Can Numerical Expressions Be Simpler? Implementation and Demonstration of a Numerical Simplification System for Spanish

Susana Bautista, Horacio Saggion
NIL: Natural Interaction based on Language, TALN: Natural Language Processing Group
Universidad Complutense de Madrid, Universidad Pompeu Fabra
Madrid, Barcelona, Spain
E-mail: subautis@fdi.ucm.es, horacio.saggion@upf.edu

Abstract

Information in newspapers is often showed in the form of numerical expressions which present comprehension problems for many people, including people with disabilities, illiteracy or lack of access to advanced technology. The purpose of this paper is to motivate, describe, and demonstrate a rule-based lexical component that simplifies numerical expressions in Spanish texts. We propose an approach that makes news articles more accessible to certain readers by rewriting difficult numerical expressions in a simpler way. We will showcase the numerical simplification system with a live demo based on the execution of our components over different texts, and which will consider both successful and unsuccessful simplification cases.

Keywords: Numerical Expressions, Simplification, Spanish

1. Introduction

Access to information is a fundamental human right which was asserted in the United Nations' Universal Declaration of Human Rights (United Nations). However, the way in which information is written or presented has a great impact on text readability and comprehension for specific groups of people.

A number of guidelines have been proposed to make a text "easy to read and understand"— see for example Plain Language (2005), the European Guidelines for the Production of Easy-to-Read Information (Freyhoff, G., Hess, G., Kerr, L, Menzel, E., Tronbacke, B. and Veken, K.V.D., 1998) or the Web Content Accessibility Guidelines (W3C, 2008). Adapting texts to make them easy to understand for specific target user groups is generally done manually, making massive production of easy-to-read texts practically impossible. Automatic text simplification is a technology which has the potential to speed up the process of transforming a text into a quasi equivalent which could be more understandable.

One important type of information which is particularly abundant in newspaper articles is numerical information, be it dates, measurements, quantities, percentages, or ratios. Numerical information poses comprehension problems for many people, including people with disabilities, older people, or persons with low literacy levels. If we have a look at daily news and pay attention to the numerical information, we can see the number and variety of numerical expressions used to present the information.

In order to address the problem of numerical expression simplification require a set of rewriting strategies to transform a difficult numerical expression into a simpler "equivalent" which is linguistically correct. For example, the expression '52.9%' could be rewritten as 'just over a half' preserving the intended meaning and losing only a bit of precision. Loss of precision is not necessarily detrimental, for several reasons. Loss of precision can be signalled linguistically by numerical hedges such as 'around', 'more than' and 'a little under', so it need not be misleading. It is worth noting that numerical expression simplification is a normal practice in newspaper article editing and an important summarization operation.

The following pair of (1) original and (2) simplified sentences is an example of the simplification transformations of numerical expressions.

1. *El informe, que recoge datos de la ONU, destaca que entre enero y junio de este año hubo **1.271** víctimas, **un 21%** más que en el primer semestre de 2009.*
2. *El informe, que recoge datos de la ONU, destaca que entre enero y junio de este año hubo **más de 1000** víctimas, **más de 20%** más que en el primer semestre de 2009.*

We can see in the example that the numerical expression **1.271** has been transformed in the expression **más de 1000** (more than 1000) and the numerical expression **un 21%** (a 21%) has been transformed in the expression **más de 20%** (more than 20%). In this way, we preserve the meaning despite losing a little precision.

The purpose of this paper is to motivate, describe, and demonstrate a rule-based lexical component that simplifies numerical expressions in Spanish newspapers. To the best of our knowledge, there are not previous attempts to simplify numerical expressions in Spanish

texts to make texts more readable. We propose an approach that makes news articles more accessible to readers by applying a set of grammars that transform difficult numerical expressions into simpler expressions at the risk of losing some precision.

2. Related Work

The first text simplification systems had been directed mainly at reducing lexical and syntactic complexity of texts (Chandrasekar, R., Doran, C. and Srinivas, B. , 1996)(Devlin, S and Tait, J, 1998).

A few rule-based systems have been developed for text simplification focusing on different kinds of readers (e.g., subjects with low literacy levels, aphasic people) (Chandrasekar, R., Doran, C. and Srinivas, B, 1996) (Siddharthan, A., 2003), (Bautista, S., Gervás, P. and Madrid, R.I, 2009) (Aluísio, S. M., Specia, L., Pardo, T.A., Maziero, E. and Fortes, R., 2008). These systems apply to each sentence a set of manually created simplification rules. These are usually based on parser structures and limited to certain simplification operations.

Other kinds of simplification systems are corpus-based. They can learn from corpora the relevant simplification operations (Zhu, Z., Bernhard, D. and Gurevych, I., 2010) (Specia, 2010). In particular, Petersen and Ostendorf (2007) address the task of text simplification in the context of second language learning.

The treatment of numerical information in different areas, such as health care, forecast, representation of probabilistic information or vague information, has been studied by experts in previous works (Peters, E., Hibbard, J., Slovic, P. and Dieckmann, N., 2007), (Dieckmann, N., Slovic, P. and Peters, E., 2009), (Bisantz, A.M., Schinzing, S. and Munch, J., 2005), (Mishra, H., Mirshra, A. and Shiy B., 2011).

Focusing on simplifying numerical information, Bautista et al. (2011) and Power and Williams (2012) are among the first to study the possibility of simplifying this kind of expressions, concentrating mainly on the use of modifiers.

A study of simplification of numerical expressions in Spanish was undertaken using a parallel corpus of original and manually simplified texts with the aim of developing a rule-based system (Bautista, S., Drndarevic, B., Hervás, R., Saggion, H. and Gervás, P., 2012). A prototype of a system based on a rule-based lexical transformation component which included our numerical simplification prototype and a syntactic simplification module was developed and evaluated for simplicity and meaning preservation (Drndarevic B, Stajner S, Bott S, Bautista S and Saggion H., 2013).

3. System Development

Our methodology consists in the next steps. For details the reader is referred to (Bautista S., Saggion H., 2014).

1. Developing and evaluating a component for the identification of numerical expressions in Spanish texts.
2. Analyzing a parallel corpus of original and manually simplified news articles, aimed at extracting types of simplification operations to be automated.
3. Analyzing the answers collected from a survey, which asked subjects to simplify numerical expressions from the corpus and upon which we identified simplification strategies used.
4. Building a rule-based lexical component for automatic simplification of numerical expressions.
5. Evaluating the automatically simplified output.

In the next sections we give an overview of the simplification algorithm. We focus on the identification of the numerical expressions in the Spanish texts and their automatic annotation.

3.1 Identifying Numerical Expressions in Spanish

In order to ground our approach we relied on a parallel corpus of original and simplified documents. From the Project Simplext (Saggion, H., Gómez-Martínez, E., Esteban Etayo, E., Anula, A. and Bourg, L., 2011) we selected a set of 40 original and manually simplified news articles in Spanish. Simplifications had been produced by trained human editors, aware of the needs of a person with cognitive disabilities and following a series of easy-to-read guidelines suggested by Anula (2007). This corpus has been used in all the steps described hereafter.

With the aim of developing a rule-based lexical component to simplify numerical expressions, we create the first specialized component to identify numerical expressions in Spanish texts. We base its development on two widely used tools for Natural Language Processing research: FreeLing (Padró, Ll., Collado, M., Reese, S., Loberes, M. and Castellón, I., 2010), the best known system for linguistic treatment of Spanish, and the General Architecture for Text Engineering (GATE) (Maynard, D., Tablan, V., Cunningham, H., Ursu, C., Saggion, H., Bontcheva, K. and Wilks, Y., 2002), which provides support for corpus analysis and development of simplification rules through its Java Annotation Pattern Engine (JAPE). We use the two components separately: FreeLing is used to analyse a document in order to produce tokens, sentences, and parts-of-speech tags from which we create an XML representation. The latter can be used within the GATE system directly.

3.2 Automatic Annotation of Numerical Expressions

In order to develop the numerical expression recognizer we rely on the Java Annotation Pattern Engine (JAPE), a regular expression recognizer over annotations implemented in GATE. We define a set of JAPE grammars in order to tag the different types of numerical expressions in the original texts, with their possible modifiers. A JAPE grammar consists of a set of phases, each of which consists of a set of pattern and action rules. The phases run sequentially and constitute a cascade of finite state transducers over annotations.

In the next example, we have a rule called "AlrededorPartitivos", which identifies a special kind of numerical expression called "Partitivos" preceded by a modifier. The pattern of the rule indicates that text annotated with a "word" tag whose "lemma" feature has the value "alrededor_de" (i.e. "around") and is followed optionally by another word. In addition, it will match any text annotated with a "word" annotation whose "tag" feature has the value "Zd". Examples of this kind of numerical expressions are: "alrededor de 9000 millones (around 9000 million)" or "alrededor de una docena (around a dozen)".

```
Rule: AlrededorPartitivos
((({word.lemma == "alrededor_de"})
({word})?):modifier ({word.tag ==
"Zd"})):annotate
-->
:modifier.MOD_EXP = {semantics =
"alrededor"},
:annotate.ALREDEDORPART =
{semantics="partitivos"}
```

We use these rules to recognize different kinds of numerical expressions which we need to manually analyse in order to understand how they can be simplified. For analysing the corpus we have relied on the ANNIC system (Aswani, N., Tablan,V., Bontcheva, K. and Cunningham, H., 2005), which allows us to see annotations in context. This system lets us make searches in the tagged corpus with labels generated from the defined rules in our grammars and improve rule coverage though an iterative cycle. The final grammars, which contain 45 different rules, cover 13 different cases of numerical expressions corresponding to the 4 numerical tags identified by FreeLing.

In Figure 1 we can see an example of original text from the corpus with numerical expressions identified.

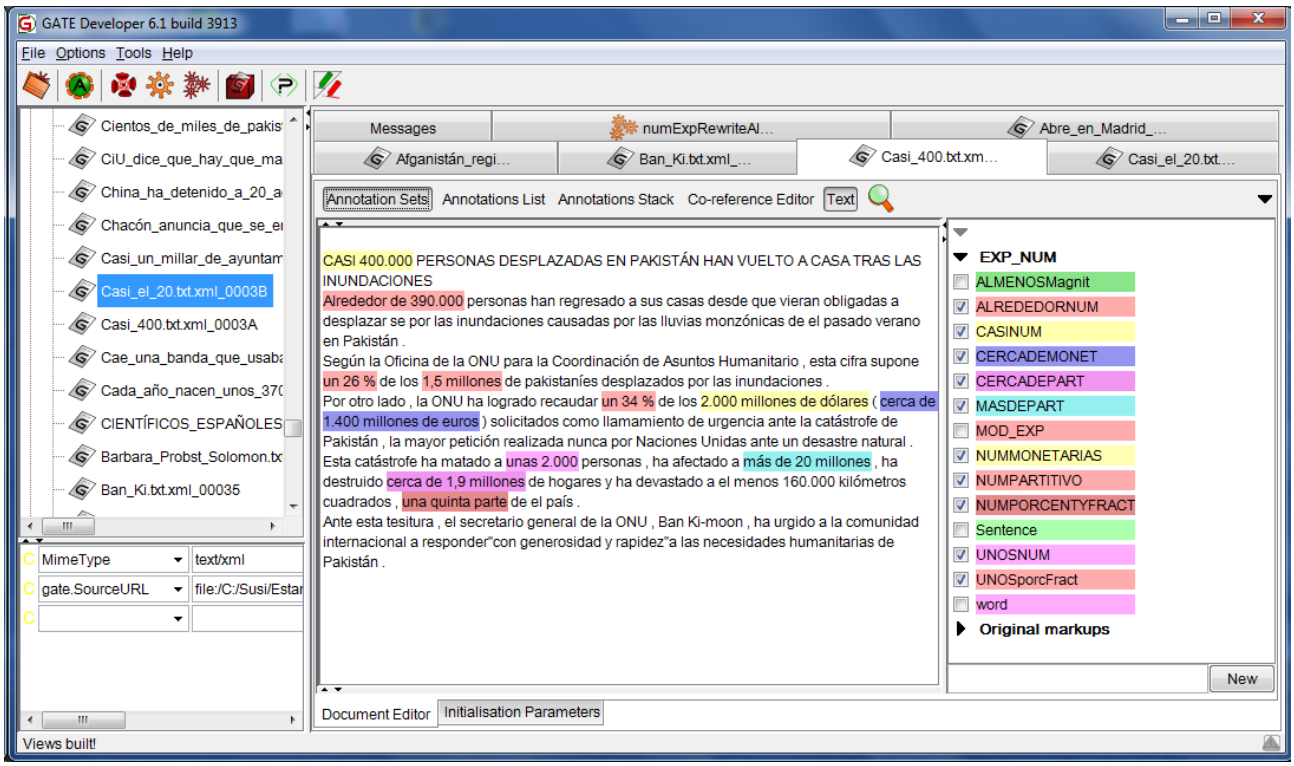

Figure 1: Original text with numerical expressions identified.

3.3 A Rule-base Lexical Component

Our simplification system is composed of the following components in the sequence described following:

1. Text processing using FreeLing
2. Transformation of FreeLing output into XML representation
3. Application of grammars for numerical expression recognition
4. Simplification of target numerical expression
5. Sentence rewriting

These components have all been integrated in a plug-in developed in Java which can be used within the GATE system. FreeLing is used to carry out the basic analysis of the text: word and sentence recognition and parts-of-speech tagging.

The third component is the set of grammars described before, which were integrated into a named entity recognizer in GATE. This module produces annotations of type NumExp in the document which contains a feature that indicates the type of pre-modification of the numerical expression. The fourth component implements the simplification strategy which is the most commonly observed in the undertaken survey (see (Bautista S., Saggion H., 2014)): the numerical quantity is always rounded and a set of rules is applied to the modifier chosen to account for loss of precision. If measurement units are present in the original expression, they are also processed. The simplified version is made up of a selected modifier, the rounded numerical expression and optionally the units if these were present in the original text.

Finally, the last module rewrites the text replacing the original numerical expression with the components added in the previous module. So, for each numerical expression, the feature with the simplified version is processed to replace it. After replacement, a post-processing of the text is carried out to solve any errors which arose from the treatment of the text by FreeLing, e.g. the transformation of contractions, such as “del” (“of the”) into their components “de + el” (“of the”).

In the Appendix we can see an example of the original and simplified text produced by our system with the numerical expressions marked in bold.

4. Evaluation

We have applied the developed rules to a subset of 10 unseen documents from the Simplext corpus comprising 59 sentences. We have corrected the automatic annotations produced by the system in order to obtain a gold standard dataset for evaluation using the GATE tool. We have tested the performance of the rules obtaining a precision of 0.94, a recall of 0.93 and an F-measure of 0.93, which we consider quite acceptable. For numerical expressions with low frequency of occurrence, the results are worse but they are better for frequently observed numerical expressions.

We have analyzed the linguistic accuracy of the output, and the results were positively rated, with 83.56% (almost 84%) of the simplified sentences considered correct, where all containing numerical expressions, and the meaning was preserved reasonably well in the process of simplification.

The results were analyzed qualitatively revealing that most common errors were due to bad treatment of comparative numerical expressions.

5. Discussion

We have attempted to identify a variety of numerical expressions in order to have a good coverage. We acknowledge that concentrating on fewer types in order to boost precision could be a useful strategy when users are involved.

In the example given in the Appendix we can see there is no global simplification strategy in our system. Different rules are applied depending on the type of numerical expressions. For example, one of the first rules applied deletes decimals in the number and adds a modifier. Other rule prefers to use frequent percentages because they are easier to understand.

One of the main problems we encountered is the treatment of comparative numerical expressions within a single sentence. For example, the original sentence “The numbers of dissolutions are maintained at 2010 similar to those of 2009, 22,435 versus 21,875, with a slight increase of 2.56%” (“Las cifras de disoluciones se mantienen en 2010 similares a las de 2009, 22.435 frente a 21.875, con un ligero incremento del 2,56%.”) is simplified by our system into “The numbers of dissolutions are maintained at 2010 similar to those of 2009, more than 20000 versus more than 20000, with a slight increase of almost 3%” (“Las cifras de disoluciones se mantienen en 2010 similares a las de 2009, más de 20000 frente a más de 20000, con un ligero incremento del casi 3%.”). This is due to the fact that our system applies the same simplification strategy to both numerical expressions without taking into account the context in which they occur, thus losing the sense of the original sentence. Such cases, where our current simplification approach fails, could be treated by further adjusting our JAPE grammars.

Experimental psychology and cognitive neuropsychology have studied number processing and calculation over the last two decades. Many researchers have studied the cognitive processes that are responsible for number

processing and calculation, with the aim of contributing to the improvement of teaching and learning processes. For example, Herrera and Macizo (2012), and Salguero and Alameda (2003), present findings that show that the frequency of use of a word or a number is an influential variable in the reading process. In addition, it seems that numerical expressions most frequently used require less recognition time. It is important to address in our future work the simplification of numerical expressions replacing with frequently used constructs.

As part of our future work, we intend to take syntactic context into consideration when simplifying numerical expressions because it might influence in the kind of simplifications applied. In addition, we have observed adding a modifier the meaning is not always kept, so we have to improve the system considering in more detail the loss of precision to decide whether a modifier is needed. Another line of work is to evaluate the impact of the results by carrying out a comparative analysis of the comprehension of original versus simplified numerical expressions.

6. Acknowledgements

We thank Stefan Bott and Biljana Drndarevic for their help in our work. The research described in this paper is partially funded by the Ministerio español de Educación y Ciencia (TIN2009-14659-C03-01 Project), the FPI grant program for the first author, while the research of the second author is funded by fellowship RYC-2009-04291 from Programa Ramón y Cajal 2009 and project number TIN2012-38584-C06-03 (SKATER-UPF-TALN) from Ministerio de Economía y Competitividad, Secretaría de Estado de Investigación, Desarrollo e Innovación, Spain. We also acknowledge support from the project ABLE-TO-INCLUDE (CIP-ICT-PSP-2013-7/621055).

7. References

Department for Education. (2003). *Skills for life.* United Kingdom.

Aluísio, S. M., Specia, L., Pardo, T.A., Maziero, E. and Fortes, R. (2008). Towards Brazilian Portuguese automatic text simplification systems. *ACM Symposium on Document Engineering 2008: 240-248.*

Anula, A. (2007). Tipos de Textos, Complejidad Lingüística y Facilitación Lectora. *Actas del Sexto Congreso de Hispanistas de Asia*, (pp. 45-61).

Aswani, N., Tablan,V., Bontcheva, K. and Cunningham, H. (2005). Indexing and Querying Linguistic Metadata and Document Content. *Proceedings of 5th International Conference on Recent Advances in Natural Language Processing.* Borovets, Bulgaria.

Bautista, S., Drndarevic, B., Hervás, R., Saggion, H. and Gervás, P. (2012). Análisis de la Simplificación de Expresiones Numéricas en Español mediante un estudio empírico. *Linguamática* .

Bautista, S., Gervás, P. and Madrid, R.I. (2009). Feasibility analysis for semiautomatic conversion of text to improve readability. *Proceedings of the Second International Conference on Information and Communication Technology and Accessibility.* Hammamet, Tunisia.

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

Bautista, S., Saggion, H. (2014). Making Numerical Information more Accessible: Implementation of a Numerical Expressions Simplification Component for Spanish. *Special Issue of the International Journal of Applied Linguistics (ITL) on readability and text simplification.*

Bisantz, A.M., Schinzing, S. and Munch, J. (2005). Displaying uncertainty: Investigating the effects of display format and specificity. *The Journal of the Human Factors and Ergonomics.*

Chandrasekar, R., Doran, C. and Srinivas, B. (1996). Motivations and methods for text simplifications. *Proceedings of the 16th International Conference on Computational Linguistics*, (pp. 1041-1044). Copenhagen, Denmark.

Chandrasekar, R., Doran, C. and Srinivas, B. . (1996). Motivations and methods for text simplifications. *Proceedings of the 16th International Conference on Computational Linguistics*, (págs. 1041-1044). Copenhagen, Denmark.

Clark, D. (2010). *Young people reading and writing today: Whether, what and why.* National Literacy Trust.

Deeqa Jama, G. (2010). *Literacy: State of the nation.* National Literacy Trust.

Devlin, S and Tait, J. (1998). The use of a Psycholinguistic database in the Simplification of Text for Aphasic Readers. *Linguistic Databases* , 161-173.

Dieckmann, N., Slovic, P. and Peters, E. (2009). The use of narrative evidence and explicit likelihood by decision markers varying in numeracy. *Risk Analysis* .

Drndarevic B, Stajner S, Bott S, Bautista S and Saggion H. (2013). Automatic Text Simplification in Spanish: A Compartive Evaluation of Complementing Modules. *International Conference on Intelligent Text Processing and Computational Linguistics.* Samos, Greece.

Drndarevic, B. and Saggion, H. (2012). Reducing Text Complexity through Automatic Lexical Simplification: an Empirical Study for Spanish. *Sociedad Española para el Procesamiento del Lenguaje Natural* , 13-20.

Freyhoff, G., Hess, G., Kerr, L, Menzel, E., Tronbacke, B. and Veken, K.V.D. (1998). *European guidelines for the*

production of easy-to-read information. Retrieved 06 22, 2013, from http://www.osmhi.org/contentpics/139/European Guidelines for ETR publications.pdf

Grice, H. (1975). Logic and Conversation. *Syntax and Semantics* , 41-58.

Herrera, A. and Macizo, P. (2012). Cómo leemos los números? (How we read numbers?). Ciencia Cognitiva , 44-47.

Krifka, M. (2002). Be brief and vague! And how bidirectional optimality theory allows for Verbosity and Precision. *Sounds and Systems: Studies in Structure and Change: A Festschrift for Theo Vennemann*, (pp. 439-458). Berlin.

Maynard, D., Tablan, V., Cunningham, H., Ursu, C., Saggion, H., Bontcheva, K. and Wilks, Y. (2002). Architectural Elements of Language Engineering Robustness. *Journal of Natural Language Engineering-Special Issue on Robust Methods in Analysis of Natural Language Data* , 257-274.

Mishra, H., Mirshra, A. and Shiy B. (2011). In praise of vagueness: malleability of vague information as a performance booster. *Psychological Science* .

Padró, Ll., Collado, M., Reese, S., Loberes, M. and Castellón, I. (2010). Freeling 2.1: Five years of open-source language processing tools. *Proceedings of the 7th International Conference on Language Resources and Evaluation.* Valletta, Malta.

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

Petersen, S.E. and Ostendorf, M. (2007). Text simplification for language learners: a corpus analysis. . *Proceedings of Workshop on Speech and Language Technology for Education.*

Power, R. and Williams, S. (2012). Generating numerical approximations. *Computational Linguistics* .

Saggion, H., Gómez-Martínez, E., Esteban Etayo, E., Anula, A. and Bourg, L. (2011). Text Simplification in Simplext. Making Text More Accessible. *Procesamiento del Lenguaje Natural , 47*, 341-342.

Salguero, M. and Alameda, J. (2003). El procesamiento de los números y sus implicaciones educativas (Number processing and its educational implications). *XXI Revista de Educación (Educational Journal)* , 181-189.

Siddharthan, A. (2003). Syntactic Simplification and Text Cohesion. *Ph.D dissertation, research and language and computation.*

Specia, L. (2010). Translating from Complex to Simplified Sentences. *Proceedings of Computational Processing of the Portuguese Language.* Porto Alegre, RS, Brazil.

The Plain Language Action and Information Network (PLAIN). (2005). Retrieved 10 14, 2013, from Plain Language: http://www.plainlanguage.gov

United Nations. (n.d.). *Universal Declaration of Human Rights*. Retrieved 10 14, 2013, from Universal Declaration of Human Rights: http://www.un.org/es/documents/udhr/

W3C. (2008). *Web Content Accessibility Guidelines*. Retrieved 10 14, 2013, from http://www.w3.org/TR/WCAG20/

Williams, S. and Power, R. (2009). Precision and mathematical form in first and subsequent mentions of numerical facts and their relation to document structure. *12th European Workshop on Natural Language Generations.* Athens, Greece.

Zhu, Z., Bernhard, D. and Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. *Proceedings of the 23rd International Conference on Computational Linguistics.* Beijing, China

.

Appendix

The following pair of original and simplified texts in Spanish is an example of the system output in the simplification of numerical expressions.

Original

CASI EL 20% DE LAS AGRESIONES QUE SUFREN LOS MÉDICOS CAUSAN LESIONES

El 18,55% *de las agresiones que sufrieron los médicos españoles en sus consultas el año pasado tuvieron como consecuencia una lesión, según los datos del Observatorio de Agresiones de la Organización Médica Colegial, que indican también que* ***el 13,4%*** *de los facultativos afectados por esta situación pidieron por esta causa la baja laboral. 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. El ámbito médico más afectado por las agresiones de pacientes, es, en virtud del observatorio creado por los colegios de facultativos, el de Atención Primaria, donde se contabilizaron en 2010* ***el 65%*** *de los atentados a profesionales sanitarios. Y el grupo de edad más castigado, el que va desde los* ***46*** *a los* ***55*** *años.*

Simplified

CASI EL 20% DE LAS AGRESIONES QUE SUFREN LOS MÉDICOS CAUSAN LESIONES

*C****asi 19%*** *de las agresiones que sufrieron los médicos españoles en sus consultas el año pasado tuvieron como consecuencia una lesión, según los datos del Observatorio de Agresiones de la Organización Médica Colegial, que indican también que* ***más de 13%*** *de los facultativos afectados por esta situación pidieron por esta causa la baja laboral. 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. El ámbito médico más afectado por las agresiones de pacientes, es, en virtud del observatorio creado por los colegios de facultativos, el de Atención Primaria, donde se contabilizaron en 2010* ***más de 60%*** *de los atentados a profesionales sanitarios. Y el grupo de edad más castigado, el que va desde los* ***casi 50*** *a los* ***casi 60*** *años.*