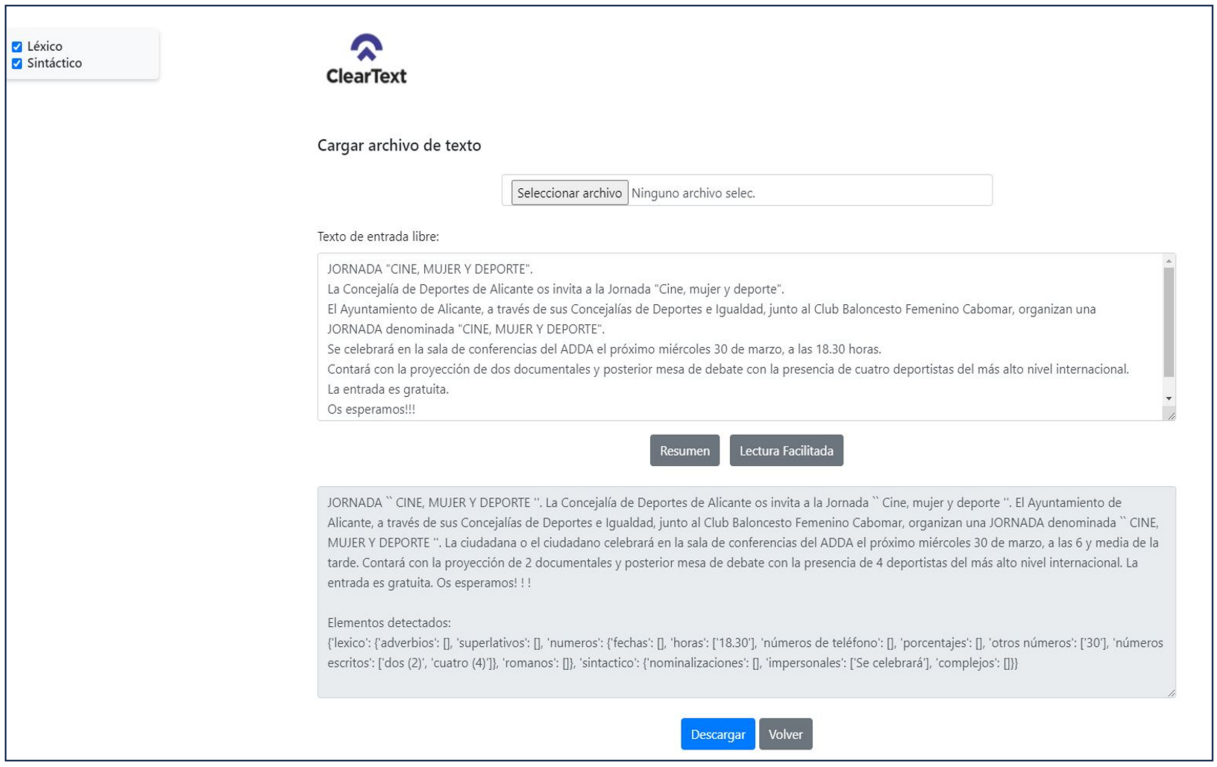


Figure 1: Simple.Text Tool.

- Ordinal numbers (both written and in figure): 92
- Roman numbers: 53
- Phone numbers: 6

After this initial human detection, an evaluation of the system's detection and transformation was performed.

## 7 *Discussion of Results*

This section discusses both the detection and transformation of numerical expressions in the current version of the Simple.Text system.

The results for the detection of numerical phenomena are presented in Table 2, which includes a description of the accuracy, precision, recall and F1-score (Derczynski, 2016) for the detection of every single numerical expression analysed. Regarding the detection of the numerical categories, we observe an overall good detection except for telephone numbers, dates, times, ordinal numbers and Roman numbers. Some of these issues are caused due to the different ways in which authors express these phenomena in the text. For example, telephone numbers separated with a full stop (e.g. 123.456.789) or in between hyphens (123-45-67-89) were not identified. Similarly, telephone numbers correctly written according to the rules in the original texts were not identified as such, but as quantities. Dates expressed in the format DD.MM.YYYY were not detected and times with the abbreviation h (hours) adjacent to the last numeral character prevented the identification of times (e.g. 07:00h). Text 1043 is a representative example with 22 cases of times expressed in this way but not detected. Similarly, quantities followed by symbols such as €, *km*, etc. prevented the detection of such figures. Roman numbers is the only category with precision below 1. This happened due to the detection as Roman numbers of entities that were not numbers (e.g. the abbreviation *CC*, "Centro Comercial", meaning "shopping centre" was identified as a Roman number). This could be counteracted with the dictionary covering abbreviations, which is a step we will potentially take in the near future.

Concerning the transformation of the categories (see Table 3), all of them perform correctly (e.g. 1) except one quantity that is not rounded (e.g. 1.125) and 16 Roman numbers that are transformed in an incorrect way (e.g. Jaime II as Jaime 2 instead of Jaime segundo). Both of these are system errors and

| Category | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| Figures | 92.96 | 1 | 92.96 | 96.35 |
| Rounding quantities | 98.03 | 1 | 98.03 | 99.01 |
| Explain big numbers | 96.62 | 1 | 92.62 | 98.28 |
| Dates | 0.5 | 1 | 0.5 | 66.66 |
| Times | 60.61 | 1 | 60.61 | 75.48 |
| Percentages | 1 | 1 | 1 | 1 |
| Ordinal numbers | 66.30 | 1 | 66.30 | 79.73 |
| Roman numbers | 70.66 | 70.66 | 1 | 82.81 |
| Telephone numbers | 16.66 | 1 | 16.66 | 28.57 |

Table 2: System evaluation, data detection. All the data expressed in percentages.

| Category | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| Figures | 99.88 | 1 | 99.60 | 99.88 |
| Rounding quantities | 1 | 1 | 1 | 1 |
| Explain big numbers | 1 | 1 | 1 | 1 |
| Dates | 1 | 1 | 1 | 1 |
| Times | 1 | 1 | 1 | 1 |
| Percentages | 1 | 1 | 1 | 1 |
| Ordinal numbers | 1 | 1 | 1 | 1 |
| Roman numbers | 69.81 | 1 | 69.81 | 82.22 |
| Telephone numbers | 1 | 1 | 1 | 1 |

Table 3: System evaluation, data transformation. All the data expressed in percentages.

this evaluation will help us fix these issues in a later system version.

Another issue we encountered is the fact that we need more context or meaning to determine if, for instance, "1999", is a quantity or a year. Out of 198 correct roundings, 114 were years and not quantities. Therefore, 57,57% of correct figure transformations are technically not correct with respect to the text. It remains imperative to establish a method for resolving ambiguity in such instances in future meaning and disambiguation modules in the project.

In that regard, previous works already highlight the importance of simplifying taking into account the local context of the sentence (Bautista and Saggion, 2014). For instance, in a context where a comparison is taking place, if rounding is applied, no information will be transmitted. See the example provided by the authors: "The numbers of dissolutions are maintained at 2010 similar to those of 2009, 22,435 versus 21,875, with a slight increase of 2.56%" (Las cifras de disoluciones se mantienen en 2010 similares a las de 2009, 22.435 frente a 21.875, con un ligero incremento del 2,56%.). This case puts in the forefront the fact that regular expressions, which disregard context, have their shortcomings in cases such as the one exemplified. Thus, syntactic awareness is key to avoiding simplification fails. All in all, although limiting the scope of the transformed phenomena, we ensure that the transformations are correct with a rule-based system.

## *8 Conclusions and Future Work*

The main contribution of this research is the implementation of the entirety of Easy-to-Read guidelines dealing with numbers to a rule-based system (i.e. Simple.Text), within the context of the Clear.Text project. The advantages of rule-based systems lie in their precision and ability transform accordingly with a very cost-effective approach. While we value the use of Large Language Models (LLMs), we understand that these are not strictly necessary for clear-cut and well-defined specific tasks. When compared to an LLM, with this approach we gain explainability and resources.

With the rule knowledge that we have gathered for the simplification of numerical entities, we could define two tasks to solve in the future: to identify both (1) the numerical entities and (2) their category in a given text, which directly refers to the transformation that it should undertake for its resolu-

tion and therefore, its simplification. Then, instead of having a set of rules that are limited by its disconnect to context, we could build a corpus to train a machine learning (ML) model that infers these rules. In this way, explainability would not be sacrificed, as many traditional ML models offer explainability.

Overall, we could improve the tool by creating a hybrid system where the detection and classification could be performed with machine learning, deep learning or even BERT, and the transformation phase to be performed with rules, which ensures a precise and accurate transformation. In this way, we would not need a large simplification corpus to train a LLM.

Future work also includes the refinement of the system's current rules, the continuation of the implementation of the entirety of Easy-to-Read guidelines and the evaluation of the system with control and cognitive disabled groups.

More specifically, regarding percentages, there are some exceptions that will be treated using an ad hoc dictionary specifically created for that purpose, for example, "50%" will be replaced by "la mitad" (half), as our collaborators indicate that this construction is easier to comprehend than "5 out of 10". Similarly, fractions will be treated and solved as percentages, that is, with constructions that transmit the same information, for example, "uno de cada tres" (one out of three) instead of "1/3".

Groups of numbers represented in one word, such as "decena" (ten), "docena" (dozen), "millar" (thousand), "centena" (hundred), "centenario/a" (centenarian) or "milenario/a" (millennial), among others, could also be difficult to comprehend. Although these are not explicitly acknowledged in Section 6.2, Rules 19-25 in the Easy-to-Read guidelines (AENOR, 2018), they could be addressed in future work related to numerical expressions using a dictionary.

The resources created by this project will be available on Huggingface[7] and the research group's GitHub[8], as well as the official webpage of the project.

[7] https://huggingface.co/gplsi

[8] https://github.com/gplsi

## *Acknowledgements*

This research was conducted as part of the ClearText project (TED2021- 130707B-I00), funded by the MCIN/AEI/10.13039/501100011033 and the European Union NextGenerationEU/PRTR. Additionally, we acknowledge the collaboration of COOLANG.TRIVIAL: Technological Resources for Intelligent VIral AnaLysis (PID2021-122263OB-C22) funded by MCIN/AEI/10.13039/501100011033/ and by "ERDF A way of making Europe" as well as SOCIALFAIRNESS.SOCIALTRUST: Assessing trustworthiness in digital media (PDC2022-133146-C22) funded by MCIN/AEI/10.13039/501100011033/ and by the "European Union NextGenerationEU/PRTR".

## *References*

AENOR. 2018. Norma española experimental une 153101 ex. lectura fácil: Pautas y recomendaciones para la elaboración de documentos.

Bautista, S., B. Drndarevic, R. Hervás, H. Saggion, and P. Gervás. 2012. Análisis de la simplificación de expresiones numéricas en español mediante un estudio empírico. *Linguamática*, 4(2):27–41.

Bautista, S., R. Hervás, P. Gervás, R. Power, and S. Williams. 2011. How to make numerical information accessible: Experimental identification of simplification strategies. In *Human-Computer Interaction–INTERACT 2011: 13th IFIP TC 13 International Conference, Lisbon, Portugal, September 5-9, 2011, Proceedings, Part I 13*, pages 57–64. Springer.

Bautista, S., R. Hervás, P. Gervás, R. Power, and S. Williams. 2013. A system for the simplification of numerical expressions at different levels of understandability. In L. Rello, H. Saggion, and R. Baeza-Yates, editors, *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, pages 39–48, Atlanta, Georgia, June. Association for Computational Linguistics.

Bautista, S., R. Hervás, P. Gervás, and J. Rojo. 2017. An approach to treat numerical information in the text simplification process. *Universal Access in the Information Society*, 16:85–102.

Bautista, S. and H. Saggion. 2014. Can numerical expressions be simpler? implementation and demostration of a numerical simplification system for Spanish. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 956–962, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Bott, S. and H. 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.

Derczynski, L. 2016. Complementarity, f-score, and nlp evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 261–266.

Drndarević, B. and H. Saggion. 2012. Towards automatic lexical simplification in Spanish: an empirical study. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 8–16.

Pita-Fernández, S. 1996. Determinación del tamaño muestral. *Cadernos de atención primaria*, 3(3):138–141.

Rello, L., S. Bautista, R. Baeza-Yates, P. Gervás, R. Hervás, and H. Saggion. 2013. One half or 50%? an eye-tracking study of number representation readability. In *Human-Computer Interaction–INTERACT 2013: 14th IFIP TC 13 International Conference, Cape Town, South Africa, September 2-6, 2013, Proceedings, Part IV 14*, pages 229–245. Springer.

Vázquez, Y. G., A. F. Orquín, A. M. Guijarro, and S. V. Pérez. 2010. Integración de recursos semánticos basados en WordNet. *Procesamiento del lenguaje natural*, (45):161–168.