TEXT SIMPLIFICATION USING DEPENDENCY PARSING FOR SPANISH

Miguel Ballesteros
Departamento de Ingeniería del Software e Inteligencia Artificial, Universidad Complutense de Madrid
C/ Profesor José García Santesmases, s/n, E-28040 Madrid, Spain
miballes@fdi.ucm.es

Susana Bautista, Pablo Gervás
Instituto de Ingeniería del Conocimiento, Universidad Complutense de Madrid
C/ Profesor José García Santesmases, s/n, E-28040 Madrid, Spain
subautis@fdi.ucm.es, pgervas@sip.ucm.es

Keywords: Text simplification, Dependency parsing, Spanish.

Abstract: In this paper we investigate the task of text simplification for Spanish. Our purpose is a system to simplified text based on rules using dependency parsing. Our main motivation is the need for text simplification to facilitate accessibility to information by poor readers and by people with cognitive disabilities. This study consists of the first step towards building Spanish text simplification systems helping to create easy-to-read texts.

1 INTRODUCTION

Text simplification aims at providing human readers with a better understanding of a written text through its simplification. Our goal is to build a system to promote access to Spanish texts for people at the rudimentary and basic literacy levels, as well as for those with cognitive disabilities.

In Spain a vast number of people belong to the so called rudimentary and basic literacy levels. These people are only able to find explicit infomation in short texts or process slightly longer texts and make simple inferences. According to some studies [1] to measure the literacy level of the population, 30% of the population have difficulty understanding texts beyond a certain complexity.

Reading comprehension entails three elements: the reader who is meant to comprehend; the text that is to be comprehended and the activity in which comprehension is a part of (Snow et al., 2002). In addition to the content presented in the text, the vocabulary load of the text and its linguistic structure, discourse style, and genre interact with the reader's knowledge. When these factors do not match the reader's knowledge and experience, the text becomes too complex for comprehension to occur. In this paper we will focus on the syntactic structure of a text to maximize the comprehension of written texts through the simplification of their linguistic structure. This may involve simplifying lexical and syntactic phenomena, by substituting words that are more usual, and by breaking down and changing the syntactic structure of the sentence. As a result, it is expected that the text can be more easily understood (Siddharthan, 2003)(Max, 2006). Text simplification may also involve dropping parts or full sentences and adding some extra material to explain a difficult point (Petersen and Ostendorf, 2007).

It has already been shown that long sentences, conjoined sentences, embedded clauses, passives, non-canonical word order, and use of low-frecuency words, among other things, increase text complexity for language-impaired readers (Siddharthan, 2002),(Klebanov et al., 2004),(Devlin and Unthank, 2006),(Bautista et al., 2009),(Caseli et al., 2009). There are different initiatives that make available guidelines to make text easier to comprehend: the Plain Language [2] or "European Guidelines for the Production of Easy-to-Read Information"[3] or "Web

[1] http://www.facillectura.es

[2] http://www.plainlanguage.gov

[3] http://www.osmhi.org/contentpics/139/European

Content Accessibility Guidelines"[4]. In principle, these recommendations can be applied to any language.

In this paper we present the results of the early steps in the study of syntactic simplification for Spanish and a rule-based syntactic simplification system for this language. We follow a subset of the whole set of guidelines to define our rules: use short sentences mostly, include one main idea per sentence and do not try to express more of an idea or theme in each sentence.

This paper is organized as follows. In Section 2 we describe related approaches for text simplification. Section 3 presents our proposal and the evaluation measures. In Section 4 we show our results. Section 5 presents the conclusions and some future work.

2 PREVIOUS WORK

In this section we present related approaches for text simplification, state of the art about multilingual dependency parsing, and the corpus that we used for our experiment.

2.1 Text Simplification

Existing text simplification systems can be compared along three axes: the type of system- rule-based or corpus-based-, the type of knowledge used to identify the need for simplification, and the goals of the system.

A few rule-based systems have been developed for text simplification (Chandrasekar et al., 1996),(Siddharthan, 2003),(Bautista et al., 2009), focusing on different readers (poor literate, aphasic, etc). These systems contain a set of manually created simplification rules that are applied to each sentence. These are usually based on parser structures and limited to certain simplification operations. Siddharthan proposes a syntactic simplification architecture that relies on shallow text analysis and favors time performance. The general goal of the architecture is to make texts more accessible to a broader audience. Max (Max, 2006) applies text simplification in the writing process by embedding an interactive text simplification system into a word processor. At the user's request, an automatic parser analyzes an individual sentence and the system applies handcrafted rewriting rules. This system requires human intervention at every step.

Corpus-based systems, on the other hand, can learn from corpus the relevant simplification operations and also the necessary degree of the simplification for a given task (Petersen and Ostendorf, 2007). Petersen addresses the task of text simplification in the context of second-language learning. A data-driven approach to simplification is proposed using a corpus of paired articles in which each original sentence does not necessarily have a corresponding simplified sentence, making it possible to learn where writers have dropped or simplified sentences. A classifier is used to select the sentences to simplify, and Siddharthan's syntactic simplification system is used to split the selected sentences. Inui et al. (Inui et al., 2003) proposes a rule-based system for text simplification aimed at deaf people.

Some language technology systems attempt to simplify documents for various purposes. A variety of simplification techniques have been used, for example substituting common words for uncommon words (Devlin and Tait, 1998), activising passive sentences and resolving references (Canning, 2000), reducing multiple-clause sentences to single-clause sentences (Chandrasekar and Srinivas, 1997; Canning, 2000; Siddharthan, 2002) and making appropriate choices at the discourse level (Williams et al., 2003).

There also commercial systems like Simplus [5] and StyleWriter [6], which aim to support Plain English writing.

2.2 Dependency Parsing

A dependency is a binary syntactic asymmetrical relation between the words of a sentence that is relevant to the structure of the sentence (Kübler et al., 2009). Based on this main idea, we can define what would be the dependency parsing. The words in a sentence depend on each other, so that the direct object of a verb depends directly on the verb and an adjective depends on a name. Finally, the purpose of dependency analysis is to build a tree where leaves represent each of the words comprising the phrase and the edges represent the dependencies between them, this tree is called the dependency tree.

There is a lot of work done in dependency parsers, and some shared tasks had as main theme Multilingual dependency parsing like the CoNLL-X Shared Task(Buchholz and Marsi, 2006). Each year the Conference of Computational Natural Language Learning(CoNLL) features a shared task, the 10th CoNLL Shared task was Multilingual dependency parsing.

Guidelines for ETR publications (2).pdf

[4] http://www.w3.org/TR/WCAG20/

[5] http://www.linguatechnologies.com/english/home.html

[6] http://www.editorsoftware.com/writing-software

```
1   Le      él      p   pp  num=s|per=3|gen=c|case=d    2   ROOT
2   dijo    decir   v   vm  num=s|per=3|mod=i|tmp=s 0   CI
3   a       a       s   sp  for=s   2   CI  2   CI
4   Auri    Auri    n   np  _       3   _   3   _
5   que     que     c   cs  _       7   _   7   _
6   se      él      p   p0  per=3   7   _   7   _
7   iba     ir      v   vm  num=s|per=3|mod=i|tmp=i 2   CD
8   al      al      s   sp  gen=m|num=s|for=c   7   CC
9   fútbol  fútbol  n   nc  gen=m|num=s 8   _   8
10  ,       ,       F   Fc  _       2   PUNC    2   PUNC
11  ella    él      p   pp  num=s|gen=f|per=3   13  SUJ
12  le      le      p   pp  num=s|per=3|gen=c|case=a    13
13  miró    mirar   v   vm  num=s|per=3|mod=i|tmp=s 2   _
14  con     con     s   sp  for=s   13  CC  13  CC
15  Los     Los     d   da  gen=m|num=p 16  _   16
16  ojos    ojo     n   nc  gen=m|num=p 14  _   14
17  fuera   fuera   r   rg  _       16  _   16  _
18  de      de      s   sp  for=s   17  _   17  _
19  las     el      d   da  gen=f|num=p 20  _   20
20  órbitas órbita  n   nc  gen=f|num=p 18  _   18
```

Figure 1: A tagged sentence from AnCora.

There were a lot of research groups, each group implemented a parser and there were a lot of languages to parse. The Corpus that they used for Spanish parsing is AnCora (Palomar et al., 2004), (Taulé et al., 2008) and we used it too for our experiment. The aim of this task was to extend the state of the art available at that time in dependency parsing. In 2007, another shared task about multilingual dependency parsing was accomplished: The CoNLL-XI Shared Task(Mcdonald et al., 2007), but in this case Spanish was not present as a language for parsing.

2.3 AnCora Corpus

We used the AnCora (Palomar et al., 2004), (Taulé et al., 2008) treebank, a corpus of 95,028 wordforms and 3,512 sentences that contains open-domain texts annotated with their dependency analyses. The ConLL X Shared Task (Buchholz and Marsi, 2006) used AnCora as treebank for the Spanish parsing and better scores were around 80% LAS(Labelled Attachment Score). AnCora was tagged automatically with morphosyntactic information (PoS tags) and manually checked. It has been used as a training corpus for learning based systems.

AnCora is in CoNLL Data Format[7], as shown in the Figure 1. A sentence given in that format has all the information about the dependency tree and some other lexical information. AnCora has a dependency tag set of 20 different tags, but the frequency of most of the labels is very low. We saw that the 'CD'(Direct Object) tag, the 'CI'(Indirect Object) tag, and the 'CC'(Adjunct) tag appear in all the sentences with more than 15 wordforms.

[7]http://nextens.uvt.nl/ conll/

3 DEPENDENCY BASED TEXT SIMPLIFICATION

We propose a rule-based syntactic simplification system. It uses a dependency parsed tree and it is limited to a simplification operation applied to the dependency trees, prunning the tree focusing on the dependency labels. The operation is applied sentence by sentence of the corpus, producing simplified versions of the sentences.

3.1 Dependency Tree Pruning

We were wondering which tag is the most appropriate to be removed, and we focused on the small subset of 3 tags ("CC", "CD", and "CI") that appears in most of the sentences. The only tag that could be deleted without losing the main information of the sentence is the "CC" tag. It expresses complementary information about an action, like *when, where, how*, and *why*. But "CC" tag never reports about *who* or *what*. Removing "CC" tag, we are not always loosing the information about i.e. *when* or *where* because this kind of information not always depends on a verb.

In the following section we present our algorithm that removes the "CC" tag from sentences tagged as a dependency tree and produces a simplified version of the sentence. The simplified version would be grammatically correct and easier to read and understand.

3.2 Pruning Algorithm

We implemented an algorithm that takes the dependency tree in the CONLL Data Format and returns a

plain text with the simplified sentence. If the dependency tree is well-formed, with 100.0% label and dependency tag accuracy, or at least it is correctly tagged for the tags that our algorithm takes into account, the resulting sentence will be grammatically correct.

The algorithm runs through the dependency tree and it makes the following steps.

1. The algorithm removes all the nodes that have as dependency tag the "CC" tag.
2. The algorithm removes all the nodes that have as parent the node removed in 1. The algorithm iterates in 2 while there are more nodes that have had a parent removed.
3. It generates a plain text sentence by removing all the semantic and syntactic information of the dependency tree.

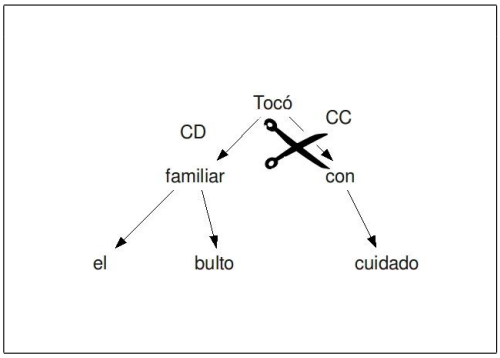

Figure 2: A tagged sentence from AnCora.

Figure 2 shows a very easy example of a sentence: *Tocó el familiar bulto con cuidado*, (in English, *He/She touched the familiar shape carefully*). The resulting sentence must be: *Tocó el familiar bulto*, (in English, *He/She touched the familiar shape*). Our algorithm removes the information about how he/she touched it.

In section 4 we present the evaluation design of our system with two measures of evaluation.

3.3 Evaluation Design

In this section we present, two measures of evaluation, the first one is not a group objective evaluation, it consists in a group of people, they all have university studies. The second evaluation measure consists in a group of children between ten to eleven years old.

3.3.1 Questionnaire for Adults

As a evaluation measure, we surveyed a group of people (20 persons), about how good is the text simplification made by our algorithm. They all have university studies and they all have as their mother tongue Spanish. None know how the simplification algorithm works. We selected 20 sentences from the Ancora corpus. We showed them the whole sentence and the simplified sentence, then we asked them the following questions, they had to answer "yes" or "no":

- Q1: *Is the main idea of the sentence retained?*
- Q2: *Was all the removed information unnecessary?*
- Q3: *Have only details without importance been deleted?*
- Q4: *Do you understand better the simplified sentence than the normal sentence?*

The results of the survey, are presented in Section 4.1.

3.3.2 Questionnaire for Children

As a second evaluation measure, we decided to do a group objective evaluation, carried out with a group of children between ten to eleven years old. There were 24 children, and they had a questionnaire for each pair of children. We selected 20 sentences from the AnCora corpus, we showed them the simplified version and the original version, they had to answer 'yes' or 'no' to the following question for each sentence: *Do you understand better the simplified sentence than the normal sentence?*

The results of the survey, are presented in Section 4.2.

4 RESULTS

In this section we present the global results of our experiment, Showing the results of the two measures of evaluation, and some global statistical results on corpus.

4.1 Results Adult Questionnaire

Table 1 shows the results of the evaluation made by the group of people. In the table we show the answers 'yes' or 'no' for each question.

The first question, Q1: *Is the main idea of the sentence kept?*, is the most important one. The survey give us 67.58% of people that say "yes" for any sentence in question Q1. But it is important to notice that in 50% of the sentences people answer "yes" in 86% or more. We can conclude that most of the people thought that in most of the sentences the main idea, and the meaning, of the sentence is preserved.

Table 1: Results obtained by the survey.

Question	YES	NO
Q1	67.58%	32.42%
Q2	27.66%	72.34%
Q3	46.72%	53.28%
Q4	60.76%	39.24%

If we focus on question Q2: *Was all the removed information unnecessary?*, people thought that not all information was dispensable. It is probably because our algorithm made very aggressive simplifications in many cases. Looking at questions Q1 and Q2, we can see that most people feel that we are loosing some information but they think that the overall meaning is preserved.

Seeing the third question Q3: *Have been only deleted details without importance?*, it is important to notice the differences between Q2 and Q3. The negative answer to this question indicates lower quality of the compressed sentence, but Q2 is more general about the idea that we loose some data but maybe it is not highly important. If we look at the results we can conclude that in some of the phrases where we loose some data, we are not loosing the most important information.

If we focus on the last question Q4: *Do you understand better the simplified sentence than the normal sentence?*. This question asks about how well the people understand the simplified version compared to the normal sentence. Most of the people think that the simplified sentences are easier to read. It is important to notice that some of the sentences are not really difficult to read in the original version and because of that, some people answer "no" to this question.

Finally, as a conclusion of the experiment, we see that most of the people think that the main idea of the sentences is preserved, which is one of our goals, and they also think that the simplified version is easier to read and understand than the original version which is our second goal.

4.2 Results Children Questionnaire

The results of the survey on children are presented on Table 2. We had 240 answers, 20 answers for each sentence by each pair of children. The children answered "yes" in 125 of the 240 cases. Therefore, we have 52.08% of children who believed that the simplified sentence was easier to read than the original version.

We can see the differences between the 4th question Q4 in the first evaluation measure, and the results given by the survey in this evaluation measure. In question Q4 people are not in the group objective, so they can not say that they understand better the sentences because they understand at the same level. In this second evaluation measure the children may have some problems to understand the sentences fluently, so our system can help them to understand the information better. In fact children may have difficulty in understanding even the simplified sentence because they are not able to read some difficult concepts that are presented in the original version and the simplified version.

Table 2: Results obtained by the survey.

Children	YES	NO
24	52.08%	47.92%

We can conclude that our system helps the group of children to understand the sentences better, which is our main goal.

4.3 Overall Statistics in Corpus

In this subsection we show the results obtained after simplifying the whole corpus, using our algorithm sentence by sentence. The AnCora corpus has 3,512 sentences, and the algorithm makes simplification in 2,737 sentences, that is 77.93% of the total. The algorithm did not simplify the whole corpus, because sentences that do not have a "CC" tag are not simplified. The results of the experiment are given in Table 3 which shows the number of wordforms, the average sentence length and the longest sentence length of the original corpus and the simplified corpus.

Table 3: Results on Sentence Length (SL).

	Original	Simplified
Total Wordforms	95,028	58,415
Average SL	27.06 wf	16.63 wf
Longest SL	143 wf	94 wf

5 CONCLUSIONS AND FUTURE WORK

The potentialities of text simplification systems for education, for example, are evident. For students, it is a first step for more effective learning. For people with poor literacy, we see text simplification as a first step towards social inclusion, facilitating and developing reading and writing skills to interact in society. The social impact of text simplification is undeniable.

Our system is a first approximation to an automatic system that runs through dependency trees and returns a simplified version of the sentence parsed. We can conclude that it is possible to simplify correctly texts using dependency parsing, in the particular case of Spanish. The simplified sentence is grammatically correct. But on the other hand, choosing any label, using dependency parsing the algorithms make aggressive simplifications.

We made a simple version of the algorithm and we only focused on the dependency tree. We are working to increase the number of simplification operations and some future work might be oriented towards defining lexical simplification operations like to swap the "difficult" words with a simple synonym using a version of WordNet (Fellbaum, 1998) for Spanish like EuroWordNet (Vossen, 1998).

REFERENCES

Bautista, S., Gervás, P., and Madrid, R. (2009). Feasibility Analysis for SemiAutomatic Conversion of Text to Improve Readability. In *The Second International Conference on Information and Communication Technologies and Accessibility*.

Buchholz, S. and Marsi, E. (2006). Conll-x shared task on multilingual dependency parsing. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL–X)*, pages 149–164.

Canning, Y. (2000). Cohesive simplification of newspaper text for aphasic readers. In *3rd annual CLUK Doctoral Research Colloquium*.

Caseli, H. M., Pereira, T. F., Specia, L., Pardo, T. A. S., Gasperin, C., and M.Aluisio, S. (2009). Building a Brazilian Portuguese Parallel Corpus of Original and Simplified Texts. In *In Proceedings of CICLing*.

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

Chandrasekar, R. and Srinivas, B. (1997). Automatic induction of rules for text simplification. *Knowledge-Based Systems*, 10.

Devlin, S. and Tait, J. (1998). *Linguist Databases*, chapter The use of a Psycholinguistic database in the Simplification of Text for Aphasic Readers, pages 161–173. CSLI.

Devlin, S. and Unthank, G. (2006). Helping aphasic people process online information. In *Assets '06: Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, pages 225–226, New York, NY, USA. ACM.

Fellbaum, C., editor (1998). *WordNet: an electronic lexical database*. MIT Press.

Inui, K., Fujita, A., Takahashi, T., Iida, R., and Iwakura, T. (2003). Text simplification for reading assistance: a project note. In *Proceedings of the second international workshop on Paraphrasing*, pages 9–16, Morristown, NJ, USA. Association for Computational Linguistics.

Klebanov, B. B., Knight, K., and Marcu, D. (2004). Text simplification for information-seeking applications. In *On the Move to Meaningful Internet Systems, Lecture Notes in Computer Science*, pages 735–747. Springer Verlag.

Kübler, S., McDonald, R. T., and Nivre, J. (2009). *Dependency Parsing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Max, A. (2006). Writing for language-impaired readers. In *CICLing*, pages 567–570.

Mcdonald, R. K. R., Nilsson, J., Riedel, S., and Yuret, D. (2007). The conll 2007 shared task on dependency parsing.

Palomar, M., Civit, M., Díaz, A., Moreno, L., Bisbal, E., Aranzabe, M., Ageno, A., Martí, M., and Navarro, B. (2004). 3lb: Construcción de una base de datos de árboles sintáctico–semánticos para el catalán, euskera y español. In *Proceedings of the XX Conference of the Spanish Society for Natural Language Processing (SEPLN)*, pages 81–88. Sociedad Española para el Procesamiento del Lenguaje Natural.

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

Siddharthan, A. (2002). Resolving attachment and clause boundary amgiguities for simplifying relative clause constructs. In *Proceedings of the Student Research Workshop, 40th Meeting of the Association for Computacional Linguistics*.

Siddharthan, A. (2003). *Syntactic Simplification and Text Cohesion*. PhD thesis, Research on Language and Computation.

Snow, C. E., States., U., Science, and Corporation), T. P. I. R. (2002). *Reading for understanding : toward an R&D program in reading comprehension / Catherine Snow*. Rand, Santa Monica, CA :.

Taulé, M., Martí, M., and Recasens, M. (2008). AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of 6th International Conference on Language Resources and Evaluation*.

Vossen, P., editor (1998). *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, MA, USA.

Williams, S., Reiter, E., and Osman, L. M. (2003). Experiments with discourse-level choices and readability. In *In Proceedings of the European Natural Language Generation Workshop (ENLG) and 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL03)*, pages 127–134.