# Text Simplification Focused on Numerical Expressions

Susana Bautista, Raquel Hervás, and Pablo Gervás

Facultad de Informática, Universidad Complutense de Madrid, Spain
subautis@fdi.ucm.es, raquelhb@fdi.ucm.es, pgervas@sip.ucm.es

**Abstract.** Public information services and documents should be accessible to the widest possible readership. Information in newspapers often takes the form of numerical expressions which pose comprehension problems for many people, including people with disabilities also affected by other barriers such as poverty, illiteracy or lack of access to advanced technology. A first approach to solve this social problem is making numerical information accessible by rewriting difficult numerical expressions in a simpler way.

## 1 Motivation

Text Simplification is a subfield of Natural Language Processing which aims to improve information access for reading-impaired people. The concept of universal accessibility may be extended in order to include new research lines on different kinds of accessibility (accessible technology, architecture, information). The way in which information is written or presented can exclude many people, especially those who have problems to read, write or understand. The process of simplifying texts by hand is extremely time and effort consuming, so any attempts to automate part of this process can leverage the access to information.

Information in newspapers often takes the form of numerical expressions (e.g., economic statistics, demographic data) which pose comprehension problems for many people, including people with disabilities also affected by other barriers such as poverty, illiteracy, lack of access to advanced technology or people with limited education. A UK Government Survey in 2003 estimated that 6.8 million adults had insufficient numeracy skills to perform simple everyday tasks such as paying household bills and understanding wage slips, and that 23.8 million adults would be unable to achieve grade C in the GCSE maths examination for 16-year-old school children [1].

One approach to solve this important social problem of making numerical information accessible is to rewrite difficult numerical expressions more simply. Such an approach would require a set of rewriting strategies yielding expressions that are linguistically correct, easier to understand than the original, and as close as possible to the original meaning. For example, '25.9%' could be rewritten as 'just over a quarter'.

## 2 Related Work

A United Nations report[1] recommends that public information services and documents should be accessible to the widest possible readership. There are different initiatives that propose guidelines that may help when rewriting a text to make it more comprehensive. Some of them are Plain Language[2], the European Guidelines for the Production of Easy-to-Read Information[3] and the latest Web Content Accessibility Guidelines (WCAG 2.0) with a wide range of recommendations for making web content more accessible[4]. There have been various attempts to provide adequate content to these groups of readers through simplifications of already existing materials or writings for a specific target group. Such is the case, for example, with Simple English Wikipedia[5] and Encyclopedia Britannica Elementary[6] in English or Noticias Fácil[7] in Spanish. However, manual simplification is overly slow and costly to be seen as an effective way of producing sufficient amount of the desired reading material. That is why we have seen numerous attempts at developing automatic or semi-automatic text simplification systems in recent years, mainly applied to English [2], but also to Portuguese [3] and Spanish [4]. Previous work has approached the domain of text simplification with the aim of providing simplified texts for humans with reading difficulties, such as foreign language learners [5] or aphasic people [2]. As for numerical expressions, some work has been done, though mainly targeted at experts who generate easy-to-read numerical information and not individuals experiencing numeracy difficulties [6]. Power and Williams [7] were among the first to concentrate on the simplification of such expressions, focusing mainly on the use of hedges as one useful simplification strategy.

## 3 Simplification Strategies for Numerical Information

Our work focuses on lexical expressions containing numerical information. We consider a "numerical expression" as a phrase that expresses a quantity, optionally modified by a numerical modifier called *hedge* as in *more than a quarter* or *around 97%*, where *more than* and *around* are examples of numerical hedges. A survey was carried out with experts in numeracy who were asked to simplify a range of numerical expressions showed like percentages [8]. Our aim was to collect candidate rewriting strategies to obtain guidelines for performing this task automatically. We analyzed two working hypotheses: (1) when experienced writers choose numerical expressions for readers with low numeracy, they tend to prefer round or common values (1/2, 25%) to precise values, and (2) the

[1] http://www.un.org/disabilities/documents/gadocs/standardrules.pdf
[2] http://www.plainlanguage.gov
[3] http://inclusion-europe.org/
[4] http://www.w3.org/TR/WCAG/
[5] http://simple.wikipedia.org/wiki/Main_Page
[6] http://school.eb.co.uk/failedlogin?target=/elementary
[7] http://www.noticiasfacil.es/ES/Paginas/index.aspx

choice between different simplification strategies (fractions, ratios, percentages) is influenced by the value of the proportion, with values in central (48%) and extreme (97.2%) ranges favouring different strategies. Responses were consistent with our intuitions about how common values are considered simpler and how the value of the original expression influences the chosen simplification. In addition, we analyzed the use of numerical hedges in the process of rewriting difficult numerical expressions in simpler ways. Hedges indicate that the original number has been approximated and, in some cases, the direction of approximation. As could be seen in the results, the use of hedges can be influenced by three kinds of parameters: the kind of simplification (simplifications not using percentages, not using decimals or using any kind of numerical form), the simplification strategy, and the loss of precision allowed when the numerical expression is rounded to simplify it. The details of the use of the hedges in the simplification process can be found in Bautista et al. [9]. With the knowledge acquired from our study we have improved our algorithm to simplify numerical expressions. A system was developed for automated simplification of numerical expressions in texts. Experts in simplification tasks were asked to validate the automatic simplifications. In the first prototype, only numerical expressions defined as percentages are adapted. From an input text, the percentage numerical expressions are detected, a target level of difficulty is chosen and the simplified version of the text is generated by replacing the original with the adapted expression. The details of the system operation can be found in Bautista et al. [10].

We have also worked in simplification of numerical expressions in Spanish. We carried out an empirical study of a parallel corpus of original and manually simplified texts with the aim of targeting simplifications concerning numerical expressions [11]. We developed a first prototype of a system with a rule-based lexical transformation component and a syntactic simplification module. System outputs were evaluated with respect to the degree of simplification (more or less similar to the original meaning), the grammaticality of the output, and the preservation of meaning. Our results indicate that the system produces simpler output compared to the original [12]. A special case of study with people with dyslexia has also been studied. We analyzed fixation and reading time to measure readability as well as comprehension questions to score understandability, using eye-tracking with Spanish texts. It is an important study that addresses the cognitive load of number representation an how the mathematical form of the numerical expressions influences comprehension [13].

## 4 Conclusions and Future Work

We have developed a first approximation to the task of simplifying numerical expressions in a text. Our works have shown that the value of the proportion influences the chosen simplification strategy, and that the final mathematical form and the use of hedges are important in the simplification process. The identification of rules governing the selection of substitution candidates constitutes an

important contribution. The empirical evaluation of the implemented systems results in human experts agreeing that the simplifications are appropriate.

Future work has to continue to solve the problems found. The role of context in establishing what simplifications must be used, different treatments for other kind of numerical expressions and different target users should be considered. The final aim is to develop an automatic simplification system in a broader sense, possibly including more complex operations like syntactic transformations and lexical substitutions to reduce the complexity of the vocabulary employed in the text or equivalent graphical representation of numerical expressions. And an evaluation with users should be done to analyze the output of the system.

In our opinion, it is possible to create accessibility guidelines that consider affordable devices, technology, cultural issues and illiteracy. We have to continue to work for designing for diversity. In the diversity is the growth and greatness, and the universal accessibility would have like main aim the user-centered design.

## References


1. Williams, J., Clemens, S., Oleinikova, K., Tarvin, K.: A national needs and impact survey of literacy, numeracy and ICT skills. Technical report (2003)
2. Carroll, J., Minnen, G., Canning, Y., Devlin, S., Tait, J.: Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In: AAAI-98. (1998)
3. Specia, L.: Translating from Complex to Simplified Sentences. In: 9th International Conference on Computational Processing of the Portuguese Language. (2010)
4. Saggion, H., Gómez-Martínez, E., Anula, A., Bourg, L., Etayo, E.: Text simplification in simplext: Making texts more accessible. In: SEPLN. (2011)
5. Medero, J., Ostendorf, M.: Identifying targets for syntactic simplification. In: Proceedings of Speech and Language Technology in Education. (2011)
6. Peters, E., Hibbard, J., Slovic, P., Dieckmann, N.: Numeracy skill and the communication, comprehension, and use of risk-benefit information. Health Affairs (2007)
7. Power, R., Williams, S.: Generating numerical approximations. Computational Linguistics **38**(1) (2012)
8. Bautista, S., Hervás, R., Gervás, P., Power, R., Williams, S.: How to Make Numerical Information Accessible. In: INTERACT-11. (2011)
9. Bautista, S., Hervás, R., Gervás, P., Power, R., Williams, S.: Experimental identification of the use of hedges in the simplification of numerical expressions. In: Workshop on Speech and Language Processing for Assistive Technologies. (2011)
10. 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)
11. 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)
12. Drndarevic, B., Stajner, S., Bott, S., Bautista, S., Saggion, H.: Automatic Text Simplification in Spanish: A Comparative Evaluation of Complementing Modules. In: COLING. (2013)
13. Rello, L., Bautista, S., Baeza-Yates, R., Gervás, P., Hervás, R., Saggion, H.: One Half or 50%? An Eye-Tracking Study of Number Representation Readability. In: INTERACT. (2013)