An Approach to Treat Numerical Information in the Text Simplification Process

Susana Bautista[1], Raquel Hervás[2], Pablo Gervás[3] and Javier Rojo[4]

[1] Universidad Complutense de Madrid, Spain
Tel.: +34 913947606 -- Fax: +134913947547
subautis@fdi.ucm.es

[2] Universidad Complutense de Madrid, Spain
Tel.: +34 913947599 -- Fax: +134913947547
raquelhb@fdi.ucm.es

[3] Universidad Complutense de Madrid, Spain
Tel.: +34 913947639 -- Fax: +134913947547
pgervas@sip.ucm.es

[4] Aloha Mental Arithmetic, Spain
Tel.: +34634 571 331
tenerife@alohaspain.com

Abstract: Public information services and documents should be accessible to the widest possible readership. In particular, information from these sources often takes the form of numerical expressions, which pose comprehension problems for many people, including people with disabilities, who are often also exposed to poverty, illiteracy, or lack of access to advanced technology. This paper presents an approach to treat numerical information in the text simplification process to make it more accessible. A generic model for automatic text simplification systems is presented, aimed at making documents more accessible to readers with cognitive disabilities. The proposed approach is validated with a real system to simplify numerical expressions in Spanish. This system is then evaluated and the results show that it is appropriate for the task at hand.

Keywords: Accessibility, Numerical Information, Adaptation and Text Simplification

1. Introduction

As a result of the development of ``Information Technology Society'', there is a tendency to digitalize all kinds of information, such as recipes, payslips, news, etc, with the aim of making them more accessible to the users. However, studies show that we are still far away from the ideal of a uniformly digitalized society where information is accessible to everyone. Information often takes the form of numerical expressions (e.g. economic statistics, demographic data), which pose comprehension problems for many people. These include people with disabilities who are often also afflicted by poverty, illiteracy, and lack of access to advanced technology, or people with limited education.

The way in which information is written or presented can exclude many people, especially those who have problems to read, write or understand. One solution would be to simplify these texts so as to adapt them to the needs of particular groups of readers. However, the process of simplifying texts by hand is extremely time and effort consuming. Therefore, any attempt to automate part of this process can leverage access to information.

Numeracy in the International Survey of Adult Skills[1] is defined as ``the ability to access, use, interpret and communicate mathematical information and ideas, in order to engage in and manage the mathematical demands of a range of situations in adult life''. In terms of the assessment, this involves managing a situation or solving a problem in a real context and responding to mathematical content, information and ideas presented in multiple ways. Clearly, numeracy is in part dependent on literacy skills (reading and writing), tough the tasks in the International Survey of Adult Skills also involve more than applying arithmetical skills to information embedded in text, which was the focus of quantitative literacy in IALS (International Adult Literacy Survey)[2].

Experimental psychology and cognitive neuropsychology have dealt with the study of number processing and calculation over the last decades. Many researchers have studied the cognitive processes that are responsible for number processing and calculation, with the goal of contributing to the improvement of teaching and learning processes. For example, [1], [2] present findings about how the frequency of use of a word or number is an influential variable in the reading process. The mathematical reasoning contributes to the development of the person. This reasoning emerges as a man's need to communicate with others and express aspects of the environment such as counting or measuring [3]. A specific learning difficulty that involves the innate difficulty in learning or understanding numeracy is dyscalculia, which includes difficulties in understanding numbers, manipulating, learning math facts and a number of other symptoms related to counting money, understanding prices or remember dates [4], [5].

In Spain, the Program for the International Assessment of Adult Competencies (PIAAC) survey of adult skills[3] estimated that only 1 in 3 Spanish people are able to comprehend a long text or comparing offers, and about 71.7% adults can read and understand a simple text. In terms of numeracy, only 68.6% adults are able to perform simple mathematical calculations and only 24.5% are able to interpret statistics, graphs or solve complex problems in steps. According to the study, the vast majority of Spanish people have difficulty extracting information from real mathematical situations like tourism packages deals comparison, calculation of the final price of a purchase discount, and interpretation of graphs and statistics.

The results of the survey for Spain are compared to the average results of the OECD (Organization for Economic Co-operation and Development)[4]. Data from a survey of adult skills conducted in 2012 as part of the Program for the International Assessment of Adult Competencies (PIAAC) showed that the overall performance in literacy in Spain is 252 (average score) versus 273 for the OECD, and in numeracy Spain is 246 versus 269 of the OECD. In general, Spain is not as good as OECD average in maths, reading and science follow the data of Program for International Student Assessment (PISA)[5] in 2012.

[1] http://www.oecd.org/site/piaac/surveyofadultskills.htm
[2] http://www.oecd.org/edu/innovation-education/adultliteracy.htm
[3] https://www.mecd.gob.es/inee/Ultimosinformes/PIAAC.html
[4] http://www.oecd.org/statistics/compare-your-country.htm
[5] http://www.oecd.org/pisa/pisaproducts/pisa2012technicalreport.htm

Daily news often contains numerical information and the way such information is presented affects the readability of these texts. It is access to information that is of central interest for this paper, in particular access to numerical information. Numerical expressions are defined as expressions denoting quantities, optionally accompanied by a numerical modifier, such as “more than a quarter” or “almost 97%”, where “more than” and “almost” take the role of numerical modifiers. Such expressions are extremely frequent in the kind of news texts we use in our study.

For example, Figure 1 shows a piece of news, taken from Simplext project corpus [6], and pay attention to the number and variety of numerical expressions used (highlighted):

*AMNISTÍA DENUNCIA TORTURA Y MALOS TRATOS EN **111** PAÍSES*

*Amnistía Internacional ha documentado durante 2010 casos de tortura y otros malos tratos en **al menos 111** países, juicios injustos en **55**, restricciones a la libertad de expresión en **96** y presos de conciencia encarcelados en **48**. Con motivo del Día de los Derechos Humanos, que se celebra el 10 diciembre, activistas de la organización saldrán a la calle en **más de 50** países durante 24 horas para participar en un maratón de recogida de firmas en defensa de los derechos humanos. En España, Amnistía estará presente en **más de 50** ciudades, entre las cuales se encuentran Madrid, Barcelona, Logroño, Valencia, Burgos, Huesca, Guadalajara, Sevilla, Tenerife, Palencia, A Coruña y Palma de Mallorca. El director de Amnistía Internacional en España, Esteban Beltrán, señaló a los periodistas en Madrid que “los derechos humanos son demasiado importantes como para dejarlos en manos de los gobiernos”.*

*AMNESTY REPORTS TORTURE AND ABUSE IN **111** COUNTRIES*

*During 2010, Amnesty International has documented torture and abuse in **at least 111** countries, unfair trials in **55**, restrictions on freedom of expression in **96**, and prisoners of conscience imprisoned in **48**. On the occasion of Human Rights Day, celebrated on 10th of December, activists of the organization will take the streets in **more than 50** countries to participate in a 24-hour signature collection campaign in defense of human rights. In Spain, Amnesty is going to be present in **more than 50** cities, including Madrid, Barcelona, Logroño, Valencia, Burgos, Huesca, Guadalajara, Sevilla, Tenerife, Palencia, A Coruña and Palma de Mallorca. Amnesty International's director in Spain, Esteban Beltrán, pointed out to journalists in Madrid that “human rights are too important to leave them in the hands of governments”.*

***Fig.1** An Example of a piece of news taken from Simplext project corpus*

In a relatively short text composed by five sentences, one can find a total of seven different numerical expressions, which is more than one expression per sentence on average. These include expressions with quantities with or without modifiers, and so on. Such information load, as well as the variety of different numerical expressions, may affect the reader's understanding of the text and prevent him from discovering cause and effect relations of the events presented in the news article.

Building on the difficulties present-day society faces, the authors propose a generic model to carry out the text simplification process. From this generic model, a kind of specific simplification such as numerical expressions simplification is studied and the specific stages of the model within the simplification stage are proposed. In addition, the proposed model is validated with a real system to simplify numerical expressions in Spanish news, analyzing each stage of the model.

The remainder of the paper is organized as follows. An overview of the most relevant work in the field of language accessibility and text simplification, from the point of view of manual and automatic simplifications approaches is presented in Section 2.Section 3 presents the generic model of text simplification, including an instance of the model with a real system to simplify numerical expressions in Spanish texts. Section 4 presents the evaluation of the presented system. The paper concludes with a discussion in Section 5. Conclusions and plans for future work are presented in Section 6.

2. Language Accessibility and Text Simplification

Automatic text simplification is a specific task in natural language processing. The aim is to modify, enhance and adapt an existing text in such a way that the grammar and structure of the prose is greatly simplified, while the underlying meaning and information remains the same. Different transformations at syntactic and lexical level are applied to achieve a simplified version of the original text.

A United Nations report[6] recommends that public information services and documents should be accessible to the widest readership possible. Many different initiatives propose guidelines that may help when rewriting a text to make it more comprehensible. Some of these initiatives are Plain Language[7], the European Guidelines for the Production of Easy-to-Read Information[8], and the latest Web Content Accessibility Guidelines (WCAG 2.0)[9] with a wide range of recommendations for making web content more accessible.

These European Guidelines outline some steps in order to prepare an easy-to-read document. There are two different situations: simplifying an original text or creating a new simplified text. In both cases, it is important to know the aim of the text and the target user. The main steps are presented below:

1. Define the objective of the text.
2. Make a list with the main points to cover in the topic of the text.
3. Prepare a draft of the text.
4. Check with the target user in order to validate the draft to prepare the final version. This test is to correct, improve and finish the best version for them.

There have been various attempts at providing adequate content for readers who present any kind of difficulties through simplification of already existing materials for a specific target group. Such is the case, for example, with Simple English Wikipedia[10] where more than 100,000 articles from the English Wikipedia have been adapted using simple English word and grammar. Another similar effort is Encyclopedia Britannica for Kids[11] where the information is adapted for kids, using different vocabulary and grammar to the classic encyclopedia Britannica. For Spanish, there is a website called *Noticias Fácil*[12] where daily news are shown using a simple language.

[6] http://www.un.org/disabilities/documents/gadocs /standardrules.pdf
[7] http://www.plainlanguage.gov
[8] http://inclusion-europe.org/
[9] http://www.w3.org/TR/WCAG/
[10] http://simple.wikipedia.org/wiki/Main_Page
[11] http://kids.britannica.com/
[12] http://www.noticiasfacil.es/ES/Paginas/index.aspx

There are different organizations and programs in Spain that help promote easy-to-read materials, such as Easy-to-Read Association (Asociación Lectura Fácil)[13] in Barcelona and Live Easy-to-Read (Vive la fácil lectura)[14] in Extremadura. Easy-to-Read initiatives work towards creating plain language versions of legal documents and news texts for institutions and companies that want to improve communication with their target audience. These initiatives also promote the publication of books for people with reading difficulties. When simplifying texts, their content, language, illustrations, and graphic design are taken into consideration.

Nevertheless, manual simplification is too slow and costly to be an efficient way of producing a sufficient amount of desired reading material. There are different simplification tasks, such as syntactic transformations, where the structure of a sentence or a part of it is transformed; or lexical substitutions, where only certain words are modified. That is why numerous attempts have been made at creating automatic or semi-automatic text simplification systems, mainly applied to English [7], [8], but also to Japanese [9], Portuguese [10], and Spanish [6]. Many systems are directly applied to offer simpler reading material for specific target users, such as foreign language learners [11], readers with aphasia [12], low literacy individuals [10], etc.

Automatic text simplification has been directed mainly at two levels: Syntactic constructions, where the structure of a sentence or a part of it is transformed and lexical choices where only certain words are modified. Some readers find difficult texts including long sentences, passives, coordinate and subordinate clauses, abstract words, low frequency words, and abbreviations. The rule-based paradigm has been used in the implementation of some systems [13], [14], [15]. The transformation of texts into easy-to-read versions can also be phrased as a translation problem between two different subsets of language: the original and the easy-to-read version. Corpus-based systems can learn from corpora the simplification operations [16], [17]. For example, Gasperin et al. [18] present a corpus-based approach to selecting sentences that require simplification in the context of Brazilian Portuguese text simplification system. Based on a parallel corpus they apply a binary classifier to decide if a sentence should be simplified. There are also recent data driven and machine learning approaches for text simplification. Zhu et al. [19] present a statistical simplification model covering splitting, dropping, reordering and substitution integrally. Woodsend and Lapata [20] implement a data-driven model based on quasi-synchronous grammar, a formalism that can naturally capture structural mistakes and complex rewrite operations. Klerke and Sogaard [21] have tried an unsupervised approach to automatic text simplification of Danish sentences.

A variety of simplification techniques have been applied to the simplification of lexical choices, like the substitution of common words for uncommon words [22]. Caroll et al. [8] contribute with an additional lexical simplification module, and introduce the paradigm, often repeated thereafter, of simplification based on synonym substitution. They use WordNet [23] to obtain a set of potential synonyms of content words in the input text, and determine the simplest out of the set by looking up Kucera-Francis frequencies in the Oxford Psycholinguistic Database [24]. Word frequency is, therefore, seen as a measure of lexical complexity, and this approach has been adopted in a number of works. Bautista et al. [25] use a similar approach but introduce word length as an additional indicator of word difficulty. De Belder et al. [26] were the first to introduce a word-sense-disambiguation element to their lexical simplification system in order

[13] http://www.lecturafacil.net/content-management-es/
[14] http://www.facillectura.es/

to account for numerous cases of polysemy, especially common among the most frequent words.

With respect to numerical information, some work has been done to address this problem, though mainly targeted at experts who generate easy-to-read numerical information and not individuals experiencing numeracy difficulties [27]. Power and Williams [28] and Bautista et al. [29] were among the first to concentrate on the simplification of such expressions, focusing mainly on the use of numerical modifiers (also termed hedges) as one useful simplification strategy.

Simplification may in some cases entail loss of precision, though this is not necessarily a bad thing, for several reasons.

1. Loss of precision can be signaled linguistically by numerical hedges such as “around”, “more than” and “a little under”, so it need not be misleading.
2. As Krifka has argued, competent writers and speakers frequently approximate numerical information and readers and hearers can readily recognize this, even when no hedge is present, especially when numbers are round [30]. For instance, in “the distance from Oxford to Cambridge is 100 miles” it is clear that 100 miles is an approximation. Williams and Power [31] showed that writers tend to approximate numerical quantities early in a document, then give more precise versions of the same quantities later.
3. As Krifka argues in the same paper [30], an inappropriately high level of precision would flout Grice's Maxim of Quantity [32] by giving too much information. There cannot be many situations in which we need to know that the distance from Oxford to Cambridge is 100.48 miles.
4. As argued by MacKay in his book on sustainable energy [33], simplification brings cognitive benefits, making numbers easier to remember and reason with. So, far from being detrimental, number simplification can have positive advantages for numerate people as well as less numerate.

Power and Williams [28] conducted a study of a news corpus in English, analyzing how authors vary mathematical forms and their precision when expressing numerical information. In a given document, the same quantity was often written in different ways, varying its expression (e.g. fraction or percentage) and its precision through the use of modifiers or rounding. Additionally, they developed a system based on restrictions to decide how to adapt the original proportion. They carried out an evaluation of the model and they found that most given values produced by participants in the survey were predicted by their model, with an overall coverage over 90%; quality as also high. These results supported their assumptions about the composition of the scale systems for fractions and percentages, which determine the range of generated given values. They have given examples of how pragmatic considerations might affect each component of a solution (scale system, given value, arithmetical relation), thus providing a framework for investigating such questions systematically.

The work of Bautista et al. [29] studies the preference for common values when rounding numerical expressions. This study also analyzes the use of different simplification strategies, depending on the value of the original proportion. Their system is built for English, it was not targeted at a particular group of readers, and simplification was applied based on the levels of difficulty as described in Mathematics Education Program of the Qualifications and Curriculum Authority [34]. They carried out a survey in which experts in numeracy were asked to simplify a

range of proportion expressions, with three different readerships in mind. Responses were consistent with their intuitions about how common values are considered simpler and how the value of the original expression influences the chosen simplification.

The Simplext project aims at producing a text simplification system for Spanish. The automatic simplification system is developed following the easy-to-read principles and apply natural language processing techniques. Manual simplification methodology in Simplext follows work by Anula [35], [36] proven to contribute to the reduction of complexity in written language. Two types of simplifications are considered: vocabulary and syntactic structures simplifications. There are various problems addressed in the project form the natural language processing viewpoint: corpus alignment, development of text analysis pipelines to carry out linguistic processing of the corpus and the new documents to be simplified in application time, semi-automatic analysis of the corpus in order to identify potential simplification operations which could be implemented, development of a decision support module to take simplification decisions and implementation of metrics for text readability in the Spanish language as well as development of a text-to-text generation component to produce simple sentences from extracted information chunks. The work presented in this paper is complementary to other automatic text simplification modules which have been proposed for Spanish. In addition, the corpus achieved in the Simplext project is used.

3. A Generic Model of Text Simplification

From the information presented in previous sections, the need to text simplification and the cost and effort of manual simplification to generate simplified versions of original texts is evident. The objective in this section is to present a generic model of automatic text simplification, and to describe its different working stages. More specifically, the mathematical processing is studied and the new stages necessary to carry out the simplification of a specific type of information - numerical expressions are explained.

The following sections present the different stages the proposed model goes through, starting with the original text that is to be simplified, then validating it with an instance for the simplification of numerical expressions. Figure 2 shows the stages of the generic model. There are different variables that determine the configuration of the model at each stage. The language of the original text affects all the stages of the model because it determines which tools can be used to analyze the text and what simplification operations can be applied as these depend on the tools available for the given language. The simplification stage takes into account other variables, such as the type of text (e.g. news article, report, recipe, etc.), the target user and the level of difficulty.

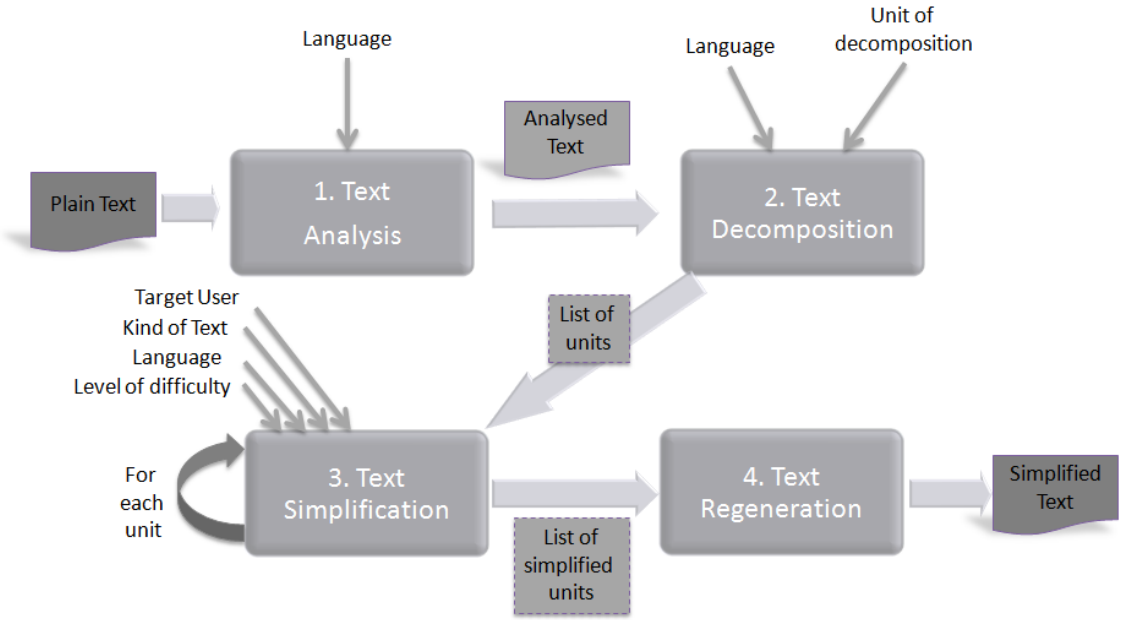


Fig.2. Stages of the Generic Model of Automatic Text Simplification. With plain text as input, the first stage consists in the analysis of the text. At the next stage, text decomposition is applied, which separates the original text into linguistic units. What follows is text simplification, which comprises different operations. Finally, text regeneration takes place and a simplified text is offered as system output. The target user, the kind of text, and the language influence the operation of different stages.

In order to validate the proposed model, an instance of the model for the simplification of numerical expressions[15] is presented. The first variable to be configured is the language of the original text, Spanish in this study, as a real system that simplifies numerical expressions in news articles in Spanish [37], [38] is used. The following sections explain how all stages work and describe the decisions that have to be made at each stage.

Figure 3 illustrates an example from the Simplext Project corpus [6], obtained from the news agency Servimedia[16]. The corpus contained 298 documents with 1149 sentences and 712 numerical expressions. In this example, it can be seen that the text consists of four sentences and contains eleven different numerical expressions. These include expressions with quantities expressed as numerals and percentages, quantities with or without modifiers, and so on.

[15] The actual instances described here as examples have been developed as plugins for GATE to take advantage of existing NLP tools

[16] http://www.servimedia.es/

***EL 10%** DE LOS ADOLESCENTES ESPAÑOLES QUE BEBEN ALCOHOL LO HACEN PARA "COLOCARSE"*

***El 61%** de los españoles de entre **12** y **18** años consumen habitualmente bebidas alcohólicas y de ellos el **10%** lo hacen para "colocarse", según afirmó la Fundación Alcohol y Sociedad, que presentó un libro en el que apuesta por cambiar la prohibición por la educación para abordar el problema de la ingesta alcohólica entre menores. El trabajo, titulado "Hablemos de alcohol. Por un nuevo paradigma en el beber adolescente", señala que pese a que en España está prohibido vender bebidas alcohólicas a los menores - en algunas comunidades hasta los **18** años y en otras hasta los **16** -, **casi el 94%** de ellos aseguran que no tienen problemas para adquirir estos productos. En la misma línea, el texto, presentado por su coordinador, Javier Elzo, sostiene que la mitad de los jóvenes que beben hoy en España dicen emborracharse como mínimo una vez cada dos meses y que **el 69%** se iniciaron en este hábito entre los **13** y los **16** años.*

***10%** OF THE SPANISH ADOLESCENTS WHO DRINK ALCOHOL DO IT TO GET "WASTED"*

***61%** of Spanish people between the ages of **12** and **18** regularly consume alcoholic drinks and of these, **10%** do it to get "wasted" as stated by the Alcohol and Society Foundation, who presented a book, in which it positions itself in favour of turning prohibition into education to deal with the problem of underage drinking. The work, entitled "Let's Talk about Alcohol. For a New Paradigm in Adolescent Consumption", notes that, although it is forbidden to sell alcoholic drinks to minors in Spain - in some regions until they are **18** years of age and in others **16** - **almost 94%** of them assure that they generally do not have problems purchasing these products. Along the same lines, the text, presented by its coordinator, Javier Elzo, maintains that half of the young people in Spain who drink alcohol nowadays declare that they get drunk at least once every two months, and **69%** started drinking between the ages of **13** and **16**.*

***Fig3.** An example of text taken from Simplext project corpus.*

3.1. Stage 1: Text Analysis

Depending on the particular goals to be achieved by the simplification process, different types of analysis of the input text may be required. In general terms, most systems apply basic steps of natural language processing such as: splitting the text into sentences, part-of-speech tagging and syntactic analysis of the text. The input at this stage is a plain text, with its language as a variable that can be configured. It is necessary to know what language we are dealing with in order to determine the tool to be used for tagging individual words.

Tokenization: Each sentence has to be separated into its constituent tokens. This process also involves some difficulties such as contractions in English like "It's blue" (It is) or "She's flu" (She has) or genitives like "Ann's house" and "My sisters' weddings".

Splitting into Sentences: The input text has to be separated into its constituent sentences. There are some obvious problems that occur at this stage. Such is the case of full stops that do not separate sentences (e.g. in abbreviations) or sentences such as titles that do not end with a full stop.

Part-of-Speech Tagging: Every word is assigned its morphological category (e.g. verb, noun, etc.), and sometimes also certain attributes expressed through inflection (e.g. gender, number, tense, etc.). A number of methods can be used to carry out this task, primarily sequential methods, such as Hidden Markov Models, dependency trees and regular grammars (such as

finite-state transducer cascades, a method used for symbolic processing). There are different tools that implement these methods to perform this task.

Syntactic Analysis: A syntactic analysis of the text tagged with the morphological category of its words is usually carried out. There are two basic types of syntactic analysis. One is the analysis of syntactic constituents, which divides the sentence into phrases, which in turn are recursively composed of smaller phrases. Such analysis can be represented in the form of trees with non-terminal nodes. The other type of syntactic analysis is the analysis of syntactic dependencies, which establishes dependencies between words. This analysis can also be represented in the form of trees, but these only contain terminal nodes. Natural language syntax is the way in which individual words are combined to form more complex units. On the one hand, syntax defines which sentences are grammatical and which are not, and on the other, it influences the propositional semantic interpretation.

In the particular instance of the model for Spanish, part-of-speech tagging and syntactic analysis of the text chosen have to be carried out. The tool used for part-of-speech tagging in order to assign corresponding morphological categories to all the words in the text was FreeLing [39], an open source language analysis tool suite. The output of this stage is a list of sentences where every word is tagged with its morphological category.

FreeLing uses a Hidden Markov Approach, producing an EAGLES tag[17] for each word in the document. Since the interest is mainly in numerical information, the focus is on tags of type Z which are allocated to quantities, ratios, fractions, percentages, etc. Four different kinds of Z tags are identified:

1. Partitive numerals which have the tag Zd (for example, “un millón” (“a million”) or “una centena” (“a hundred”)).
2. Monetary expressions having the tag Zm, their lemma being the quantity and the monetary unit (for example, “2000 dólares” (“2000 dollars”), where the lemma is $USD: 2000).
3. Fractions and percentages, which are allocated the tag Zp, where the lemma is a normalized proportion (for example, 34% with the lemma 34/100).
4. Physical measures having the tag Zu. Their lemma is a normalized notation of the unit and the quantity (for example, 30km/h with the lemma SP km/h:30).

To illustrate an example, let us have a look at a pair of numerical expressions belonging to sentences of a text from the corpus (in Spanish): ``...el 65% de los atentados...'' and `` ... va desde los 46 a los 55 años...''. Table 1 shows the tags obtained, where the first column is the word itself, the second column is its lemma, the third is the morphological category tag assigned by FreeLing, and the fourth column is the probability of the word having the specified tag.

[17] http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html

Table 1. An example of the Part-of-Speech obtained by FreeLing.

Word	Lemma	Tag	Frequency
El	el	DA0MS0	1
65%	65/100	Zp	1
De	de	SPS00	0.999919
Los	el	DA0MP0	0.97623
atentados	atentado	NCMP000	0.75
Va	ir	VMIP3S0	1
Desde	desde	SPS00	1
Los	el	DA0MP0	0.97623
46	46	Z	1
A	a	SPS00	0.99585
Los	el	DA0MP0	0.97623
55	55	Z	1
Años	año	NCMP000	1

At this stage, syntactic analysis is applied to the whole text, thus obtaining an analyzed text that is to be simplified at the next stage. For example, Table 2 shows how a part of a sentence, ``...el 65% de los atentados...'', was analyzed by FreeLing. Each word has its tags with different features like identifier, span, lemma, form, lexical information, etc.

Table 2. Example of part of the XML file.

```
<word
id=”179” span=”993-995” guess=”no”
recov=”el” tag=”DA0MS0” lemma=”el”
form= “el”>
el
</word>
<word
id=”180” span=”995-999” guess=”yes”
recov=”65%” tag=”Zp” lemma=”65/100”
form=”65%”>
65%
</word>
<word
id=”181” span=”1000-1002” guess=”no”
recov=”de” tag=”SPS00” lemma=”de”
form=”de”>
de
</word>
<word
id=”182” span=”1003-1006” guess=”no”
recov=”los” tag=”DA0MP0” lemma=”el”
form=”los”>
los
</word>
<word
id=”183” span=”1007-1016” guess=”no”
recov=”atentados” tag=”NCMP000”
lemma=”atentado”
form=”atentados”>
atentados
</word>
```

3.2. Stage 2: Text Decomposition

The second stage consists of decomposing the text into the linguistic units that are to be the target of the simplification process, such as words, sentences, or paragraphs. Although this is a fairly simple task, it is of paramount importance for the entire chain of further processing. This stage may also involve a task of selection: selection of specific parts of the text that are going to be the subject of simplification, and their separation from other parts of the text that will not be affected.

For instance, if the goal of the simplification process is to replace difficult words with simpler ones, at this stage, only difficult words will be picked out of the input text as target units for simplification. Certain types of simplification may involve the elimination or insertion of information - such as graphical depictions of the content, or dictionary definitions of difficult words. In those cases, the stage of Text Decomposition would have to identify the particular elements to be eliminated or the specific points of the original text where information has to be inserted.

In this stage, the input is the result of the syntactic analysis of a Spanish text. It is necessary to know the unit of decomposition considered in order to configure this stage. The obtained output is a list of linguistic units - e.g.sentences, words, numerical expressions - that make up the original text that is to be simplified.

In the current instance of the model, the text is decomposed into sentences, beginning with the headline and continuing with the sentences in the body of the text. It can be seen that the text (shown in Figure 3) consists of four sentences, the headline and the three sentences that make up the body of the text, separated by full stops. For each sentence, one has to find and collect the target units for the simplification process, in this case, numerical expressions. Two different stages are involved, one relying on the results of the lexical-syntactic analysis to identify candidate expressions, and another one to annotate the numerical expressions with the information that the simplification process will require. Figure 4 presents the specific steps within the text decomposition and text simplification stages.

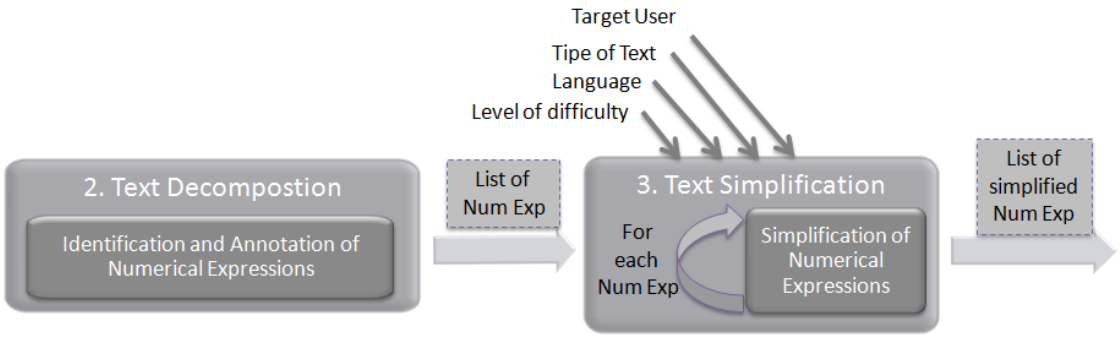


Fig. 4. Stages of the automatic text simplification model with a focus on numerical information. At this stage different operations are applied: identification of the different types of numerical expressions; annotation of numerical expressions with information pertaining to modifier, quantity, unit, etc; and the simplification of numerical expressions through the application of defined rules.

3.2.1. Stage 2.1: Identification and Annotation of Numerical Expressions

At this stage extra information of the different types of numerical expressions are identified and annotated in the original text from the text analysis made in the previous stages, sentence by sentence. Numerical information can be expressed in different mathematical forms, such as percentage ("*78%*"), fraction ("*1/4*"), ratio ("*four in six*"), etc. Numerical information can also be found in non-numerical expressions, such as "*almost everybody"*, "*during the entire day"*, etc. That is why this stage is important - it is important to identify the numerical load of a text so as to be able to modify it. Depending on the language of the original text, the target user for which text adaptation is being applied, and the tools used at the preceding analysis stages, numerical expressions can be identified in different ways, starting with morphological tags assigned to words through part-of-speech tagging, the use of grammars, rules, etc.

In the present case, the analysis carried out by means of FreeLing allows to identify different numerical expressions tagged according to the numeral classification system of the parser. For the example presented, the sentences being considered contain 11 different numerical expressions, with or without modifiers. From each sentence we focus on different kind of numerical expressions.

The process of information extraction consists of identifying the different components of the numerical expressions, including modifiers, such as "*almost*", "*more than"*, " *around"*, etc.; the quantity ("*500*", "*23%*", "*6.7*"); the type of mathematical representation, such as percentage, fraction, exact number, decimal number, ratio, etc.); the use of measurement units, such as "*km*", "*litre*", "*millions of euros*", etc.; and the kind of numerical expression. Having identified the different component of the numerical expressions, one can then choose the corresponding simplification operations at the next stage.

For the purpose of annotating the expressions, JAPE (Java Annotation Patterns Engine) grammars [40] have been used. JAPE applies finite state transduction over annotations based on regular expressions. This is how the information present in the different types of numerical expressions identified by the parser is annotated. The grammars used at this stage of the prototype are explained in depth in Bautista et al. [37].

To illustrate an example, let us look at the expression of the example text: the expression "casi el 94%" ("almost 94%") bears the tag "Zp", which means that it is a percentage, and its modifier "casi" ("almost") and the quantity "94%" can be further annotated. The parser is responsible for calculating other parameters such as the lemma or the units. Once all the numerical expressions in the text have been annotated, Text Simplification follows.

The output of this stage is a list of numerical expressions.

3.3. Stage 3: Text Simplification

This stage starts with a list of linguistic units from the text obtained at the previous stage, and simplification techniques are applied to each individual linguistic unit. Simplification operations can be applied to the whole unit or some of its parts.

This stage still depends on language though it is important to bear in mind other types of factors, such as the type of text, the target user and the level of difficulty, which determine what simplification operations are to be applied. There are different simplification tasks, such as

syntactic transformations, where the structure of a sentence or a part of it is transformed; or lexical substitutions, where only certain words are modified. This way, a series of transformations is applied to every linguistic unit of the original text to obtain a simplified unit. When the intended simplification requires the insertion of additional information, this stage would have to produce the required new information to be placed in the position already staked out by Stage 2 of Text Decomposition. Context and text discourse would have to be taken into consideration in this case.

The output at this stage is a list of simplified linguistic units, obtained as a result of applying different transformations to each sentence of the original text.

In the current instance of the model, the text simplification stage starts with a list of numerical expressions that have been previously identified and annotated. The working language is Spanish, the type of text is news articles, and the target user is a person with cognitive problems and reading difficulties that arise when trying to understand mathematical concepts.

Starting with the information annotated and obtained at the previous stages, it is time to simplify each numerical expression from the original text. A set of specifically defined rules for the simplification of numerical expressions are applied. These rules vary depending on the type of the numerical expression that is being treated and the type of text that is being simplified. These variables determine the level of difficulty at which the simplified text should be, because adaptations at different levels can be considered depending on the grade of mathematical concept and the simplification chosen.

The simplification rules apply to modifiers, the numerical quantity, and the measurement units. The rules were defined based on previous studies with experts with the aim of developing a set of simplification strategies (see [41]). For example, a survey was carried out with numeracy experts who were asked to simplify a range of different numerical expressions. From these surveys candidate rewriting strategies were collected and guidelines for defining rules for the prototype were obtained. The simplification strategy which is most commonly observed has been implemented in the proposed system: the quantity is always rounded and a set of rules is applied to choose a modifier which accounts for loss of precision. In order to obtain the rounded number corresponding to the original quantity, auxiliary calculations using different methods of the package Math of Java are used (*round, pow, rint*). For example, if the original value of the quantity is “0.953”, its rounded value is calculated “1.0” for the simplified expression.

In addition, the use of numerical modifiers in the process of rewriting difficult numerical expressions in simpler ways was analyzed. Modifiers indicate that the original number has been approximated and, in some cases, the direction of approximation. The details of the use of modifiers in the process of simplifying numerical expressions can be found in Bautista and Saggion [41]. For example, if the original expression contains a modifier, it is kept and the quantity is rounded. For the rest of the cases, the system compares the original quantity to the rounded quantity, and depending on both values, a modifier is selected. The modifiers chosen are the most commonly used ones in the survey. When the original quantity is already a rounded quantity, there are no modifications in the numerical expression. Table 3 summarizes the selection of modifiers for the simplified numerical expression.

Table 3. Selection of modifier for the simplified numerical expression in different cases. Each case is illustrated by an example.

Original Expression	Rounded Quantity	Case	Modifier	Simplified Expression
Alrededor de 5689 millones (around 5689 million)	6000 millones (6000 million)	There is a modifier in the original expression	Original modifier is kept	*Alrededor de 6000 millones*
27.3%	25%	No modifier in original expression, and original > rounded	Add modifier "más de" (*more than)*	*Más de 25% (more than 25%)*
476	500	No modifier in original expression, and original < rounded	Add modifier "casi" (*almost*)	*Casi 500 (almost 500)*
3000	3000	No modifier in original expression, and original = rounded	No modifier	3000

The simplification rules applied by the system can be summarized from the conclusions obtained in the experimental identification of simplification strategies carried out with experts.

1. The expressions represented in words are transformed into digits.
2. If the original expression has a modifier, then it is kept and the quantity is rounded.
3. If the original expression does not have modifier, then a modifier is chosen (Table 3) and the quantity is rounded.

In this stage each numerical expressions is simplified, and as a result a list of simplified numerical expressions are obtained which will be used in the next stage.

3.4. Stage 4: Text Regeneration

At this final stage, the only thing left to do is to recompose the text, either by putting together the simplified versions of the linguistic units that are the output of all the previous stages, or, if there was a selection process involved during decomposition, using the simplified versions of the target units in combination with the rest of the input text to reconstruct a whole simplified version. A simplified text is thus obtained as the final output of the system.

In the presented instance of the model, the text is recomposed based on the list of simplified numerical expressions from the previous stage. The output of the model is, therefore, a simplified version of the original text. For each sentence, all numerical expressions have been identified and if there was a simplified version obtained in the process of simplification, in this moment the original expression in the sentence is replaced by the simplified one. Figure 5 presents the output of the system after applying the model for the simplification of numerical expressions to the original example text presented in Section 3.

***EL 10%** DE LOS ADOLESCENTES ESPAÑOLES QUE BEBEN ALCOHOL LO HACEN PARA ``COLOCARSE''*

***Más de 60%** de los españoles de entre **más de 10** y **casi 20** años consumen habitualmente bebidas alcohólicas y de ellos **el 10%** lo hacen para "colocarse", según afirmó la Fundación Alcohol y Sociedad, que presentó un libro en el que apuesta por cambiar la prohibición por la educación para abordar el problema de la ingesta alcohólica entre menores. El trabajo, titulado "Hablemos de alcohol. Por un nuevo paradigma en el beber adolescente", señala que pese a que en España está prohibido vender bebidas alcohólicas a los menores - en algunas comunidades hasta los **casi 20** años y en otras hasta los **casi 20**-, **más de 90%** de ellos aseguran que no tienen problemas para adquirir estos productos. En la misma línea, el texto, presentado por su coordinador, Javier Elzo, sostiene que la mitad de los jóvenes que beben hoy en España dicen emborracharse como mínimo una vez cada 2 meses y que el **casi 70%** se iniciaron en este hábito entre los **más de 10** y los **casi 20** años.*

***10%** OF THE SPANISH ADOLESCENTS WHO DRINK ALCOHOL DO IT TO GET "WASTED"*

***More than 60%** of Spanish people between the ages of **more than 10** and **almost 20** regularly consume alcoholic drinks and of these, **10%** do it to get "wasted" as stated by the Alcohol and Society Foundation, who presented a book, in which it positions itself in favour of turning prohibition into education to deal with the problem of underage drinking. The work, entitled "Let's Talk about Alcohol. For a New Paradigm in Adolescent Consumption"', notes that, although it is forbidden to sell alcoholic drinks to minors in Spain - in some regions until they are **almost 20** years of age and in others **almost 20** -**more than 90%** of them assure that they generally do not have problems purchasing these products. Along the same lines, the text, presented by its coordinator, Javier Elzo, maintains that half of the young people in Spain who drink alcohol nowadays declare that they get drunk at least once every two months, and **almost 70%** started drinking between the ages of **more than 10** and **almost 20**.*

***Fig.5.** Output of the system where the numerical expressions have been simplified.*

3.5. Combining Several Simplification Approaches

In some cases, it may be necessary to combine more than one approach to simplification to achieve the desired result. For instance, replacing difficult words with easier ones may be combined with rewriting complex syntactic constructions into simpler ones. In these cases, each different approach requires a different instantiation of Stage 2 - to identify and select the target units for the particular approach - and Stage 3 - to apply the particular transformation required in each case.

Combinations of radically different approaches - for instance, when summarization techniques based on the extraction of complete sentences are combined with lexical or syntactic simplification within the sentences - may also require different instantiations of Stage 1 - text analysis.

In all cases, a complex instantiation of Stage 4 - Text Regeneration - is required to integrate together the results of the various approaches that have been applied into a single cohesive text.

4. Evaluation

The system for the simplification of numerical expressions in Spanish, which has been used as instance of the proposed model was evaluated.

First, the performance of the rules was tested. The developed rules were applied to a subset of 10 unseen documents form the Simplext corpus comprising 59 sentences. The automatic annotations produced by the system were then corrected in order to obtain a gold standard dataset for evaluation using the GATE tool. From a total of 57 tags used, only 6 tags have been improved, which corresponds to a 10.52% of error from our set of rules. A precision of 0.94, a recall of 0.93 and an F-measure of 0.93, were obtained which were considered quite acceptable. For numerical expressions with low frequency of occurrence, the results are worse but they are better for frequently observed numerical expressions.

In addition, the linguistic accuracy of the output was analyzed, and the results were positively rated, with 83.56% (almost 84%) of the simplified sentences considered correct, where all containing numerical expressions. The meaning was also preserved reasonably well in the process of simplification. The corpus used for the evaluation had 57 texts with 73 sentences. This corpus is part of the Simplext corpus however it does not correspond to the subset of the corpus used for developing the rules. One of the authors of this paper was in charge of rating the linguistic accuracy of the output.

The qualitative analysis of the results revealed that the most common error that results in poor correction of the output sentence was bad treatment of comparative numerical expressions. In this case, errors were found in 4 different sentences where the treatment of comparative numerical expressions was incorrect. For example, for this original sentence "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%." ("Bankruptcy figures in 2010 are similar to those of 2009: **22,435** versus **21,875**, a slight increase of 2.56%"), the output of the system was "Las cifras de disoluciones se mantienen en 2010 similares a las de 2009, **más de 20.000** frente a **más de 20.000**, con un ligero incremento del casi 3%" ("Bankruptcy figures in 2010 are similar to those of 2009: **more than 20,000** versus **more than 20,000**, a slight increase of almost 3%''). This kind of error can also be seen in the text used such as example, in the part of the sentence "en algunas comunidades hasta los **casi 20** años y en otras hasta los **casi 20"** ("in some regions until they are **almost 20** years of age and in others **almost 20**"). Here, it can be seen that the meaning preservation of the original sentence is lost. To measure the meaning preservation the evaluator compared the original and the simplified sentences using her own judgment. Noting such errors, we plan as a system upgrade some kind of a posteriori filtering to rule out bad simplifications is planned as a system upgrade.

Finally, in order to evaluate the output of the system, experts, primary and secondary school teachers who work daily with pupils who need adaptations participated in the evaluation. In addition, these experts are trained in pedagogy and curriculum adaptation, so they are a group especially suitable for the evaluation of the system.

To carry out this evaluation, a questionnaire was designed using Google Forms[18], which easily allows to create forms and collect responses to the questions posed. Participants were presented with 15 pairs of sentences (original and simplified by the system)[19] with 34 different

[18] https://docs.google.com/forms/d/1cfISwwcUGdZBI9XjnvlgYLG0Sp9tcrfjw5cT2Jv8veo/viewform

[19] The 15 pairs of sentences are presented in the Annex at the end of the paper

types of numerical expressions (numerals, partitives, percentages and monetary quantities), with an average of 33.5 words per sentence and a mean of 2.26 numerical expressions per sentence. The modality of answers is yes/no answer and for each question, three things were asked:

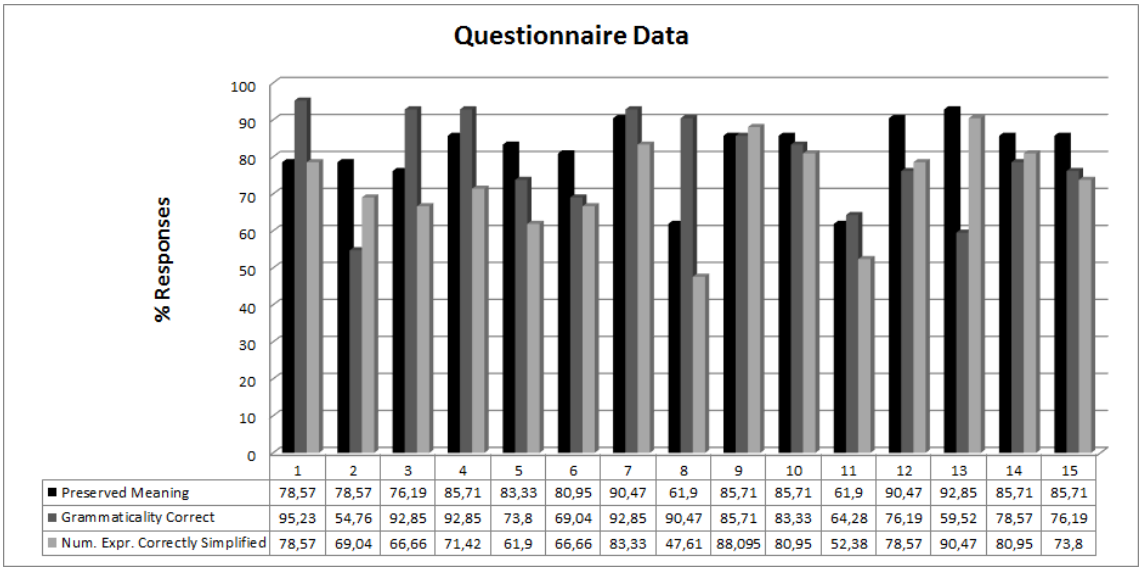
1. If the simplified sentence preserves the meaning of the original.
2. If the simplified sentence is grammatically correct.
3. If numerical expressions are correctly simplified.

A total of 42 experts participated in the evaluation, 11 men and 32 women, with ages ranging between 18 and more than 45 years old. In total, there were 15 participants between 18 and 25 years, 19 between 25 and 35 and 7 participants between 35 and 45. Finally, there was only one participant older than 45 years old. All participants were native Spanish speakers volunteer teachers from different primary or secondary schools in Spain. They were contacted by email and were sent the form online to collect the data.

The data collected in the questionnaire was analyzed, and the results showed that the participants considered that the simplified version of the sentence preserved the meaning in comparison to the original sentence with a mean of 81.58% and a standard deviation of 9.24%. In addition, they believed that the sentence with simplified numerical expressions was grammatically correct with a mean of 79.04% and a standard deviation of 12.98%. Finally, they considered that numerical expressions have been correctly simplified with a mean of 72.69% and a standard deviation of 12.3%.

For each sample proportion, statistical inference was carried out by constructing the 95% confidence intervals (CI). The amplitude of CI obtained depends on the sample percentage itself and on the sample size that supports it. As can be seen, for the first question a CI [78.6%, 84.6%] with a standard error of 1.5% was achieved. As long as value 50% is not covered by CI, this data confirm that the number of Yes answers is statistically higher than the number of No answers. This supports the hypothesis that the simplified sentence preserves the meaning of the original according to the participants' opinion. For the second question a CI [75.9%, 82.2%] with a standard error 1.6% was achieved. It is the same case and this data confirm that the simplified sentence is grammatically correct. For the last question, a CI [69.2%, 76.2%] with a standard error 1.8% was achieved. It can be seen that the number of Yes answers is statistically higher than the number of No answers, confirming that the numerical expressions are correctly simplified according to the participants.

Figure 6 illustrates the data for each of the sentences of the questionnaire.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Preserved Meaning	78,57	78,57	76,19	85,71	83,33	80,95	90,47	61,9	85,71	85,71	61,9	90,47	92,85	85,71	85,71
Grammaticality Correct	95,23	54,76	92,85	92,85	73,8	69,04	92,85	90,47	85,71	83,33	64,28	76,19	59,52	78,57	76,19
Num. Expr. Correctly Simplified	78,57	69,04	66,66	71,42	61,9	66,66	83,33	47,61	88,095	80,95	52,38	78,57	90,47	80,95	73,8

Fig. 6. Data collected for each of the sentences of the questionnaire used in the evaluation with experts to assess the simplification of numerical expressions performed by the system.

5. Discussion

This section discusses the results of the present study against the previous work in English where simplification strategies were also considered. In addition, the proposed model is discussed in comparison to a previous architecture for text simplification, and the validity of the presented model as an abstraction of existing procedures for text simplification is considered.

5.1. Simplification strategies in different languages

If the simplification strategies obtained in the authors' previous work for English [42] are compared to the strategies presented in this work, one can see some features in common.

For English, only one kind of numerical expressions were considered, while in Spanish a wider range of types of expressions are considered to be simplified.

For English it can be observed that no numerical expressions were used in extreme range, while in Spanish no-numerical information is not used. For example, the original expression *95.3%* is simplified in English by *around all,* while in Spanish is simplified by *around 95%*.

In English the fractions can be simplified by other fractions similar to the original though in Spanish the system does not calculate candidates to simplify the original fractions.

Both systems, for English and for Spanish, follow the European guideline where the expressions in words are transformed into digits. The percentages are rounded to the nearest integer value. The use of modifiers is similar in both systems.

5.2. Comparison of our model and a previous architecture for text simplification

It can be seen that the simplification system architecture presented in the work of Siddharthan [43] and the simplification generic model presented in this paper follow the same idea to generate the simplified version of an original text. The following describe the similarities and differences between the two proposals.

Siddharthan's work proposes an architecture consisting on three stages: analysis, transformation and regeneration. The first state provides the structural representation of a sentence and its part-of-speech tagging. The second stage uses transformation rules to generate plain text from the structure obtained by the previous state. The third and final state is responsible for performing syntactic simplifications referred to in each case.

Instead, the generic model presented in this work consists of four phases or stages: analysis, decomposition, simplification and text regeneration. Although this last phase is called as the third stage in Siddharthan's architecture, functionality is not the same, since different operations are performed in one state and the other. The first stage of the model proposed in this paper is responsible for the text analysis at syntactic level and part-of-speech tagging. The second stage decomposes the text identifying the linguistic units that are to be simplified. The third stage is where simplification rules are applied to generate simplified versions of the identified units. Finally, the regeneration state is responsible for reconstructing the text with simplified versions of the treated units to generate the final simplified text.

Comparing the two approaches, one can see that the initial state of analysis is common to both. The next state is different in both cases. In Siddharthan's architecture it is to generate plain text from the structures obtained in the analysis, while in the model proposed in this paper the second state corresponds to the identification of linguistic units to be simplified. The third state, where properly performed simplification transformations are performed, is called regeneration in Siddharthan's architecture while in the model it is called text simplification. The idea is the same in both cases, since they involve the simplification transformations according to some rules. Furthermore, the model proposed in this paper provides one more state where the simplified text units are restored to generate the final simplified version of the text.

5.3. The Model as an Abstraction of Existing Practice

The generic model of text simplification presented in this paper is intended as an abstraction that aims to cover a number of procedures being followed in practice by simplification systems already in existence.

For example, the system presented by Carroll et al. [8] in order to assist aphasic readers automatically simplifies English newspaper texts as available on the Internet. The system can roughly be divided into two main components: an analyzer component which provides a lexical tagger, a morphological analyzer and parser, and a simplifier component which subsequently adapts the output of the analyzer to aid readability for aphasic people using lexical and syntactical transformations.

In terms of the model described in this paper, the analyzer component would correspond to Stage 1 of Text Analysis. Lexical transformation would correspond to a particular instantiation of Stage 2 - Text Decomposition - to produce particular difficult words as target units, and a

particular instantiation of Stage 3 - Text Simplification - that applies substitutions to these words to result in simpler alternatives for these difficult words.

Syntactical transformations would correspond to a particular instantiation of Stage 2 - Text Decomposition - to produce particular syntactic constructions as target units, and a particular instantiation of Stage 3 - Text Simplification - that applies transformations to these syntactic constructions to result in simpler formulations.

For both instantiations, a final process of reconstructing the complete version of the simplified text corresponds to Stage 4 - Text Regeneration, as described in section 3.5.

Further systems can be analyzed in a similar way. In the Simplext project [6] the text is analyzed using FreeLing [39] and GATE [40], which can be mapped onto Stage 1 of the model. Subsequent application of lexical and syntactical transformations can be considered as instantiations of Stages 2 and 3 as described above.

The PorSimples project [10] developed tools for Brazilian Portuguese and aims at developing technologies to make access to information easier for low-literacy individuals. This approach establishes that text simplification can be subdivided into syntactic simplification, lexical simplification, automatic summarization and other techniques. This proliferation of operations can be seen as the integration of several instantiations of the generic model presented in this paper, with different types of simplification operations being applied at different levels of decomposition granularity (summarization at the level of the complete text, syntactic rewriting at the level of syntactic constructions, word substitution at the level of lexical terms). The regeneration stage would be able to solve possible conflicts between different modules proposing changes for the same text segment. Conflict resolution could be decided by applying some rules of priority or some kind of refereeing.

There are other types of systems which would not fit directly into the general model presented here. For example, systems which allow for some global optimization, such as integer linear programming. They use synchronous grammar that combines a manually constructed grammar for syntactic rules and an automatically acquired grammar for lexical rules and paraphrase [20], [44], [45], [46].

This analysis could be extended to other systems mentioned in Section 2. This work has shown how three different simplification systems for several languages can be described in terms of the proposed generic model of text simplification and other specific systems would not fit directly into the general model. This can be taken as an indication of a certain degree of generality which may help to improve comparability across different systems. In each particular case, the language on which it will operate, the tools to be used, the kind of text and the target user have to be defined. Over these, each system applies its analysis and depending on the objective of the system, it defines its specific simplification transformations.

6. Conclusions and Future Work

The changes in the ''Information Technology Society'' lead to considering changes in the treatment and processing of the information. For example, manual text simplification cannot cope with the process of content adaptation for diverse audiences as it requires a lot of time and effort. This reality was a motive to take advantage of technological solutions to help improve the access to information for people with special difficulties.

Numerical information processing plays a fundamental role in our lives due to numerical expressions presented in different contexts such as news, recipes, etc. Studies of OECD[20] analyze the performance in numeracy as a special case of study in order to identify the problems with mathematics or situations with numerical data. The main motivation to automatize the numerical expressions simplification process is the difficulty that certain people may have to understand this kind of information in texts.

This paper defined a computational model to carry out the automatic simplification of texts. It has focused on the treatment of numerical information as a special case of study and presented the validation of the proposed model with a real simplification system focusing on numerical expressions in Spanish texts.

In the generic model different stages in the automatic text simplification process have been presented. A special simplification stage was considered, where a range of different transformations can be taken into account. They depend on the language, type of the original text and the target user for whom the text is being adapted. These transformations are applied to sentences though they could be considered at a higher level. Possible simplifications include different kinds of operations such as summaries, paraphrases, addition or deletion of information, avoiding of metaphor, sarcasm or ironies, etc. To automatize this kind of operations, other kinds of variables such as context, semantic, etc. may be considered.

With the intent of focusing on the simplification of numerical information in texts, the stage of text simplification is redefined in the model with a focus on numerical expressions. The language of the original text influences the simplification process, as it affects the identification and annotation of numerical expressions in the text decomposition stage. Both stages are dependent on the information provided by the previous stage of text analysis. This analysis is the base for recognizing numerical expressions in the text, and for the annotation of information like modifiers, units or quantities. In the approach presented, it is possible to map syntactic and semantic levels in transformation however not to map the pragmatic level of transformation. However, in previous work [37] the impact of the context has been explored. In order to improve the proposed system, it is necessary to consider this level in future implementations.

The automatic text simplification model has been validated with a focus on numerical expressions in an empirical way, with an already implemented system for the simplification of numerical expressions in Spanish texts. In this instance the different decisions required for each stage of the model have been shown. These kinds of decisions are the ones to be considered in future implementations of other systems.

As part of future work the authors would like to validate the model proposed in this paper, with other real systems. This would require making new instances of the model with the required tools and defining rules for simplifications based on hypotheses about the use of simplification techniques. The proposed model depends on a variety of factors, such as the language of the original text, the kind of the text, the target user for whom the text is being adapted and the level of difficulty desired for the simplified text.

Other operations can be taken into account to apply at the simplification stage, such as adding information to explain certain concepts. This can be done through the use of a dictionary and the insertion of definitions of given terms. It can be argued that this might improve text comprehension. Adding graphic representations of numerical expressions could also prove

[20] http://www.oecd.org/

helpful. These representations would help understand the mathematical meaning of the given numerical expression through the use of images, graphs, etc. As an alternative to textual simplification, the possibility of adding multimedia information, such as video or audio content, is considered as a way of helping the target user to read and understand the original text.

In the approach presented in this paper, experts have been used to determine that the simplification is done correctly. In future work, the plan is to evaluate the system with disabled users in order to know if the transformations actually allow people with cognitive disabilities to understand numerical information that has previously been incomprehensible. In addition, it would be interesting to classify different rules to make adaptations to different types of target users.

It is possible to achieve universal accessibility that takes into consideration affordable devices, technology, cultural issues, and illiteracy. It is crucial to continue working towards designing for diversity, because diversity is where growth and greatness lie, and user-centered design should be the main aim for universal accessibility.

Acknowledgment: This research is funded by the Spanish Ministry of Education and Science (TIN2009-14659-C03-01 Project), and the FPI grant program. The authors would like to thank Ricardo García for his help in this work.

References

[1] Herrera, A., Macizo, P.: ¿Cómo leemos los números? (How we read numbers?). Ciencia Cognitiva 6(2) (2012) 44-47

[2] Salguero, M., Alameda, J.: El procesamiento de los números y sus implicaciones educativas (Number processing and its educational implications). XXI Revista de Educación (Education Journal) 5 (2003) 181-189

[3] Piaget, J., Inhelder, B.: Psicología del niño. Editorial Morata. (1969)

[4] Butterworth, B.: Foundational numerical capacities and the origins of dyscalculia. Trends in Cognitive Sciences 14(12) (2010) 534-541

[5] Landerl, K., Bevan, A., Butterworth, B., et al.: Developmental dyscalculia and basic numerical capacities: A study of 8{9-year-old students. Cognition 93(2) (2004) 99-125

[6] Saggion, H., Gómez-Martínez, E., Etayo, E., Anula, A., Bourg, L.: Text Simplification in Simplext: Making Text More Accessible. Procesamiento del Lenguaje Natural 47 (2011)

[7] Medero, J., Ostendorf, M.: Identifying targets for syntactic simplification. In: Proceedings of Speech and Language Technology in Education. (2011)

[8] Carroll, J., Minnen, G., Canning, Y., Devlin, S., Tait, J.: Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In: AAAI-98. (1998)

[9] Inui, K., Fujita, A., Takahashi, T., Iida, R., Iwakura, T.: Text simplification for reading assistance: A project note. In: Workshop on Paraphrasing. (2003)

[10] Specia, L.: Translating from Complex to Simplified Sentences. In: 9th International Conference on Computational Processing of the Portuguese Language. (2010)

[11] Burstein, J., Shore, J., Sabatini, J., Lee, Y.W., Ventura, M.: The Automated Text Adaptation Tool. In: HLTNAACL (Demonstrations). (2007) 3-4

[12] Devlin, S., Unthank, G.: Helping aphasic people process online information. In: Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility. (2006)

[13] Chandrasekar, R., Doran, C., Srinivas, B.: Motivations and methods for text simplification. In: Proceedings of the Sixteenth International Conference on Computational Linguistics (COLING '96). (1996) 1041-1044

[14] Siddharthan, A.: An Architecture for a Text Simplification System. In: Proceedings of the Language Engineering Conference (LEC 2002). (2002) 64-71

[15] Junior, A., Maziero, E., Gasperinm, C., Pardo, T., Specia, L., Aluisio, S.: Supporting the Adaptation of Texts for Poor Literacy Readers: a Text Simplification Editor for Brazilian Portuguese. In: Proceedings of the NAACL/HLT Workshop on Innovative Use of NLP for Building Educational Applications, Boulder, Colorado (2009) 34-42

[16] Daelemans, W., Hothker, A., Sang, E.T.K.: Automatic Sentence Simplification for Subtitling in Dutch and English. In: Proceedings of the 4th Conference on Language Resources and Evaluation, Lisbon, Portugal (2004) 1045-1048

[17] Petersen, S.E., Ostendorf, M.: Text Simplification for Language Learners: A Corpus Analysis. In: Proceedings of Workshop on Speech and Language Technology for Education (SLaTE). (2007)

[18] Gasperin, C., Specia, L., Pereira, T.F., Aluisio, S.M.: Learning when to simplify sentences for natural text simplification. In: Proceedings of the Encontro Nacional de Inteligencia Artificial (ENIA), Bento Gonalves, Brazil (2009) 809-818

[19] Zhu, Z., Bernhard, D., Gurevych, I.: A monolingual tree-based translation model for sentence simplification. In: Proceedings of the 23rd International Conference on Computational Linguistics, COLING'10. (2010)

[20] Woddsend, K., Lapata, M.: Learning to simplify sentences with quasi-synchronous grammar and integer programming. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). (2011)

[21] Klerke, S., Sogaard, A.: Simple, readable sub-sentences. In: ACL (Student Research Workshop). (2013)

[22] Devlin, S., Tait, J.: The use of a Psycholinguistic database in the Simplification of Text for Aphasic Readers. In: Linguist Databases. CSLI (1998) 161-173

[23] Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: Introduction to WordNet: An On-line Lexical Database. Int J Lexicography 3(4) (1990) 235-244

[24] Quinlan, P.: The Oxford Psycholinguistic Database. Oxford University Press (1992)

[25] Bautista, S., Gervás, P., Madrid, R.: Feasibility Analysis for SemiAutomatic Conversion of Text to Improve Readability. In: Proceedings of the Second International Conference on Information and Communication Technologies and Accessibility. (2009)

[26] De Belder, J., Deschacht, K., Moens, M.F.: Lexical simplification. In: Proceedings of the 1st International Conference on Interdisciplinary Research on Technology, Education and Communication. (2010)

[27] Peters, E., Hibbard, J., Slovic, P., Dieckmann, N.: Numeracy skill and the communication, comprehension, and use of risk-benefit information. Health Affairs 26(3) (2007) 741-748

[28] Power, R., Williams, S.: Generating numerical approximations. Computational Linguistics 38(1) (2012)

[29] Bautista, S., Hervás, R., Gervás, P., Power, R., Williams, S.: How to Make Numerical Information Accessible: Experimental Identification of Simplification Strategies. In: 13th IFIP TC13 Conference on Human-Computer Interaction (INTERACT), Lisbon, Portugal (2011)

[30] Krifka, M.: Be brief and vague! And how bidirectional optimality theory allows for Verbosity and Precision. In: Sounds and Systems: Studies in Structure and Change: A Festschrift for Theo Vennemann (Trends in Linguistics 141), Mouton de Gruyter, Berlin (2002) 439-458

[31] Williams, S., Power, R.: Precision and mathematical form in first and subsequent mentions of numerical facts and their relation to document structure. In: Proc. of the 12th European Workshop on Natural Language Generation, Athens (2009)

[32] Grice, H.P.: Logic and Conversation. In Cole, P., Morgan, J.L., eds.: Syntax and Semantics: Vol. 3: Speech Acts. Academic Press, San Diego, CA (1975) 41-58

[33] MacKay, D.J.: Sustainable Energy - without the hot air. (2009)

[34] Qualifications, Authority, C.: Annual report and accounts. Technical report, Financial statements (2010)

[35] Anula, A.: Tipos de textos, complejidad lingüística y facilicitación lectora. In: Actas del Sexto Congreso de Hispanistas de Asia. (2007) 45-61

[36] Anula, A.: Lecturas adaptadas a la enseñanza del español como L2: variables lingüísticas para la determinación del nivel de legibilidad. In: La evaluación en el aprendizaje y la enseñanza del español como LE/L2, Pastor y Roca (eds.), Alicante (2008) 162-170

[37] Bautista, S., Drndarevic, B., Hervás, R., Saggion, H., Gervás, P.: Análisis de la Simplificación de Expresiones Numéricas en Español mediante un estudio Empírico. Linguamática 4(2) (2012)

[38] Drndarevic, B., Stajner, S., Bott, S., Bautista, S., Saggion, H.: Automatic Text Simplification in Spanish: A Comparative Evaluation of Complementing Modules. In: 14th International Conference on Intelligent Text Processing and Computational Linguistics (Cicling). (2013)

[39] Padró, L., Stanilovsky, E.: FreeLing 3.0: Towards Wider Multilinguality. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2012), Istanbul, Turkey, ELRA (May 2012)

[40] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A framework and graphical development environment for robust NLP tools and applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics. (2002)

[41] Bautista, S., Saggion, H.: Making Numerical Information more Accessible: Implementation of a Numerical Expressions Simplification Component for Spanish. ITL-International Journal of Applied Linguistics. Special Issue on Readability and Text Simplification. Peeters Publishers, Belgium (2014)

[42] Bautista, S., Hervás, R., Gervás, P., Power, R., Williams, S.: A System for the Simplification of Numerical Expressions at Different Levels of Understandability. In: NLP4ITA. (2013)

[43] Siddharthan, A.: An architecture for a text simplification system. In: Language Engineering Conference, IEEE Computer Society (2002) 64

[44] De Belder, J., Deschacht, K., Moens, M.F.: Lexical simplification. In: Proceedings of Itec2010 : 1st International Conference on Interdisciplinary Research on Technology, Education and Communication. (2010)

[45] Brouwers, L., Bernhard, D., Ligozat, A., Francois, T.: Syntactic Sentence Simplification for French. In: Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) at EACL 2014, Gothenburg, Sweden (2014)

[46] Siddharthan, A., Angrosh, M.: Hybrid text simplification using synchronous dependency grammars with handwritten and automatically harvested rules. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014), Gothenburg, Sweden (2014)

ANNEX

This Annex contains the 15 pairs of sentences selected to the survey carried out. In each question the sentence tagged as (A) is the original and the sentence tagged as (B) is the simplified version.

Question 1

(A) El secretario general de la ONU, Ban Ki-moon, afirma que alrededor de 1.750 millones de personas de 104 países pobres no pueden satisfacer algunas de sus necesidades básicas.

(B) El secretario general de la ONU, Ban Ki-moon, afirma que casi 2000 millones de personas de más de 100 países pobres no pueden satisfacer algunas de sus necesidades básicas.

Question 2

(A) El Alto Comisionado de Naciones Unidas para los refugiados , Antonio Guterres , hizo un llamamiento a la comunidad internacional de unos 280 millones de dólares (205 millones de euros) para las 40 organizaciones que apoyan a los refugiados iraquíes en 12 países.

(B) El Alto Comisionado de Naciones Unidas para los refugiados, Antonio Guterres, hizo un llamamiento a la comunidad internacional de casi 300 millones de dólares (más de 200 millones de euros) para las unas 40 organizaciones que apoyan a los refugiados iraquíes en más de 10 países.

Question 3

(A) Con cerca de 76.000 visitantes profesionales procedentes de 115 países y más de 2.600 exhibidores , se trata de el certamen de referencia para este sector , en el que se presentan las novedades del juguete tradicional para el año 2011 , según informó la Asociación Española de Fabricantes de Juguetes.

(B) Con casi 80000 visitantes profesionales procedentes de más de 100 países y casi 3000 exhibidores, se trata del certamen de referencia para este sector, en el que se presentan las novedades del juguete tradicional para el año 2011, según informó la Asociación Española de Fabricantes de Juguetes.

Question 4

(A) Por comunidades, destaca la representación de empresas de la Comunidad Valenciana con 55 participantes seguida de Cataluña con 29.

(B) Por comunidades, destaca la representación de empresas de la Comunidad Valenciana con casi 60 participantes seguida de Cataluña con casi 30.

Question 5

(A) Según la Comunidad, este proyecto se dirige a los autóctonos y a los más de un millón cien mil inmigrantes que viven en la región, de los que cerca de 34.000 proceden de Bulgaria.

(B) Según la Comunidad, este proyecto se dirige a los autóctonos y a los más de 1000000 inmigrantes que viven en la región, de los que más de 30000 proceden de Bulgaria.

Question 6

(A) Aproximadamente siete de cada diez prefiere el formato papel y el 13 % se decanta por un diario "on-line".

(B) Aproximadamente siete de cada diez prefiere el formato papel y más de 10 % se decanta por un diario "on-line".

Question 7

(A) Los filmes extranjeros no se libraron de la merma y atrajeron a un millón y medio menos de aficionados: 43,7 millones en lugar de los 45,3 millones del periodo anterior.

(B) Los filmes extranjeros no se libraron de la merma y atrajeron a 1500000 menos de aficionados: casi 44 millones en lugar de los más de 45 millones del periodo anterior.

Question 8

(A) Pinturas, esculturas y cerámicas de diferentes periodos y estilos del artista conforman este conjunto de 43 piezas cedidas en comodato por 15 años por la citada fundación.

(B) Pinturas, esculturas y cerámicas de diferentes periodos y estilos del artista conforman este conjunto de más de 40 piezas cedidas en comodato por casi 20 años por la citada fundación.

Question 9

(A) Según dicho trabajo, el hallazgo tuvo lugar en la nebulosa del Cangrejo, situada a 6.300 años luz de la Vía Láctea, en la constelación de Tauro y en la Vía Láctea.

(B) Según dicho trabajo, el hallazgo tuvo lugar en la nebulosa del Cangrejo, situada a más de 6000 años luz de la Vía Láctea, en la constelación de Tauro y en la Vía Láctea.

Question 10

(A) El fomento del conocimiento de la cultura y las lenguas españolas es una de las prioridades del departamento de Exteriores que aporta al Instituto Cervantes 86 de los 102 millones de euros de su presupuesto anual.

(B) El fomento del conocimiento de la cultura y las lenguas españolas es una de las prioridades del departamento de Exteriores que aporta al Instituto Cervantes casi 90 de los más de 100 millones de euros de su presupuesto anual.

Question 11

(A) El 61 % de los españoles de entre 12 y 18 años consumen habitualmente bebidas alcohólicas y de ellos el 10 % lo hacen para "colocarse".

(B) Más de 60 % de los españoles de entre más de 10 y casi 20 años consumen habitualmente bebidas alcohólicas y de ellos 10 % lo hacen para "colocarse".

Question 12

(A) Por otro lado, la ONU ha logrado recaudar un 34 % de los 2.000 millones de dólares (cerca de 1.400 millones de euros) solicitados como llamamiento de urgencia ante la catástrofe de Pakistán.

(B) Por otro lado, la ONU ha logrado recaudar más de 30 % de los 2000 millones de dólares (más de 1000 millones de euros) solicitados como llamamiento de urgencia ante la catástrofe de Pakistán.

Question 13

(A) Alrededor de 390.000 personas han regresado a sus casas desde que vieran obligadas a desplazar se por las inundaciones...

(B) Casi 400000 personas han regresado a sus casas desde que vieran obligadas a desplazar se por las inundaciones...

Question 14

(A) El 18,55 % de las agresiones que sufrieron los médicos españoles en sus consultas el año pasado tuvieron como consecuencia una lesión...

(B) Casi 19 % de las agresiones que sufrieron los médicos españoles en sus consultas el año pasado tuvieron como consecuencia una lesión...

Question 15

(A) En virtud de estas cifras, difundidas este martes en rueda de prensa, en 2010 se registraron en España un total de 451 agresiones a facultativos, es decir, 2,07 por cada mil médicos, lo que supone, a juicio de la organización médica, un “grave problema social “para el que se pide “tolerancia cero “y que se produce en el 90,63 % de los casos en el sector público.

(B) En virtud de estas cifras, difundidas este martes en rueda de prensa, en 2010 se registraron en España un total de casi 500 agresiones a facultativos, es decir, más de 2 por cada 1000 médicos, lo que supone, a juicio de la organización médica, un "grave problema social" para el que se pide "tolerancia cero" y que se produce en casi 91 % de los casos en el sector público.