Making it Simplext: Implementation and Evaluation of a Text Simplification System for Spanish

HORACIO SAGGION, Universitat Pompeu Fabra
SANJA ŠTAJNER, University of Wolverhampton
STEFAN BOTT, Universität of Stuttgart
SIMON MILLE, Universitat Pompeu Fabra
LUZ RELLO, Carnegie Mellon University
BILJANA DRNDAREVIC, Universitat Pompeu Fabra

The way in which a text is written can be a barrier for many people. Automatic text simplification is a natural language processing technology which, when mature, could be used to produce texts which are adapted to the specific needs of particular users. Most research in the area of automatic text simplification has dealt with the English language. In this paper, we present results from the Simplext project which is dedicated to automatic text simplification for Spanish. We present a modular system with dedicated procedures for syntactic and lexical simplification which are grounded on the analysis of a corpus manually simplified for people with special needs. We carried out an automatic evaluation of the system's output, taking into account the interaction between three different modules dedicated to different simplification aspects. One evaluation is based on readability metrics for Spanish and shows that the system is able to reduce the lexical and syntactic complexity of the texts. We also show, by means of a human evaluation, that sentence meaning is preserved in most cases. Our results, even if our work represents the first automatic text simplification system for Spanish which addresses different linguistic aspects, are comparable to the state-of-the art in English Automatic Text Simplification.

Categories and Subject Descriptors: I.2.7 [**Natural Language Processing**]: Text Analysis

ACM Reference Format:
ACM 1, 1, Article 1 (April 2014), 37 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

The last two decades have seen drastic changes in the way we access information. With the availability of the Internet, the average person has access to much more information than at any other time in history. In parallel, access to information has become more and more important for most people's everyday lives. Also the way in which we access information has changed: everyone who has access to a computer connected to the Web has much more content available than one can possibly process. The selection of information is more difficult, and at the same time more crucial, than ever. There are also other issues with the access to information, one of which we want to address here: textual information may be written in a style which makes the content hard to understand. This may affect user groups like non-native speakers, persons with a low literacy rate, and people with reading or cognitive impairments. Even if some organizations, such as the United Nations and the World Wide Web Consortium Web Accessibility Initiative (W3C WAI), have stressed this problem, the general awareness

of it is still very low and most of the information on the Internet is not published with special needs taken into consideration.

One way to address this problem is by manually adapting existing textual content for people with special needs. This approach has been taken by several organizations in different countries which publish specially prepared material. However, this approach is very costly, both in terms of time and money. One possible remedy is to apply Natural Language Processing techniques to simplify text automatically. Automatic text simplification is a technology used to adapt the content of a text to the specific needs of particular individuals or target populations in a way that the text becomes more readable and understandable for them. The adapted text will most probably suffer from information loss and a too simplistic or boring style, which is not necessarily a bad thing if the original message can in the end be transmitted to the reader. Text simplification has also been suggested as a potential pre-processing step for making texts easier to handle by generic text processors such as parsers [Chandrasekar et al. 1996], or to be used in specific information access tasks such as information extraction [Klebanov et al. 2004]. But our research is more related to the first objective of making texts more accessible to specific users. This is certainly more challenging than the second use of simplification because the output will necessarily be evaluated with the same yardstick that human written texts are evaluated with.

The interest in automatic text simplification has grown in recent years and in spite of the many approaches and techniques proposed, there is still space for improvement. Consider the problem of lexical simplification in English which was proposed in a recent natural language processing evaluation for example, where most systems equipped with sophisticated tools and resources were unable to beat a simple baseline [Specia et al. 2012]. The growing interest in text simplification is evidenced by the number of languages which are targeted by researchers around the globe. Simplification systems and simplification studies do exist at least for English [Chandrasekar et al. 1996; Siddharthan 2002; Woodsend and Lapata 2011a; Wubben et al. 2012; Glavaš and Štajner 2013; Siddharthan and Angrosh 2014], Brazilian Portuguese [Aluísio et al. 2008; Specia 2010; Paetzold and Specia 2013], Japanese [Inui et al. 2003], Dutch [Ruiter et al. 2010], French [Seretan 2012; Brouwers et al. 2014], Italian [Dell'Orletta et al. 2011; Barlacchi and Tonelli 2013], and Basque [Aranzabe et al. 2012; Aranzabe et al. 2013].

In this paper we present Simplext, the first fully-fledged text simplification system for the Spanish language, together with a thorough evaluation of the system's output and comparison with state-of-the-art simplification systems for English. The research was conceived to provide an automatic tool able to adapt text content to the specific needs of people with cognitive disabilities [Saggion et al. 2011; Bott and Saggion 2014]; however our work has now been extended to address general issues in text simplification. Our system is made up of components for reducing the syntactic complexity of sentences, deleting unnecessary information (currently only parenthetical information), rewriting numbers, normalizing reporting verbs, and substituting difficult words by their simpler synonyms. Rule-based and corpus-based techniques are combined to computationally model different simplification phenomena. The evaluation we present targets three aspects which we consider essential for measuring the objectives of the simplification system: reduction of text complexity, meaning preservation during simplification, and production of grammatical output.

Here we evaluate the system for the first time as a whole taking into account interactions among the 3 modules. We also describe an evaluation with the end users. We compare our results to similar works on the simplification of English which represent the state-of-the-art and show that we can obtain comparable results for a language

which counts with fewer resources and has so far received less attention. From the insights obtained here we propose avenues for pushing forward our agenda on text simplification research.

The rest of the paper is organized as follows: in Section 2 we present related work which is relevant for our project, in Section 3 and 4 we present the system design and the system components. Sections 5 and 6 describe the evaluation design and the evaluation results, which we then discuss in Section 7. We close the paper with conclusions and and outlook on future work in Section 8.

2. RELATED WORK

The text simplification problem has been studied from various angles. In this section we present related work which treats (i) guidelines for the production of simplified material, (ii) text simplification for target user groups, (iii) general concerns of automatic text simplification together with currently used methods for syntactic and lexical simplification, and (iv) evaluation of automatic text simplification systems.

2.1. Simplification Guidelines for Human Editors

In the early nineties, Basic English, a version of English with reduced vocabulary and grammar [Ogden 1937] was proposed as a tool to facilitate international communication. Since the late nineties, several initiatives raised awareness of the complexity of the vast majority of written documents and the difficulties they pose to people with any kind of reading or learning impairments. These initiatives proposed various guidelines for writing in a simple and easy-to-read language which would be equally accessible to everyone, e.g. the "Make it Simple" European Guidelines for the Production of Easy-to-Read Information for people with Learning Disability" [Freyhoff et al. 1998], the Mencap's "Am I making myself clear?" guidelines for accessible writing [Mencap 2002] and the Guidelines for Easy-to-Read Materials [Nomura et al. 1997]. An extensively discussed question is how much the needs of different target populations overlap or not [Nomura et al. 1997]. It is generally agreed that there are more factors which unify different target groups than those which separate them [Nomura et al. 1997]. All of those guidelines share similar instructions for accessible writing. For example, they all advise the writer to use the active voice instead of passive, use short, simple words and omit unnecessary words, write short sentences and cover only one main idea per sentence, etc. None of them, however, specifies any language- or user group-dependent instructions. Simplification guidelines have also been proposed in the industry to make technical communication more standard and less ambiguous [Barthe et al. 1999].

2.2. Text Simplification for Readers

Some work on automatic simplification has aimed at creating generic simplification tools, without considering the special needs of specific user groups [Chandrasekar et al. 1996; Siddharthan 2002; Coster and Kauchak 2011b]. This is justifiable, since many aspects of text complexity affect a large range of users with reading difficulties. For example, long and syntactically complex sentences are generally hard to process. Some particular sentence constructions, such as syntactic constructions which do not follow the canonical subject-verb-object (e.g. passive constructions) may be an obstacle for people with aphasia [Devlin 1999] or autism spectrum disorder (ASD) [Martos et al. 2012]. The same is true for very difficult or specialized vocabulary. Infrequent words make the text difficult to comprehend by people with aphasia [Devlin 1999], and ASD [Norbury 2005; Martos et al. 2012]. When it comes to students with intellectual disability, existing studies show contradictory findings: Fajardo *et al.* [2014] found no effects of the word frequency on the comprehension scores (neither literal

nor inferential) in students with intellectual disability, while when studying Web site adaptations following the "Make it Simple" guidelines [Freyhoff et al. 1998], Karreman *et al.* [2007] reported both literal and inferential comprehension scores higher in the adapted version.

But there are also aspects which are quite specific to certain groups of readers. Language learners, for example, may have a good capacity to infer information, although they may have a very restricted lexicon and may not be able to understand certain grammatical constructions. Dyslexic readers, in turn, do not have a problem with language understanding *per se*, but with the understanding of the written representation of language: in addition, readers with dyslexia were found to read faster when using more frequent and shorter words [Rello et al. 2013b], graphical schemes [Rello et al. 2012], or certain number representations [Rello et al. 2013]. People with intellectual disabilities have problems processing and retaining large amounts of information [Feng 2009; Fajardo et al. 2014]. Several studies have shown that long texts can affect self-efficacy and reading motivation in students with intellectual disability [Morgan and Moni 2008; Gómez 2011]. The study of Fernsbacher and Faust [1991] indicated that adult poor readers have difficulties in suppressing irrelevant information. Therefore, text simplification systems aimed at those target populations should not only simplify the written content (by using simpler synonyms and splitting long and complex sentences into several simple ones), but should also perform some kind of content reduction (discarding irrelevant information) in order to reduce the memory load necessary for understanding the given text.

There have been approaches concentrating on language learners and foreign readers [Crossley and McNamara 2008], children [De Belder and Moens 2010; Vu et al. 2014], aphasic readers [Canning et al. 2000], people with ASD [Orasan et al. 2013], people with dyslexia [Rello et al. 2013a; Rello and Baeza-Yates 2014], people with cognitive problems [Carroll et al. 1998; Max 2006; Feng 2009], people who need assisted reading [Inui et al. 2003] or people with a generally low literacy rate [Aluísio et al. 2008; Watanabe 2010]. There have also been attempts to simplify very complex text genres for average readers, for example in the case of patent texts [Bouayad-Agha et al. 2009].

2.3. Automatic Text Simplification (ATS)

Automatic text simplification can be directed to human readers or be used as a preprocessing step for other NLP tasks, such as parsing [Chandrasekar et al. 1996], machine translation [Chandrasekar 1994], semantic role labelling [Vickrey and Koller 2008], or information retrieval [Klebanov et al. 2004; Ong et al. 2007]. It is worth pointing out a series of facts which make text simplification a somewhat special NLP task and which pose specific challenges. First of all, text simplification is not a clearly defined monolithic task, but rather a series of coordinate tasks which combine for a common goal. It is similar to, but sufficiently different from, other NLP tasks, such as automatic translation [Lopez 2008], summarization [Saggion and Poibeau 2013], sentence compression [Clarke and Lapata 2006] and paraphrasing [Barzilay and Lee 2004]. Although there may be close similiarities at first sight, the definition of some of these sub-tasks may also differ from other NLP tasks in important ways. Extractive text summarization, for instance, tries to retain the most informative parts of an input text while simplifying content reduction aims at eliminating text parts with superfluous information. The two things seem to be the two sides of the same coin, but content reduction has to be more careful in not eliminating steps in the argumentative structure of a text. Users of text summarization systems must often compensate a deficit of text coherence in the output texts with their cognitive ability to reconstruct logical connections from their real word knowledge. This cannot be expected from the target users of text simplification. Further on, texts can be simplified at very different

linguistic levels: simplification may try to reduce sentence length, syntactic embedding depth, lexical complexity, lexical variety, the level of detail of the transmitted information, etc. Where sentence length reduction is concerned, sentence compression techniques should be adapted here since for text simplification the material taken out from one sentence should be used to create new linguistic units [Angrosh et al. 2014]. Even extra-linguistic factors can be used to make reading easier, for example by explaining unfamiliar words or linking parts of the text to external resources like dictionaries or encyclopedia.

In the last years the availability of the Simple English Wikipedia has made a big impact [Coster and Kauchak 2011b]. Even if the Simple English Wikipedia (SEW hereafter) is not fully parallel to the "ordinary" English Wikipedia (EW), the SEW covers a subset of the EW and out of this subset parallel sentences can be extracted, which to a large extent express the same information. Thus a quasi-parallel corpus can be extracted, which allows for a range of purely data-driven approaches. This new dataset allows for the use of techniques that were not applicable before because of the lack of sufficient data. For languages other than English, it is, however, still relatively difficult to obtain large-scale parallel resources.

2.3.1. Syntactic Simplification. Syntactic simplification tries to reduce the structural complexity of sentences, i.e. sentence length and syntactic embedding depth. The first approaches to syntactic simplification were based on linguistic intuitions and were implemented as hand-written rules [Chandrasekar et al. 1996; Siddharthan 2002]. Later approaches gradually employed more data-driven methods [Chandrasekar and Srinivas 1997; Petersen and Ostendorf 2007]. In languages other than English, syntactic simplification approaches are still rule-based [Aranzabe et al. 2012; Aranzabe et al. 2013; Orasan et al. 2013]. The Brazilian PorSimples project [Aluísio et al. 2008; Aluísio and Gasperin 2010] created a dataset of original and simplified texts in Portuguese which allowed researchers to study and implement a simplification system. The same dataset, although small, could also be used for experiments with statistical machine translation software [Specia 2010]. The more recent availability of the dataset extracted from the EW and the SEW has led to a series of experiments and approaches in simplification. Zhu *et al.* [2010] used a tree-based simplification model which is derived from statistical machine translation (SMT) techniques to simulate four simplification operations: split, drop, copy, and reorder. Coster and Kauchak [2011a] used standard SMT software and applied it to the simplification problem with the addition of a dedicated and task-specific deletion module. Also Woodsend and Lapata [2011b] and Wubben et al. [2012] treated text simplification as a translation problem from "normal" language to simplified language. They used quasi-synchronous grammars and linear integer programming for this purpose. They also compared the use of the revision histories of the SEW to learning from bi-text and found that the use of revision histories yields better results.

In the last years, there have also been several hybrid approaches that give better results, such as the data-driven model from Narayan and Gardent [Narayan and Gardent 2014] that combines deep semantics and machine translation, or models that combine data-driven and rule-based approaches [Siddharthan and Angrosh 2014; Angrosh and Siddharthan 2014].

2.3.2. Lexical Simplification. Lexical simplification is usually understood as a word-substitution task, where the goal is to find a synonym which is in some sense simpler than the original word. This task requires a resource which allows the lookup of synonyms. WordNet has often been used to this end [Carroll et al. 1998; Lal and Rüger 2002; Burstein et al. 2007], but synonym dictionaries [Bautista et al. 2011] or thesauri can also be used. The most common metric for lexical simplicity used in these

approaches is word frequency, since frequency can be assumed to correlate well with familiarity. An additional factor is word length: long words tend to be harder to read [Rello et al. 2013b]. For this reason word length can be taken as an additional or alternative predictor for perceived lexical difficulty. It was found to be a decisive factor by Flesch [1948] and it is used in the calculation of the Flesch-Kincaid formula (Section 2.4.2). Lexical simplification has to cope with the problem of lexical ambiguity and the suitability of a synonym depends on the specific word sense of a target word. For this reason De Belder *et al.* [2010] proposed the application of word sense disambiguation for lexical substitution.

Recently some purely data-driven approaches have exploited the availability of the SEW: Yatskar *et al.* [2010] used edit histories from the SEW and the combination of SEW and EW in order to create a set of lexical substitution rules. Biran *et al.* [2011] also used the SEW/EW combination (without the edit history of the SEW), in addition to the explicit sentence alignment between SEW and EW to identify pairs of words which occur in similar contexts using WordNet as a filter for inducing lexical substitution rules (e.g., "canine" can be replaced by "dog"). In this latter approach, a form of word sense disambiguation was carried out by comparing, using a distance measure, candidate word vectors in context. Here the distance between the target context and a potential lexical substitute was used to filter out potentially harmful rule applications.

It should be noted that the machine translation based approaches we mentioned above [Coster and Kauchak 2011a; Specia 2010] as well as the hybrid approaches [Narayan and Gardent 2014; Siddharthan and Angrosh 2014; Angrosh and Siddharthan 2014] are also able to handle lexical simplification, even if implicitly, since the translation model maps words from the non-simplified language to words of the simplified language.

2.4. Evaluation of the Automatic Text Simplification Systems

The ideal way of evaluating ATS systems aimed at providing more accessible information to a certain target population would be to test its effectiveness on their reading time and comprehension. However, as the access to a specific target population might be difficult, most of the studies perform only the expert (non-final user) evaluation of their systems, providing the human scores for grammaticality, meaning preservation and simplicity of the system's output. Given that such evaluation is performed only on the sentence level, it is usually combined with the automatic evaluation of simplicity of the whole text measured in terms of its readability. Data-driven ATS systems which have the possibility of comparing the system's output with the gold standard (manual simplification) additionally use some of the most common machine translation (MT) evaluation metrics.

2.4.1. Expert (Non Final User) Evaluation. The output of the ATS systems is commonly evaluated by human judgments of its grammaticality (fluency), meaning preservation (adequacy) and simplicity, e.g. [Wubben et al. 2012; Feblowitz and Kauchak 2013; Coster and Kauchak 2011a; Angrosh and Siddharthan 2014]. Fluency measures grammatical correctness of the output, simplicity measures how simple the output is, and the meaning preservation measures how well the meaning of the simplified sentence corresponds to the meaning of the original sentence. All three scores are usually measured on a five-point Likert scale, the exceptions being [Narayan and Gardent 2014] with a 0–5 scale, and [Glavaš and Štajner 2013] with a 1–3 scale. In all cases, the higher score indicates the better output.

2.4.2. Readability Indices. Since the second half of the last century, over two hundred readability formulae have been developed for the English language [DuBay 2004].

They were initially used to assess the grade level of textbooks, but later they were also adapted for different domains and purposes, e.g. to measure readability of technical manuals [Smith and Senter 1967] and US health-care documents intended for the general public [McLaughlin 1969]. In spite of various criticisms for using only features like average sentence and word length, some of the oldest readability formulae (e.g., the Flesch Reading Ease score [Flesch 1948]) are still widely used, due to their simplicity and good correlation with reading tests.

Recent developments in natural language processing offered the possibility for automatic computation of new readability formulae which use more sophisticated lexical and syntactic features. The works on statistical readability assessment [Si and Callan 2001; Collins-Thompson and Callan 2005] used unigram language models for estimating the grade level of US text books. Schwarm and Ostendorf [2005] and Petersen and Ostendorf [2009] used statistical language modeling and support vector machines to show that more complex features (e.g., average height of the parse tree, average number of noun and verb phrases) give better readability prediction than the traditional Flesch-Kincaid readability formula. Feng *et al.* [2009] introduced some new cognitively motivated discourse-level features (e.g. entity mentions, lexical chains, etc.) showing that they are better correlated with the comprehension of people with intellectual disabilities than the traditionally used Flesch-Kincaid Grade Level index [Kincaid et al. 1975]. In spite of those findings, ATS systems are commonly evaluated with the traditional readability formulae, such as the Flesch-Kincaid Grade level index [Woodsend and Lapata 2011a; Wubben et al. 2012; Glavaš and Štajner 2013; Vu et al. 2014] or the Flesch Reading Ease Score [Zhu et al. 2010; Woodsend and Lapata 2011a], probably due to the fact that they can easily be computed automatically with a high precision.

While all of the aforementioned formulae were made for assessing the level of English texts, similar studies have started to appear for other languages as well: German [Vor der Brück et al. 2008], Portuguese [Aluísio et al. 2010], French [François and Watrin 2011], Italian [Dell'Orletta et al. 2011], Swedish [Roll et al. 2007], and Basque [Gonzalez-Dios et al. 2014]. However, there have been no similar studies for the Spanish language. Therefore, we used some traditional Spanish readability formulae [Spaulding 1956; Anula 2007] and adapted them to be computed automatically (see Section 5.1).

2.4.3. MT Evaluation Metrics. Recently, many studies which propose data-driven ATS systems include an additional assessment of the systems' output by comparing it with gold standard manual simplifications, borrowing the MT evaluation metrics such as BLEU (as, for example, in [Specia 2010; Zhu et al. 2010; Woodsend and Lapata 2011a; Coster and Kauchak 2011a; Wubben et al. 2012; Feblowitz and Kauchak 2013; Narayan and Gardent 2014; Vu et al. 2014]), TERp (as in [Woodsend and Lapata 2011a; Vu et al. 2014]), or NIST (as in [Specia 2010; Zhu et al. 2010]).

BLEU [Papineni et al. 2002] is the most widely used MT evaluation metric which measures similarity between the system's output and a human reference. It is based on the exact n-gram matching and heavily penalises word reordering or sentence shortening. NIST [Doddington 2002] is, like BLEU, also based on exact n-gram matching, with the difference that it gives different weights to different n-grams (depending on how likely they are to occur) and that its brevity penalty is less severe (small differences in the length of the system's output and the human reference do not impact the overall score as much as in BLEU). TERp [Snover et al. 2009] measures the number of 'edits' needed to transform the MT output (automatically simplified version of the original sentence in our case) into the reference translation (human simplified sentence in our case). TERp is an extension of TER – Translation Edit Rate [Snover et al. 2006] that utilizes phrasal substitutions (using automatically generated paraphrases), stem-

ming, synonyms, relaxed shifting constraints and other improvements [Snover et al. 2009]. The higher the value of TERp (and each of its components), the less similar the manually simplified and the automatically simplified sentences are.

We opted not to use those MT metrics for the evaluation of our system as it is known that those metrics are appropriate only for comparing systems of similar architectures and are not meant for comparing systems of radically different architectures. In our case, we need to compare the output of the system which performs only lexical and syntactic simplification with the manual simplification which, in addition to those two operations, also includes a high number of strong paraphrasing, summarizations, and deletions [Drndarevic et al. 2013; Štajner et al. 2013a; Štajner 2014]. The sentence-wise BLEU score between original and manually simplified sentences was reported to be as low as 0.17 [Štajner 2014].

We also did not adopt content-based metrics which are generally used in the evaluation of text summarization systems such as ROUGE [Lin 2004], FRESA [Saggion et al. 2010], or PYRAMIDS [Nenkova and Passonneau 2004] since our system is very conservative about the application of content reduction, a single manual simplification (the ideal simplification) whereas ROUGE and PYRAMIDS usually require more than one ideal target to compare to.

3. THE SIMPLEXT TEXT SIMPLIFICATION APPROACH

Our approach to text simplification is modular, in order to respond to the fact that text simplification can be applied at different linguistic levels. It is also restricted by the availability of parallel (original-simplified) data in Spanish. As we explained in Section 2.3.1, in the last years purely empirical methods of text simplification have become very popular for English. Purely data-driven approaches, however, require very large data collections from which they can learn, but to the best of our knowledge there is no such dataset for Spanish. Within the Simplext project we compiled a corpus of 200 news texts and created manual simplifications for them. The corpus contains news from four domains: national news, international news, society and culture. The simplified part of this corpus is based on very specific simplification recommendations [Anula 2011] for human editors and represents the kind of simplification we want to produce faithfully.

Examples of original sentences and their manual simplifications are shown in Table I. Example 1 shows an instance of simplification of the vocabulary. The word *sucursal* (branch) in the original sentence is replaced by its synonym *oficina* (office) (10 times more frequent according to the Spanish Royal Academy's frequency list)[1] in the simplification. It also shows a syntactic transformation in order to have a simplified sentence with the SVO syntactic pattern, which in Spanish is the natural order of syntactic elements in a sentence.

Example 2 in addition to the replacement of *sucursal* by *oficina* shows an interesting case of summarization of detailed information: the replacement of *en 28 países del mundo* (in 28 countries around the world) by *en muchos países del mundo* (in many countries around the world). The example also presents a splitting operation.

Case 3 exemplifies a delete operation by which a full sentence is not included in the simplified text. This is a frequent operation in the Simplext corpus with over 70% of simplified documents presenting it [Štajner et al. 2013b].

Example 4 shows a splitting operation together with the replacement of the verb *arranca* (starts) by its more common synonym *comienza* (begins) (7 times more frequent according to the Spanish Royal Academy's frequency list).[1]

[1] Royal Spanish Academy: CREA Database [online]. Spanish Reference Corpus. http://www.rae.es

Table I: Examples of Manual Simplifications in Simplext.

Ex.	Original	Simplified
1	*Abre en Madrid su primera* **sucursal** *el mayor banco de China y del Mundo.* (Opens in Madrid its first branch the biggest bank of China and the World.)	*El banco más importante de China y del mundo abre una* **oficina** *en Madrid.* (The most important bank of China and the world opens an office in Madrid.)
2	*El ICBC ha abierto ya 203* **sucursales** *en un total de 28 países de todo el mundo, también en España desde este lunes.* (The ICBC has opened 203 branches in a total of 28 countries around the world, also in Spain since this Monday.)	*El Banco de China tiene* **oficinas** *en muchos países del mundo. Ahora, también tiene una* oficina *en España.* (The Bank of Chine has offices in many countries around the world. Now it also has an office in Spain.)
3	*Como muestra de su envergadura, según datos de 2009, el ICBC tenía en nómina a un total de 386.723 empleados, sólo en China, en un total de 16.232 sucursales.* (As a sign of its size and according to data from 2009, the ICBC had a total of 386,723 employees in China only, in 16,232 branches.)	
4	**Arranca** *la liga masculina de Goalball, el único deporte específico para ciegos.* (Starts the men's league of Goalball, the only specific sport for the blind.)	**Comienza** *la liga masculina de Goalball. El Goalball es el único deporte específico para ciegos.* (Begins the men's league of Goalball. Goalball us the only specific sport for the blind.)
5	*La ONU prevé el fin de muertos por malaria para 2015.* (The UN expects the end of dead by malaria for 2015.)	*La ONU cree que ninguna persona morirá por malaria a partir de 2015. La ONU es la Organización de las Naciones Unidas. La malaria es ua enfermedad que se transmite gracias a un mosquito.* (The UN believes that nobody will die of malaria from 2015. The UN is the United Nations Organization. Malaria is a decease transmited by a mosquito)

Example 5 is a case of clarification, where the human editor includes "definitions" of "difficult" terms such as the abbreviation *ONU* and the term *malaria*.

With a corpus size of 200 pairs of documents it was clearly not possible to apply purely data-driven methods. For this reason and in order to take advantage of this valuable material, we carried out two corpus studies: the first tried to isolate as much as possible the "simplification operations" produced by human editors [Bott and Saggion 2011; 2014], and the second concentrated specifically on lexical changes [Drndarevic and Saggion 2012]. Identifying and isolating human changes found in the corpus is not a trivial task, since human editors tend to produce rather strong re-wordings of the content in the original text, instead of applying clear-cut editing steps (a thing that a computer would do and that our simplification system is expected to perform). The simplification guidelines on the basis of which the editors worked did not list very precise operations, either. They instead gave recommendations which could be interpreted by humans (and took advantage of the editors' experience and creativity), and were hard to translate into computational operations. These findings convinced us that the creation of an annotation scheme was necessary, which could capture and classify the operations observed in the parallel corpus as neutrally as possible. The initially available sample of the Simplext corpus (145 sentence pairs of original text with manual simplifications) was then annotated according to our annotation scheme.

This corpus study was used as the basis for the design of the system components and its general architecture. As it will be shown in the next sections, one of the components of the system is a rule-based procedure that performs syntactic simplification, like previous work for English (e.g. [Siddharthan 2011]) our system receives as input a dependency graph/three (both syntactic dependencies and precedence relations are present in the input representation). However and unlike previous work, our approach is transductive – as opposed to transformation-based – in that the input structure is always kept while new structures (i.e. dependency graphs) are generated by the iterative application of a set of rules. Thus, the original input is always available to generate new dependency graphs which serve to produce the final simpler sentences.

4. COMPONENTS OF THE SIMPLEXT SYSTEM

The Simplext system consists of three modules: a syntactic simplification component, a synonym-based simplification component, which uses a thesaurus and distance measures from distributional semantics, and a rule-based lexical simplification component. The choice and the design of the modules was based on the two initial corpus studies we already mentioned.

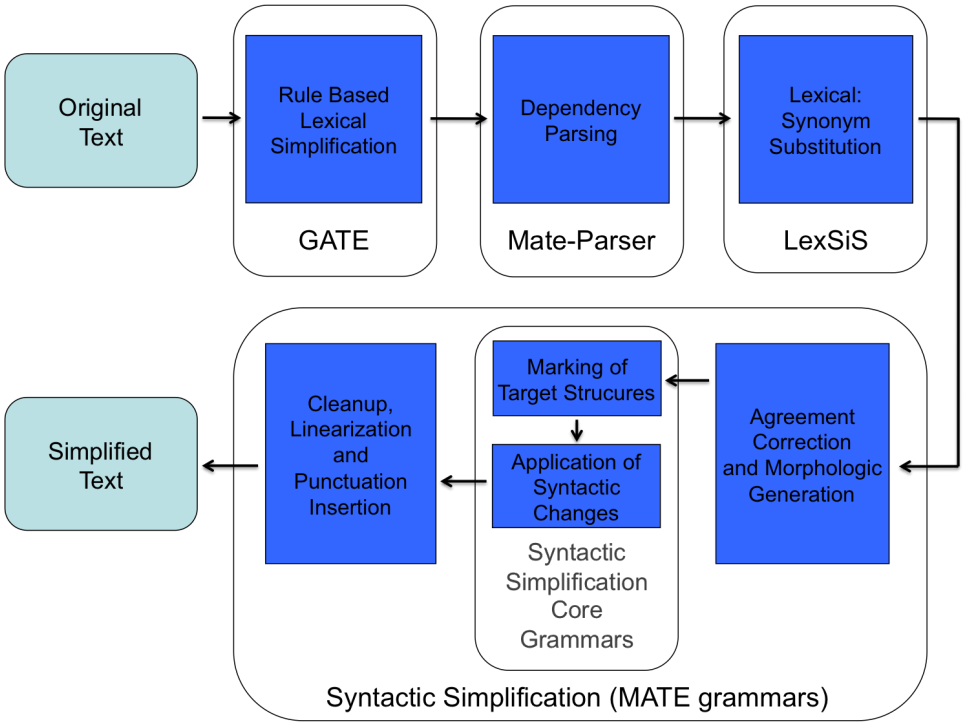

Fig. 1: The Architecture of the Simplext System

The Simplext system is implemented as a pipeline, as represented in Figure 1. The text is first dependency parsed with Bohnet's MATE parser [Bohnet 2009], then lexical simplification and syntactic simplification are applied in sequence. The lexical simplification process outputs lemma forms for substituted words, which makes morphological generation of inflected forms necessary. Morphological generation is integrated in the syntactic simplification module, as part of the MATE graph transduction framework

[Bohnet et al. 2000]. In Figure 2 we present an interface of the system as a simplification gateway. Therein the user can paste a Spanish text and receive a simplification upon pressing the “Simplificar texto” button.

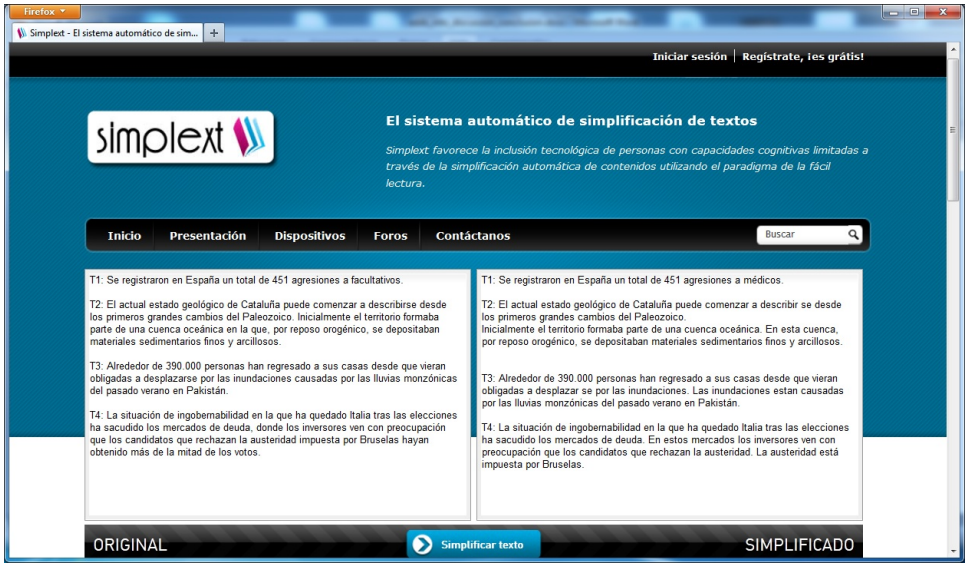

Fig. 2: Simplext Simplification Portal

The figure shows how the sentences from four different sources were transformed into their simplifications. The system is hosted as an Amazon service which we can activate upon request (note that hosting the service is expensive so we only activate it for specific demonstrations; the system has been deployed at the following address: http://www.simplext.net).

4.1. Syntactic Simplification

The syntactic simplification component utilizes of a hand-written computational grammar and focuses on the reduction of structural complexity. Several types of sentence splitting operations are performed; in particular, we turn subordinate and coordinate structures, such as relative clauses, gerund constructions and VP coordinations, into separate sentences, producing shorter and syntactically less complex outputs. The following pair of original (1a) and simplified (1b) sentences exemplify the simplification of a participle and a clausal coordination construction.

(1) a. *El primer encuentro dedicado a esta iniciativa será el partido inaugural, celebrado hoy en Doha con los equipos de Qatar y Uzbekistán, y los dos siguientes duelos de Qatar estarán también dedicados a la campaña del fútbol asiático contra el hambre.*
(The first encounter dedicated to this initiative will be the opening match, celebrated today in Doha with the teams of Quatar and Uzbekistan, and the following two encounters of Quatar are also dedicated to the campaign of Asian football against hunger.)

(1) b. *El primer encuentro dedicado a esta iniciativa será el partido inaugural. Este partido está celebrado hoy en Doha con los equipos Qatar y*

Uzbekistán. Los 2 siguientes duelos de Qatar estarán también dedicados a la campaña del fútbol asiático contra el hambre.
(The first encounter dedicated to this initiative will be the opening match. The match is celebrated today in Doha with the teams of Quatar and Uzbekistan. The following 2 encounters of Quatar are also dedicated to the campaign of Asian football against hunger.)

The pair (2a) and (2b) shows the simplification of a relative clause.

(2) a. *Los vecinos pueden acercarse a las unidades móviles, que se instalarán en treinta avenidas de la ciudad.*
(The neighbors can approach the mobile units, which will be installed on thirty avenues through the city.)

(2) b. *Los vecinos pueden acercarse a las unidades móviles. Estas unidades se instalarán en muchas avenidas de la ciudad.*
(The neighbors can approach the mobile units. These units will be installed on thirty avenues through the city.)

In order to transform (1a) into (1b) and (2a) into (2b), the syntactic simplification module operates on syntactic dependency trees, and tree manipulation is modeled as graph transduction. The graph transduction rules are implemented in the MATE framework [Bohnet et al. 2000], in which the rules are gathered in grammars that apply in a pipeline: the first grammar applies to an input as shown in Figure 3[2], and then each grammar is applied to the output produced by the previous grammar. There are currently 8 grammars (and around 140 rules):

— 2 grammars that deal with the lexical substitutions performed during lexical simplification;
— 3 grammars that perform the syntactic simplification;
— 3 small grammars for cleaning the output and returning a well formed sentence.

First of all, the **lexical substitution grammars** control the syntactic agreements between the substitute words and the original words of the sentence. For instance, if a masculine noun is replaced by a feminine one, we have to make sure that the gender of the determiners, adjectival modifiers, or attributes is changed accordingly. It is also the case when an invariant element is replaced by one which has to agree in gender or number. For instance, in Figure 3, the word *treinta* (thirty) is invariant, but is replaced by a variant quantifier *mucho* (a lot) by the lexical simplification module, as indicated in the attribute-value pairs associated to the node. The first grammar gets the morpho-syntactic features from the governor (in this case its gender and number) and prepares the node for a two-level morphology model or a full form dictionary, as shown in Figure 4(a); the second grammar generates the correct form, Figure 4(b).

Second, the **syntactic simplification grammars** modify the structure of the sentences. Five types of syntactic simplification take place (see examples (1) and (2) at the beginning of this section for illustration):

— participial modifiers are separated from their governing noun to form a new sentence;
— non-defining relative clauses preceded by a comma or the ones which modify an indefinite noun are also separated from it to form a new sentence;
— quoted direct objects are systematically positioned *after* the speech verb that introduces the quote;

[2]For the sake of clarity, we do not show the precedence relations between the words, but the word order is kept all through the process.

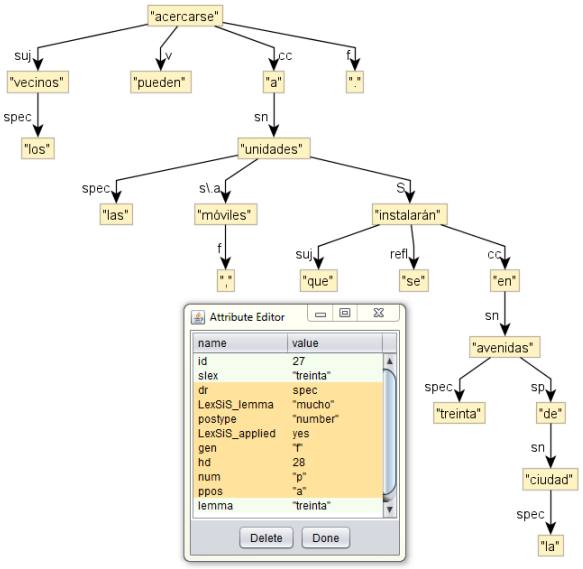

Fig. 3: A syntactic input structure corresponding to example (2a).

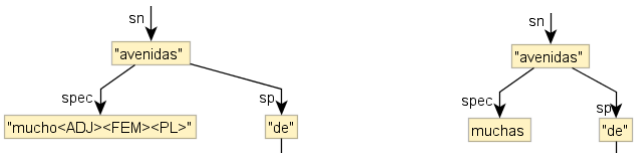

Fig. 4: Sample partial output structures of the lexical substitution grammars

— sentences which contain coordinated main verbs are split (one sentence per verb);
— sentences with long coordinated objects are split.

Syntactic simplification takes place in three steps, which correspond to three grammars. The first grammar identifies all possible simplifications and marks the concerned nodes in the syntactic tree. The second grammar chooses the simplifications to be performed, in order to avoid that more than one applies to the same subtree; for instance, a coordination of main verbs will only be split if none of their objects is involved in a coordination that triggers a simplification. Each type of simplification is associated to a set of transformations: add an auxiliary *estar* (be), change the label of a node, duplicate a noun, add a determiner, invert the positions of some chunks or remove some nodes, etc. Figure 5 shows the output of the second grammar for the structure of Figure 3. A chunk corresponding to the whole relative clause is defined, and the relative pronoun *que* (that) contains all the necessary information (as attribute and values) for the modifications to take place: which rule has applied, the fact that the node label has to be changed, the name, gender, number of the antecedent, the

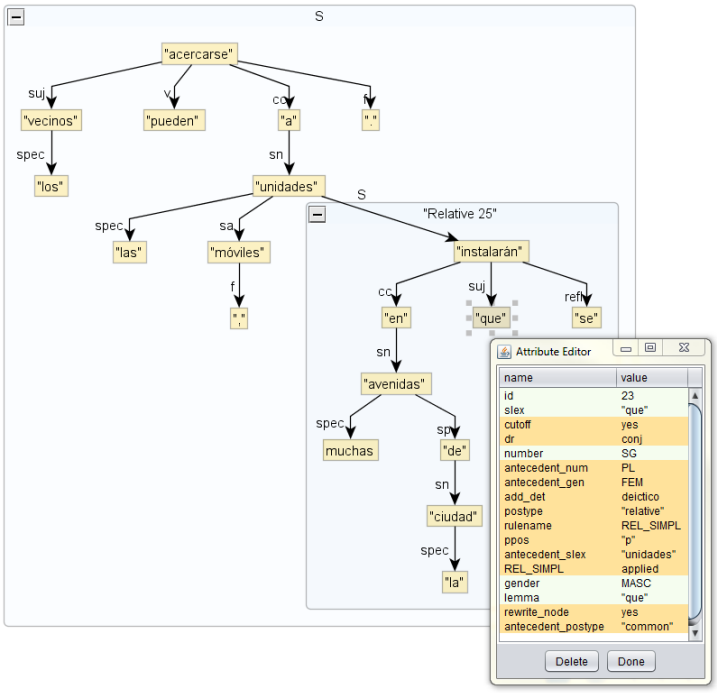

Fig. 5: A sample output of the second syntactic simplification grammar.

fact that this node needs a deictic determiner, etc.[3] The third grammar takes care of performing the modifications in the tree; the output produced by this grammar when applied to the structure in Figure 5 is shown in Figure 6. At this point, dependencies

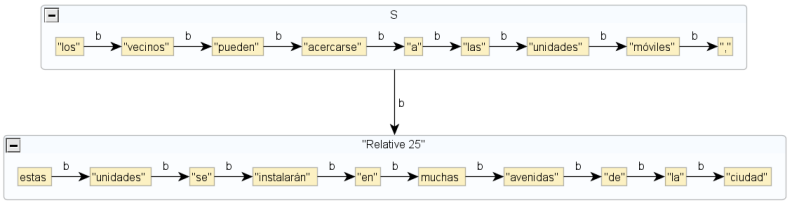

Fig. 6: A sample output of the syntactic simplification module

are not needed anymore, so they are removed; only the order between the nodes and chunks is explicit. The grammar successfully cuts the sentence into two parts, and substitutes the relative pronoun by its antecedent with a deictic such as *estas unidades* (these units) at the beginning of the second one.

Figure 7 shows a rule from the third syntactic simplification grammar, as it appears in the MATE development environment. This rule is applied when splitting a sentence that has coordinated main verbs with the same subject. In such cases, the subject is

[3]In combination with the rules, we use dictionaries which contain language-specific information such as the form of determiners and auxiliaries; that is, with one single rule we can insert any type of determiner by getting the adequate form in the dictionary.

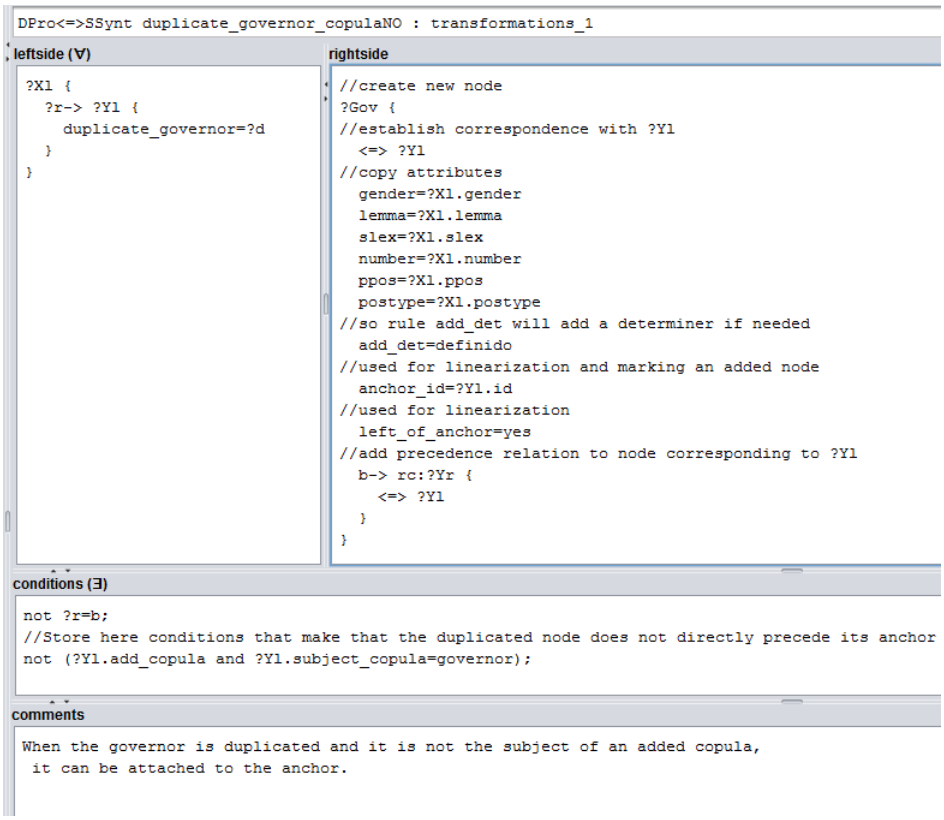

Fig. 7: A sample graph transduction rule

generally elided on the second verb and any following verbs, and it has to be introduced in order to complete each sentence, which is precisely what this rule does. The graph transduction rules map a part on the input graph with the *leftside*;[4] in this case, the rule looks for two nodes *?Xl* and *?Yl* linked by a relation *?r*, *?Yl* having the attribute *duplicate_governor* with any value (the interrogation marks indicate the variables). On the *rightside*, we indicate what to do; in this rule, a new node *?Gov* is created with all its attributes, and a relation *b* is added to another node *?Yr* that has already been built by another rule (*rc:* means 'right context' - as opposed to the context identified by the *leftside* of the rule). The rule has to comply with the other conditions stated in the 'conditions' field.

Finally, the **cleaning grammars** simply take care of removing some superfluous nodes, resolve ordering contradictions and add missing punctuation.

Some simplification operations implemented here are similar for other languages (e.g., the case of relative clauses or of main coordinated verbs). Our rules are designed in such a way that the grammars are generic allowing for easy adaptation to other languages. In order to run these grammars on English texts for instance, the *leftside*s should be adapted to the outputs of the parsers, but the *rightsides* would be almost

[4]Note that *leftside* and *rightside* are technical terms here, which are used to describe a rule in MATE. These terms also appear in the MATE IDE shown in Figure 7. *Leftside* describes the part of the rule that tries to match a part of the input graph. The *rightside* describes the output of the rule, which is a new sub-graph which is built if certain conditions are met.

exactly the same, given that we can use dictionaries to store language-specific features. Note that this kind of graph transduction grammars have already been used for paraphrasing legal texts, as part of a broader natural language generation pipeline. Our system could equally be used as a module of a language system such as the one described in [Mille and Wanner 2008], since our input structures are the same as one of their intermediate representations. However, to adapt it to other domains would require more work, because different types of simplifications would be at stake.

The grammars presented here have been previously evaluated in terms of precision and recall, looking at correct rule applications [Bott et al. 2012b] and all the possible target constructions where the rules should have been applied. The evaluation was done for separate grammatical construction types. The precision was calculated as the ratio between correct applications and all applications of each rule, while recall was defined as the ratio between correct applications and the target constructions which should have been simplified. For the most frequent target constructions, relative clauses and gerund constructions, we obtained a precision of 0.39, 0.63, respectively, while the values for recall were 0.66 and 0.21. As for coordination constructions, we could obtain precision and recall of 0.42/0.58 for object coordination and 0.65/0.50 for VP and clausal coordination. Some refinement of the grammars was carried out after the evaluation, based on error analysis. These refinements fixed some recurrent precision leaks and corrected grammaticality issues.

The evaluation we present in Section 5 below is complementary to the earlier evaluation because it takes simplicity metrics into account, as well as human judgments on the degree of simplicity. Also, it takes into account the interaction between this module and the other two, as we evaluate only those sentences which were altered by at least two modules at the same time.

4.2. Lexical Simplification

In our initial corpus study [Bott and Saggion 2011] we found that lexical changes were the single most frequent simplification operation produced by human editors. Therefore, we implemented a lexical simplification system called LexSiS [Bott et al. 2012a]. A second corpus study [Drndarevic and Saggion 2012] revealed a series of operations which were very frequent but could not be replaced by synonym substitution. To cover these, we implemented a rule-based lexical simplification system. Both components are described below.

4.2.1. Synonym Substitution in LexSiS. Synonym based lexical simplification has to solve two problems: first, the system must find a set of synonyms, which can serve as adequate word substitutes in a given context and, second, the system has to choose the synonym with the lowest lexical complexity. As a resource in which to look up synonyms we use the Spanish OpenThesaurus, a collaborative effort to create a thesaurus which is freely available under a GNU License. Even if the collaborative nature of the OpenThesaurus implies a certain lack of quality control, we showed in [Saggion et al. 2013] that its use does not lead to a significantly worse performance than the use of the Spanish WordNet, a resource with a much stricter quality control.

In order to address the problem of adequacy for a given context, LexSiS performs a kind of Word Sense Disambiguation: OpenThesaurus lists several groups of possible synonyms which separate different senses of the target word. Consider the example of the entry for *droga* (drug), which is ambiguous between a medical and a recreational reading:

droga|3
- medicina|medicamento|fármaco
- anfeta|anfetamina|estimulante

- estupefaciente|narcótico

OpenThesaurus lists three word senses here. One corresponds to the medical use and includes the words *medicamento* (medicament) and *medicina* (medicine). The other two correspond to the non-medial sense(s) and list words like *estimulante* (stimulant) and *narcótico* (narcotic). According to the distributional hypothesis [Firth 1957; Harris 1968], which we adopt here, different uses of a word like *droga* tend to occur in different lexical surroundings. LexSiS uses a Word Vector Space model [Sahlgren 2006], which represents the lexical distributional information for individual words. In this model, each dimension represents a possible context word and the extension of this dimension represents the frequency with which it can be found within a symmetric 9 word window (4 words to the left and 4 words to the right) in a corpus. The vector for individual senses of this word are derived by summing over the word vectors of all the words listed in one word sense. Each word-sense vector is then compared to a vector representing the local target context and the vector with the lowest cosine distance is chosen. Nevertheless, since the thesaurus is not perfect, often the word senses are not properly distinguished and often words are listed in the wrong word senses. Therefore, we apply an additional threshold which discards those words whose vectors are too distant from the target context.

Finally, the system must choose one of the words from the list of words that represent a word sense. For measuring the lexical simplicity we used a weighted measure which combines word frequency and word length.[5] The word with the highest simplicity score is chosen for the final lexical simplification. If none of the words listed in the chosen word sense has a higher simplicity score than the original word, no simplification is performed.

As an illustration, LexSiS is able to perform lexical substitutions, like the one which can be observed in (3), where the less frequently used word with Latin origin *urbes* (cities, urban areas) has been changed to *ciudades* (cities). In fact, even if the lemma *ciudad* is two characters longer than *urbe*, the former is nearly 80 times more frequent than the latter. Note that, in addition, this example contains an instance of syntactic simplification and one instance of rule based lexical simplification (*dos*/two transformed into the numeral *2*, see Section 4.2.2).

(3) a. *dos ciudades llegan a la fase final de un concurso convocado por la Comisión Europea para reconocer a aquellas* **urbes** *que más se han destacado en promover la accesibilidad universal.* (two cities reach the final phase of the competition organized by the European Commission to recognize those **urban areas** that have most stood out in their promotion of universal access.)

(3) b. **2** *ciudades llegan a la fase final de un concurso para reconocer a aquellas* **ciudades** *que más se han destacado en promover la accesibilidad universal. El concurso está convocado por la Comisión Europea.*

[5]The simplicity score is computed as $score_{simp} = \alpha_1 \times score_{word\_length} + \alpha_2 \times score_{freq}$, where $\alpha_1 = -0.39$ and and $\alpha_2 = 1.11$. Even if the scores for word length and frequency correspond to different underlying distributions and can thus not be directly compared, word frequency receives a stronger weight than word length. The frequency based simplicity score $score_{freq}$ is computed as the logarithm of the word count per lemma. The score for word length, $score_{word\_length}$, is calculated as $\sqrt{word\_length - 4}$ for words with more than 4 characters. Since nearly all highly frequent words tend to have between 1 and 4 letters, we assign a $score_{word\_length}$ of 0 to all words of length 4 and shorter. For a justification of this formula and the weight setting, please see [Bott et al. 2012a].

(2 cities reach the final phase of the competition to recognize those **cities** that have most stood out in their promotion of universal access. The competition is organized by the European Commission.)

A previous dedicated evaluation [Bott et al. 2012a] showed that the synonym substitution performed by LexSiS preserves the meaning in 72.49% of the cases and can produce a simpler replacement in 40.88% of the cases. In 58.93% of the instances the substituted word was judged as either equally complex or simpler than the original.

The system also outperformed the frequency baseline (where the most frequent word listed in the thesaurus is used as a substitute) employed in the SemEval simplification task in 2012 [Specia et al. 2012]; in the mentioned task the baseline turned out to be very hard to beat. This can be attributed to the fact that the Word Sense Disambiguation module improves over the baseline in terms of meaning preservation: the baseline, which ignores sense distinctions, only shows a meaning preservation of 66.12%, which compares favorably to the aforementioned 72.49% achieved by LexSiS.

It should be stressed that LexSiS only requires language resources which are relatively easy to find for most languages. The most crucial of these is a thesaurus lexical resource. We have shown in [Saggion et al. 2013] that also WordNet can be used for this purpose with similar results. Further on, LexSiS needs a lemmatizer and a sufficiently large corpus to train the Vector Space Model. With these relatively modest requirements, the module is portable to a range of other languages, which might not have large collections of parallel text, such as the Simple English Wikipedia which was used in some alternative approaches to lexical simplification [Yatskar et al. 2010; Biran et al. 2011].

4.2.2. Rule Based Lexical Simplification. An analysis of a subset of 40 pairs of original and manually simplified texts from the Simplext corpus revealed a number of restricted simplification operations none of which belong to either the lexical or syntactic components and which serve to normalize reporting verbs, reduce sentence content, and clarify, normalize or reduce numerical information.

All the insights obtained in this study have been reported in [Drndarevic and Saggion 2012] and [Drndarevic et al. 2013] and are here briefly summarized.

One of the findings of our analysis was that parenthetical information is generally eliminated from the sentences, so we implemented a rule that recognizes and eliminates the corresponding constructions from the texts.

Another interesting observation was the one concerning how reporting verbs, which are very common in newspaper articles, are simplified. We observed that ten different reporting verbs in the 40 original texts of the Simplext corpus (i.e., *warn*, *confirm*, *assure*, *suggest*, *say*, *explain*, *inform*, *point out*, *underline*) were all transformed at least once into the verb *decir* (*say*), which is simpler and less ambiguous than any of the other verbs. However, a rule that replaces a reporting verb with the form *say* cannot be blindly applied; instead, a set of rules which check the context of the reporting verb were implemented to ensure that the substitution is valid (e.g., the substitution with the verb *say* leaves the syntactic structure correct). The original list of reporting verbs from the Simplext corpus was expanded using a thesaurus so we obtained 32 different verbs in order to have a good coverage on unseen documents (see [Drndarevic et al. 2012] for details). The decision of substituting all reporting verbs was justified with the fact that *decir* is both the most common and the most general reporting verb [Quirk et al. 1985; Bosque Muñoz and Demonte Barreto 1999] and shorter than any of its semantic equivalents, which complies with the rules present in the "Make it Simple" guidelines [Freyhoff et al. 1998]. The authors also found that substitution of any reporting verb with *decir* eliminates polysemy, as is the case with the verb *indicar*, which in Spanish means both 'point' (the literal meaning) and 'point out' (non-literal

meaning). As stated in WCAG 2.0 guidelines [W3C 2008], use of non-literal meaning should be avoided in easy-to-read writing.

Where numerical expressions are concerned, we have found a number of interesting editing operations, the most common being that numerical expressions are usually eliminated, probably because they convey too detailed information which could be erased without harming the essential message. However we found this transformation a bit risky to implement because if blindly applied, it can harm the sentence syntactic structure.

The rest of the editing operations which are regular enough, context independent and therefore safe to implement are: the transformation of numbers in words into their equivalent numerical expressions (for numbers in the range from 1 to 10), the addition of the word "year" to the numerical representation of years (e.g., "year 1999" instead of "1999"), the transformation of named periods (e.g., decades, centuries) into their corresponding meaning (e.g., "20 years" instead of "two decades"), and the reduction of dates comprising a year to the year itself (e.g., "by 2010" instead of "by the end of May 2010"). This latter operation requires accurate identification of a number of complex constructions for which 47 rules have been implemented and tested.

The last rather regular operation, which is mainly observed in international news, is the transformation of adjectives of nationality into a periphrastic structure (e.g., "the government of Pakistan" for "the Pakistani government"); this is also observed for pronominalization of these adjectives (e.g., "people from Pakistan" for "the Pakistanis").

For the implementation of the aforementioned operations, the Java Annotation Pattern Engine from the GATE system was used [Cunningham et al. 2000]. The rules, which were manually designed, rely on lexical, part-of-speech tags, and dictionary information (e.g., reporting verbs, adjectives of nationality, keywords).

An evaluation of the rules over a set of randomly selected unseen documents from the corpus revealed perfect precision although limited recall. For example, rules that transform reporting verbs achieved perfect precision and 0.74 recall, while rules that transform numerical information achieved perfect precision and 0.84 recall.

5. EVALUATION DESIGN

In this section we present the experimental setup to evaluate our system by using, first, automatic readability measures and, second, a human evaluation.

We were interested in investigating how far the improvement could be measured in terms of automatic metrics of text complexity at different linguistic levels and also in how far such automatic metrics correlate to the judgments of human readers. For the evaluation with the use of metrics, we first compared the original texts to the automatically simplified versions of the same texts in order to see if an improvement could be observed. We also compared the original texts to the human simplified versions of these, which we took as an upper bound of how much reduction in measurable complexity a system could be expected to achieve. It should be noted that automatic metrics can usually only be applied at the text level and not at the sentence level, in contrast to the evaluation with human subjects, which was designed to apply at the sentence level. As for the evaluation with human subjects, we only compared original sentences to simplifications produced by the Simplext system (human simplifications were not assessed). We wanted to test three factors: in how far automatically simplified sentences showed a degradation of grammaticality, in how far they were perceived as being simpler, and in how far they preserved the meaning of the original.

For the creation of the dataset we used the Simplext Corpus for Spanish [Saggion et al. 2011]. This corpus is composed of news from four different genres: international news (INT), culture news (CULT), national (NAC) news and society (SOC). To obtain

the evaluation samples, we applied the Simplext system to the whole corpus. While for the evaluation with metrics we used the whole corpus, for the evaluation with human subjects we used randomly selected sentences which contained at least two automatically produced simplification operations. The corpus we used for the evaluation here did not contain the simplified text we used for the initial corpus studies presented in Section 3, nor material which was used for system development.

5.1. Readability Measures

We automatically evaluated our text simplification system using seven complexity measures for Spanish: the Lexical Complexity index (LC), the Spaulding's Spanish Readability index (SSR), the Sentence Complexity index (SCI), the Percentage of Complex Sentences (CS), the Average Sentence Length (ASL), the average embedding depth of sentence (DEPTH), and the average number of punctuation marks (PUNCT). We define each of these measures below.

The readability indexes we used (LC, SSR, and SCI) were not originally formulated for the evaluation of automatic simplification but for the assessment of the reading difficulty level of human produced texts. In spite of this, they showed a good correlation with many linguistically motivated features which might be seen as reading obstacles for our target population [Štajner et al. 2014]. In the same study, the authors proposed various ways in which those indices can be used in automatic evaluation of text simplification systems in Spanish.

To the best of our knowledge, the average embedding depth of sentence (DEPTH) has never been used for the automatic evaluation of the ATS systems before. We propose it here as we believe that it complements well the other three metrics concerned with the syntactic complexity of texts (SCI, CS, and ASL). The percentage of complex sentences (CS) was implemented as the ratio between complex and simple sentences which was used as a measure of syntactic complexity in [Štajner et al. 2012]. The average number of punctuation marks (PUNCT) was originally used in one of our previous studies [Drndarevic et al. 2013].

We initially assumed that the simplified texts produced by humans would achieve higher scores on these metrics because they were intended to be simpler to read than the original. We also expected that the automatically simplified texts would score better than the original texts, since the automatic process should resemble in some respect the human performance. It is important to note, that none of the metrics were used to guide the system development, so the influence of automatic simplification on improvement according to these scores should be mediated by the fact that automatic simplification imitates human operations.

Lexical Complexity index (LC). The Lexical Complexity index (LC) is defined as a measure of lexical complexity of literary texts aimed at second language learners. Following [Anula 2007], the formula is computed using equation (1) where LDI (Lexical Distribution Index) and ILFW (Index of Low Frequency Words) are computed with equations 2 and 3 respectively.

$$LC = \frac{LDI + IFW}{2} \tag{1}$$

$$LDI = \frac{N(dcw)}{N(s)} \tag{2}$$

$$ILFW = \frac{N(lfw)}{N(cw)} * 100 \quad (3)$$

Definition of variables used in the formulae are given in Table II. According to [Anula 2007], low frequency words (lfw) are those words whose frequency rank in the Referential Corpus of Contemporary Spanish (cf. footnote 1) is lower than 1,000 (See Table III for a sample of such a list). In order to compute the formula automatically, we lemmatised the list of low frequency words.

Table II: Basic Definitions for Complexity Measure Computation.

$N(dcw)$ is the number of distinct content words in the text.
$N(s)$ is the number of sentences in the text.
$N(cs)$ is the number of complex sentences in the text.
$N(lfw)$ is the number of low frequency words in the text.
$N(cw)$ is the number of content words in the text.
$N(w)$ is the number of words in the text.
$N(rw)$ is the number of rare words in the text.

Table III: Royal Spanish Academy's Frequency List from Royal Spanish Academy Corpus.

Word	**Frequency**
de	9999518
la	6277560
que	4681839
el	4569652
...	...
teatro	21663
importantes	21597
evitar	21587
...	...
adornos	957
discute	957
ejecutado	957
ermita	957
...	...
esnifaban	1
esnifadas	1
esnifaron	1

Spaulding's Spanish Readability index (SSR). The Spaulding's Spanish Readability index (SSR) [Spaulding 1956] uses both vocabulary and sentence structure to predict the relative difficulty of reading material. We use formula (4) and definitions on Table II for SSR computation.

$$SSR = 1.609 * \frac{N(w)}{N(s)} + 331.8 * \frac{N(rw)}{N(w)} + 22.0 \quad (4)$$

As rare words (rw), we considered those words which cannot be found on the list of 1,500 most common Spanish words provided in [Spaulding 1956]. Similarly to the case of the LC, we lemmatised the given list in order to be able to compute the formula

automatically. For the same reason, we slightly modified the formula by not taking into consideration Spaulding's additional rules for the SSR calculation. SSR has been used for assessing the reading difficulty of fundamental education materials for Latin American adults of limited reading ability. Therefore, it would be reasonable to expect that it could be successfully used for estimating the level of simplification performed by text simplification systems that aim at making texts more accessible for this target population.

Sentence Complexity Index (SCI). The Sentence Complexity Index (SCI) was proposed by Anula [2007] as a measure of sentence complexity in a literary text aimed at second language learners. For the computational implementation of SCI see equation (5).

$$SCI = \frac{ASL + CS}{2} \tag{5}$$

Average Sentence Length (ASL). The Average Sentence Length (ASL) was calculated according to equation 6.

$$ASL = \frac{N(w)}{N(s)} \tag{6}$$

Percentage of Complex Sentences (CS). The Percentage of Complex Sentences (CS) was calculated according to equation 7. We defined "complex sentences" as those which have more than one verb cluster (a cluster being a sequence of adjacent verbs without intervening words of other categories, such as *ha comido* (has eaten) or *quiere comer* (wants to eat)).

$$CS = \frac{N(cs)}{N(s)} \tag{7}$$

Embedding depth (DEPTH). For the calculation of embedding depth (DEPTH) we took the most deeply embedded node in the dependency tree for each sentence produced by the dependency parser [Bohnet 2009] and measured the distance between this node and the root of the tree as the number of intervening nodes (plus the leaf node itself). This measure does not discriminate between different syntactic constructions which may present different degrees of perceived complexity, but it is still a very useful metric to capture syntactic complexity: long sentences may be either syntactically complex or contain a lot of modifying material (adjectives, adverbs or adverbial phrases). The latter do not increase the syntactic complexity and do not result in very deep trees while the former have a strong tendency to result in deep trees. Because of this, syntactic embedding depth is a measure that complements ASL and captures syntactic complexity in terms of recursive structures.

Punctuation marks (PUNCT). "Make it Simple" European Guidelines for the Production of Easy-to-Read Information for people with Leaning Disability [Freyhoff et al. 1998] advise that texts aimed at this target population should have simple punctuation. Therefore, we calculate the average number of punctuation marks per text (PUNCT), according to the POS tagged output produced by FreeLing [Padró et al. 2010], as one of the indicators of text simplicity. Those punctuation marks which denote the beginning[6] and end of sentence were not taken into account. In this way, the

[6]Note that in Spanish, interrogative, imperative or exclamatory sentences have special punctuation marks not only at the end of the sentence but also at the beginning (¿and ¡).

presence (or absence) of sentences which were entirely deleted from the original texts or those sentences which were added in the simplified versions (as explanations of difficult terms or addition of general knowledge) during the manual simplification did not influence the results. This decision enabled a fairer comparison of the outputs of automatic and manual simplification, given that those two modules (for *deleting* and *adding* information) are not implemented in the current ATS system.

We applied the formulae to the original, manually simplified, and automatic simplifications of 120 texts from the Simplext corpus (those not used for the corpus studies), in order to test whether the formulae are good indicators of the degree of simplification and also to assess the degree of simplification achieved by our system. Results of the evaluation are presented in Section 6.1.

5.2. Evaluation with Expert Readers

For the human evaluation with non-target readers we created a dataset composed of sentences to be assessed according to their simplicity, grammaticality, and meaning preservation. Each participant had to read and rate the sentences as explained below.

5.2.1. Design. There were two conditions in the experiment: one experimental condition and one control condition. The experimental condition, "Simplified", is the condition in which the sentences were simplified using Simplext, while the control condition, "Original", is the condition in which the sentences were not modified. The experiments followed a within-subjects design, so every participant contributed to both of the conditions. The order of the conditions was counter-balanced to cancel out sequence effects. To measure the quality of our algorithm we used three dependent variables: *Simplicity Score*, *Grammaticality Score*, and *Meaning Preservation Score*.

5.2.2. Participants. We recruited a total of 25 participants. They were all native speakers of Spanish and their ages ranged from 18 to 58 years. We consider them strong readers because they all finished post-compulsory schooling and did not have any reading or cognitive disability. We decided to use strong readers because this way we ensure that a disability does not affect the results of the evaluation. None of the participants were involved in the project and none had experience in simplification tasks.

5.2.3. Materials. To study whether the sentences generated by our system were accurate and simpler we presented sentences to the participants. We used sentences and not shorter segments because the comprehension of the text generally pertains to long segments [Huenerfauth et al. 2009]. Following, we describe how we designed the materials that were used in this study.

Evaluation Dataset. For the creation of the dataset we used 120 sentences from the Simplext Corpus (not used for system development). From the system simplified output we extracted all those sentences which had undergone two or more simplification operations, stemming from at least two different simplification modules described in Section 3. In the human evaluation of ATS systems, it is a common practice to include only those sentences which have undergone at least one modification in each of the systems compared [Feblowitz and Kauchak 2013; Siddharthan and Angrosh 2014]. As the components of our system were already evaluated separately in [Bott et al. 2012a] and [Drndarevic et al. 2013], our goal here was to investigate how they interact among themselves and evaluate our ATS system as a whole. Therefore, we were interested only in those sentences which were simultaneously modified by at least two different simplification modules. This gave us a total of 150 automatically simplified sentences. We divided these sentences according to the genre to which they belong and then we randomly extracted 12 sentences from each genre.

As a result we had an evaluation dataset composed of 96 sentences, 48 simplified and 48 corresponding original sentences. For example, in the following pair of sentences we observe two changes done by Simplext: first, one long sentence is divided into two shorter ones and, second, the word in parenthesis "(Colombia)" is deleted.

(4) a. (Original) *La Casa de América de Madrid acoge el Festival Vivamérica, que este año se celebra también simultáneamente en las ciudades de Cádiz, Zaragoza y Barranquilla (Colombia).* (La Casa de America in Madrid hosts the Vivamérica Festival, which is this year also celebrated simultaneously in the cities of Cadiz, Zaragoza and Barranquilla (Colombia).)

(4) b. (Simplified) *La Casa de América de Madrid acoge el Festival Vivamérica. Este Festival este año se celebra también simultáneamente en las ciudades de Cádiz, Zaragoza y Barranquilla.* (La Casa de America in Madrid hosts the Festival Vivamérica. The Festival is this year also celebrated simultaneously in the cities of Cadiz, Zaragoza and Barranquilla.)

Test. To present the sentences we used an on-line questionnaire composed of 240 items: 96 items for rating the *Simplicity Score*, 96 for the *Grammaticality Score*, and 48 for the *Meaning Preservation Score*. For the *Meaning Preservation Score* the sentences were presented in pairs and the participants gave a score through comparison (i.e., These two sentences have the same meaning). For the *Simplicity Score* (i.e., this is a simple sentence) and *Grammaticality Score* (i.e., this sentence is grammatically correct) the sentences were presented individually. Each of the sentences were presented with a five-point Likert scale (1-Strongly disagree, 2-Disagree, 3-Neutral, 4-Agree, 5-Strongly agree).

5.2.4. Procedure. Depending on the participant the test lasted from 30 to 50 minutes. All participants undertook the test on-line at their homes. The fifth author was on-line to ease possible doubts or questions. First, the participants read the instructions presented in the test and had the opportunity to ask questions if needed. Then, they began with a questionnaire that was designed to collect demographic information. Third, they undertook the test and rated the sentences.

6. RESULTS

6.1. Readability Measures Results

Table IV shows the overall results for the automatic evaluation of the readability of the original, automatically simplified and manually simplified texts. Furthermore, we tested the statistical significance of the paired differences between each of the two corpora, for each of the readability metrics. In cases in which the metric was approximately normally distributed in both corpora, we used the 2-tailed paired t-test; we used Wilcoxon signed-rank test otherwise. The normality of the data was tested using the Shapiro-Wilk test of normality, which is preferred for small datasets (< 2000). All tests were performed using SPSS. Table V presents the mean value of the paired relative differences (MPRD) for each pair of the corpora on each metric. The MPRD were calculated according to Eq. 8, where $r_i(x)$ and $c_i(x)$ represent the value of the metric x on the ith *reference* text ($r_i(x)$), and the value of the metric x on the ith current text ($c_i(x)$). For example, in the column 'Original vs. Manual' in Table V, the original texts are the *reference* texts and the manually simplified texts are the *current* texts. In our case, the number of text pairs (N in Eq. 8) is always 120.

$$MPRD = \frac{1}{N} * \sum_{i=1}^{N} (\frac{100 * c_i(x)}{r_i(x)} - 100) \quad (8)$$

Table IV: Mean values for different readability metrics. Values are presented along with the standard error mean. Results are for 120 datapoints in each corpus.

Readability	Original	Autom. Simp.	Man. Simp.
LC	11.27 ± 0.26	9.35 ± 0.25	4.29 ± 0.27
SSR	179.89 ± 1.50	164.70 ± 1.50	120.90 ± 1.74
ASL	33.08 ± 0.56	25.43 ± 0.53	13.81 ± 0.16
CS	69.15 ± 1.39	55.11 ± 1.82	52.05 ± 2.04
SCI	51.11 ± 0.82	40.27 ± 1.08	32.93 ± 1.05
DEPTH	9.85 ± 0.14	8.50 ± 1.14	5.87 ± 0.06
PUNCT	17.22 ± 0.72	14.07 ± 0.59	3.40 ± 0.33

Table V: Mean paired relative differences (MPDR) on the whole corpora. Results are presented together with standard error mean. The differences presented in **bold** are those *not* statistically significant ($p > 0.05$), while all other results are *significant* ($p < 0.001$).

Readability	Original vs. Manual	Original vs. Autom. Simp.	Autom. Simp. vs. Man. Simp.
LC	-62.92% $\pm$ 1.90%	-17.00% $\pm$ 1.08%	-54.64% $\pm$ 2.23%
SSR	-32.74% $\pm$ 0.89%	-8.39% $\pm$ 0.46%	-25.90% $\pm$ 1.02%
ASL	-56.92% $\pm$ 0.85%	-22.32% $\pm$ 1.31%	-43.40% $\pm$ 1.15%
CS	-24.58% $\pm$ 3.02%	-20.55% $\pm$ 1.96%	**-1.33% $\pm$ 4.58%**
SCI	-34.43% $\pm$ 2.31%	-21.16% $\pm$ 1.63%	-14.52% $\pm$ 3.15%
DEPTH	-39.46% $\pm$ 0.77%	-13.12% $\pm$ 1.15%	-29.42% $\pm$ 1.03%
PUNCT	-77.28% $\pm$ 2.37%	-17.37% $\pm$ 1.41%	-72.28% $\pm$ 2.79%

The differences between the original texts and their corresponding simplified versions are very large (up to 77%) and statistically significant (at a 0.001 level of significance) on all seven readability metrics (column "Original vs. Manual", Table V). This shows that these seven metrics reflect some of the main differences between the original and simplified texts, and thus justifies their use as a part of the evaluation of our automatic simplification system. The scores for automatically simplified texts are consistently lower than those for the original texts on all seven measures (column "Original vs. Autom. Simp.", Table V). This indicates that our system produces texts which are simpler than the originals. However, the simplicity of these automatically simplified texts still does not reach the level of manually simplified texts on six out of seven used metrics (column "Autom. Simp. vs. Man. Simp.", Table V). This was expected, as all of those six metrics are heavily influenced by deletion of both whole sentences and sentence parts, which were common operations during the manual simplification [Štajner et al. 2013a], but are still not implemented in the current version of the system.

Still, automatic simplification achieves almost equal decrease in the percentage of complex sentences (CS) as the manual simplification (note that the paired relative difference in CS between those two corpora is very small and not statistically significant).

6.2. Evaluation with Expert Readers

A Shapiro-Wilk test showed that the datasets were not normally distributed ($p < 0.001$ for all datasets). Hence, to study the effect of the experimental condition on the Simplicity Score and Grammaticality Score we used the non-parametric test for repeated measures, the Wilcoxon's test. The results of the non-final user evaluation are presented in Table VI.

Table VI: Results of the non-final user evaluation (Simplicity (O): Simplicity of the original sentences; Simplicity (AS): Simplicity of the automatically simplified sentences; Gramm. (O): Grammaticality of the original sentences; Gramm. (AS): Grammaticality of the automatically simplified sentences; Meaning: Meaning preservation score on the corresponding pairs of original and automatically simplified sentences).

Score	Simplicity (O)	Simplicity (AS)	Gramm. (O)	Gramm. (AS)	Meaning
5 – Strongly agree	20%	30%	63%	24%	45%
4 – Agree	20%	20%	23%	21%	25%
3 – Neutral	30%	10%	7%	24%	10%
2 – Disagree	20%	20%	5%	18%	11%
1 – Strongly disagree	10%	20%	2%	13%	9%
Mean	3.20	3.20	4.40	3.25	3.86
Median	3	3.5	5	3	4
Mode	3	5	5	3 and 5	5
Positive	40%	50%	86%	45%	70%
Neutral	30%	10%	7%	24%	10%
Negative	30%	40%	7%	31%	20%

Although the original sentences and their corresponding automatic simplifications have the same mean for the Simplicity score (Mean = 3.2), the Wilcoxon's test for repeated measures reported a statistically significant difference between those two groups of sentences ($W = 68782.5$, $p < 0.001$*). Automatically simplified sentences were perceived as significantly simpler (Median = 3.5, Mode = 5) than their originals (Median = 3, Mode = 3). It is interesting to note that as many as 30% original sentences were rated as Neutral, while that was the case in only 10% of the automatically simplified sentences.

We found a significant difference between the conditions regarding the Grammaticality Score ($W = 301565$, $p < 0.001$*). Original sentences were perceived as significantly more grammatical than automatically simplified ones, which is common in previously proposed ATS systems [Wubben et al. 2012; Feblowitz and Kauchak 2013].

Regarding the Meaning Preservation Score, 70% of the participants agreed that the meaning was preserved (44.75% of them strongly agreed).

6.2.1. Comparison of Our Results with the State-of-the-art ATS Systems in English. Given that our experimental setup for the human evaluation follows the previously established standards for this task [Wubben et al. 2012; Feblowitz and Kauchak 2013; Angrosh and Siddharthan 2014][7], we are able to compare our results with the ones obtained for the state-of-the-art ATS in English (Table VII)[8]. Those four previous studies also evaluate the Simplicity, Fluency (which we call "Grammaticality"), and Adequacy (which we call "Meaning Preservation") on a five-point Likert scale. The mean value of the three scores is based on the ratings for: 62 original sentences and their corresponding simplifications involving 28 raters [Angrosh and Siddharthan 2014]; 100 original

[7][Narayan and Gardent 2014] use the 0–5 scale and [Glavaš and Štajner 2013] use the 1–3 scale, instead of the standard 1–5 scale and were thus excluded from this comparison.

[8]ATS systems in other languages, e.g. [Paetzold and Specia 2013; Specia 2010; Brouwers et al. 2014], were not evaluated using this kind of human evaluation.

sentences and their corresponding simplifications involving 10 raters [Feblowitz and Kauchak 2013]; 20 original sentences and their corresponding simplifications involving 46 raters [Wubben et al. 2012]; 48 original sentences and their corresponding simplifications involving 25 raters (current study).

Table VII: Comparison of the human evaluation scores obtained for our system and for the state-of-the-art ATS systems in English.

Reference	System	Fluency	Adequacy	Simplicity
[Wubben et al. 2012]	[Zhu et al. 2010]	2.59	2.82	2.93
	RevILP [Woodsend and Lapata 2011a]	3.18	3.28	2.96
	[Wubben et al. 2012]	3.83	3.71	2.88
[Feblowitz and Kauchak 2013]	[Feblowitz and Kauchak 2013]	3.80	3.09	3.55
	[Wubben et al. 2012]	3.64	3.91	3.07
	[Coster and Kauchak 2011a]	3.74	3.86	3.19
[Angrosh and Siddharthan 2014]	[Angrosh and Siddharthan 2014]	3.52	3.40	3.73
	[Woodsend and Lapata 2011a]	1.97	2.23	2.33
Current study	**Simplext**	**3.25**	**3.86**	**3.20**

As we can see from the results presented in Table VII, regarding its fluency, our system's output was rated lower than output of the systems proposed in [Wubben et al. 2012], [Feblowitz and Kauchak 2013], [Coster and Kauchak 2011a], and [Angrosh and Siddharthan 2014], but still better than those proposed in [Zhu et al. 2010] and [Woodsend and Lapata 2011a]. In terms of meaning preservation, our system was (together with the systems proposed in [Coster and Kauchak 2011a] and [Wubben et al. 2012]) rated the best. The simplicity of our automatically simplified sentences was rated equally good as in the system proposed in [Coster and Kauchak 2011a]. Only two systems ([Feblowitz and Kauchak 2013] and [Angrosh and Siddharthan 2014]) were rated better than ours in terms of the simplicity.

6.3. Evaluation with Target Readers

The PRODIS foundation[9] carried out a reading and comprehension evaluation of the texts produced by the Simplext system relying on 44 subjects with Down Syndrome (results of the experiment were reported as part of internal documentation of the Simplext project in [Rodriguez and Izuzquiza 2013]). This evaluation is complementary to the one presented above and sought to assess differences in readability and understanding of different versions of the same text with the intended users of the system. Only three texts – A (a text about pets), B (a text about a soccer museum), and C (a text about the Braile writing system) – were considered in the evaluation. Readability was measured as the time taken by the subjects to read the texts and comprehension was measured as the number of correct questions answered after reading the text. There were three conditions in the experiment: (i) "Original" – the condition in which the subject reads and answers questions about the original text, (ii) "Automatic" – the condition in which the subject reads and answers questions about the automatically simplified text, and (iii) "Manual" – the condition in which the subject reads and answers questions about the manually simplified text (see Table VIII for examples of original, manually simplified, and automatically simplified sentences for text A on pets adoption).

Each subject answered four inferential questions after reading a printed version of the text. Note that no participant read two different versions of the same text (see

[9]http://www.fundacionprodis.org/v2/en

Table VIII: Example of Original, Manually Simplified, and Automatically Simplified Sentences.

Original: *El 90% de los españoles prefiere adoptar un perro o gato antes que comprar el animal, en tanto que el 10 por ciento restante optaría por pagar porque prefiere un animal de raza, según un sondeo elaborado por eBay Anuncios sobre la percepción ante las adopciones de mascotas.* (90% of Spaniards prefer to adopt a dog or cat before buying the animal, while the remaining 10 percent would opt to pay because they prefer an animal of good breed, according to a survey prepared by eBay Ads on perceptions before pet adoption.)
Manual Simplification: *La mayoría de los españoles prefiere adoptar un perro o un gato a comprarlo. El resto de los españoles prefiere comprar el animal. Así están seguros de que el animal es de buena raza.* (Most Spaniards prefer to adopt a dog or cat to buy it. The rest of the Spaniards prefer to buy the animal. So they are sure that the animal is of good breed.)
Automatic Simplification: *El 90% de los españoles prefiere adoptar un perro o gato antes que comprar el animal, en tanto que el 10 por ciento restante optaría por pagar porque prefiere un animal de raza, según un sondeo sobre la percepción ante las adopciones de mascotas. El sondeo está elaborado por eBay Anuncios.* (90% of Spaniards prefer to adopt a dog or cat before buying the animal, while the remaining 10 percent would opt to pay because they prefer an animal of good breed, according to a survey on perceptions before pet adoption. The survey is prepared by eBay Ads.)

Table IX: Examples of Inferential Questions.

Text	Question
A - Pets	What is this news about?
B - Soccer Museum	What museum is this news about?
C - Braile Writing System	What is the Braile Writing System?

Table IX for examples of the questions for each of the three texts). While the quantitative results of the experiment showed differences between the original and simplified conditions with simplified texts obtaining on average more correct answers than the original texts, no statistical differences on readability and understanding for the three conditions could be established using statistical tests. However, the qualitative evaluation showed that the participants found very positively the existence of a tool such as Simplext, that is having a simplification solution accessible through different technological channels (e.g., computer, smart-phone, tablet). Subjects were also able to perceive differences in the texts making them more confident in the reading task. One of the limitations of the user evaluation was the size of the sample; with only 3 texts it was very difficult to establish any statistical differences in the obtained results so additional tests with more texts should be carried out before any conclusion can be reached.

7. DISCUSSION AND ERROR ANALYSIS

The evaluation according to automatic measures shows that the linguistic complexity captured by these measures can be reduced by our text simplification approach. Although statistically significant differences are observed when comparing original and automatically simplified versions (e.g., automatically simplified texts are indeed simpler), we cannot achieve the level of simplicity of human editors. This is not an unexpected result since human editors use considerable syntactic and world knowledge; they are able to transform the input by paraphrasing and applying summarization operations which are difficult to implement computationally.

According to our human evaluation, 70% of the simplifications preserve the meaning of the original sentences. In only 30% of the cases the simplification was not considered more simple than the original and in only 31% of the cases the automatically simplified sentence was not considered grammatical. The results should be interpreted with caution since all of the participants in the evaluation in Section 5.2 were native speakers

of Spanish with no intellectual disability. In addition, they had a generally high level of education. Therefore, the test subjects generally may have had no difficulties understanding the original sentences and may be more likely to perceived degraded grammaticality more strongly than any possible simplification effect. Simplified sentences tend to contain repetitions of words, which are desirable from a simplification point of view, but may be seen as stylistic shortcomings by readers that do not have problems understanding the original text from the start. Bad rule applications of the syntactic simplification module are especially disturbing for human readers, since they directly influence grammaticality. Earlier error analysis [Bott and Saggion 2014] revealed that many of the degraded output cases from the syntactic simplification module were directly traceable to parsing errors which stem from the dependency parser.

In order to estimate in how far the known problems of different parts of the pipeline affect perceived quality of the simplification - the degradation of the ratings for grammaticality, simplicity and meaning preservation between original and simplified version - we carried out an error analysis over the user ratings from the evaluation with non-final users. We found that practically none of the errors which caused distortion were caused by the rule based lexical simplification module. This module makes rather prudent and conservative changes, so it is not very surprising to find that it operates with a high precision. Therefore, the further error analysis only takes the syntactic simplification grammar and the LexSiS module into consideration.

Low simplicity ratings of automatic simplifications could be traced to both syntactic simplification errors and, to a somewhat larger extend, lexical errors produced by LexSiS. Interestingly, we could observe that syntactic simplification errors were not penalized as much as lexical simplification errors in this respect. The lexical errors were partly due to bad lemmatizations, which, for example wrongly output a verbal lemma for a target noun.

Turning to the decrease in grammaticality ratings, we found that, although this problem can be caused by either lexical or syntactic simplification, there seems to be a tendency for syntactic simplification errors to be responsible for it. On the contrary, a decrease in meaning preservation seems to have been caused more often by problems stemming from the LexSiS module, although syntactic errors are also very common. As we suspected, decrease of meaning preservation is also correlated with high statistical significance to both decrease of simplicity and decrease of grammaticality.[10].

Example (5) shows the worst simplification in the evaluation dataset (the one which got the worst simplicity, grammaticality and meaning preservation scores for the automatically simplified version). Here, two unfortunate simplification errors conspire: the word *escape* (escape) was substituted by the word *libertad* (liberty), which is listed as a possible synonym for the former, but is clearly infelicitous in the given context. Further on, there is also a syntactic simplification error which is caused by the wrong attachment point of the verb *causa* (cause) and a wrong interpretation of the adjective *abieto* as a past participle. Because of the parser error the grammar wrongly separates the complex NP subject of *causa* (causes) from the verb itself. Further on, because of the wrong attachment point (given by the parser) for the verb *causa*, the state of affairs which is caused is separated from the real grammatical subject of the verb (the use of stoves and fires) and placed in the wrong one of the two sentences which result from the splitting operation.

[10]We compared the ranks of the examples according to the each rating with Spearman's rank order correlation and found high levels of significance in both cases ($p<0.005$). Also decrease in simplicity and in grammaticality are highly correlated (ρ=0.825, df: 48, $p<0.005$)

(5) a. *La organización mundial advirtió que el uso de esos materiales para cocinar y calentar las casas en estufas o fuegos abiertos sin escape por las chimeneas causa contaminación en los espacios cerrados.*
(The international organization warned that the use of these materials for cooking and heating the houses in stoves or open fires without escape through chimneys causes contamination in closed spaces.)

(5) b. *La organización mundial dijo que el uso de esos materiales para preparar y calentar las casas en estufas o fuegos. Las estufas están abiertos sin libertad por las chimeneas causa contaminación en los espacios cerrados.*
(The international organization said that the use of the these materials for cooking and heating the houses. The stoves are open without liberty for the chimneys cause contamination in closed spaces.)

It must also be stressed that the syntactic constructions we target in this module are notoriously difficult for automatic parsing: it can be, for example, very hard for a parser to find the right head nouns it has to attach relative clauses to, which is a problem of similar complexity to the well-known PP-attachment problem.

The automatic synonym substitutions present different problems for human judges: they usually do not degrade the grammaticality of the output, but they influence the meaning directly. If our system chooses a non-synonym, the meaning of the sentence can be altered. To some degree, we could trace non-synonym substitution to errors in the lexical resource we used (OpenThesaurus). This resource often lists words which are not real synonyms or does not separate word senses accurately enough. A further source of errors is the word sense disambiguation (WSD) performed. WSD, especially in unrestricted settings like ours, is a non-trivial problem in its own right.

We also analysed the strengths of the pipeline, looking at the examples which got very high ratings in any of the aspects. The examples with the highest increase of simplicity ratings show both lexical and syntactic simplification. Interestingly, raters often perceived an increase in simplicity caused by synonym substitution, even if they did not find the meaning preservation to be perfect. One such example is shown in (6). The term *órganos competentes* (competent authorities), which appears in the original sentence quoted here, is changed in the automatic simplification to *órganos eficaces* (effective authorities). In this case the ratings for meaning preservation range from 1 to 5, with a mean of 3.72. The example was on the average rated as 0.72 point less complex than the original.

(6) . . . *establece la posibilidad de que los órganos competentes puedan fijar normas de calidad ambiental para los sedimentos . . .*
(. . . establishes the possibility that the competent authorities can fix norms for the environmental quality of the sediments . . .)

As already said, simplicity, grammaticality and meaning preservation are strongly correlated. This tendency is especially strong in the highest rated examples. So, the examples with the highest simplicity ratings (for the automatically simplified version) also tend to be the highest rated examples for grammaticality and meaning preservation.

Note that the results need to be taken with caution because the context of the simplified sentences has not been taken into account when evaluating the simplification. The results of the experiments leave space for improvement that we put forward in the next section.

Where the final user evaluation is concerned, although differences were observed in the performance of subjects with original and simplified texts (with simplified texts allowing users to correctly answer more questions on average), the size of the sample (3 texts) prevented us from finding any statistically significant differences in performance. A bigger sample would be required when carrying out further experiments. Text simplification is a problem which is well deserved to be studied, because it corresponds to a real need. Because of the multiple error sources and the unforgiving nature of the task, the current state-of-the-art cannot provide the end user with highly reliable fully automated text simplification.

However, we believe that some of the shortcomings can be improved with further investigation. Another promising option is to integrate automatic simplification in an editing environment directed to human editors, in a similar way in which machine translation can help human translators to produce high quality translations more efficiently.

8. CONCLUSIONS

The availability of massive textual repositories and the ubiquitous presence of textual material on the Web does not mean that we all fulfil our rights to access information. The way in which information is reported can have a big impact on accessibility for people with special needs, people with limited education, immigrants, etc. It is humanly impossible to create customized versions of texts for every single individual or group. Natural language processing research can, however, provide fully automatic or semi-automatic text simplification processors which can facilitate the task of transforming texts into adapted versions that are easier to read and understand for specific user groups.

In this paper we have described the research that was implemented in the Simplext project to provide automatic simplification in Spanish. The main contributions of our work are:

— the first lexical and syntactic simplification system for Spanish;
— a complete evaluation (both automatic and reader-based) of the solution; and
— a comparison with similar techniques for the English language.

Our work was empirical, based on the analysis of a parallel corpus of original and manually simplified newspaper articles. The findings from our corpus analysis were implemented whenever possible in different rules and procedures. Our technological roadmap was influenced by the need for a usable and accessible technological solution.

The limitations of the technology we used indicate many avenues for improvement. For example, during syntactic simplification we found that many errors were propagated into the simplification system due to parsing errors. We believe that adapting the parser to the specific characteristics of the target texts is essential and a method based on adding small amounts of genre specific data could improve the parser performance and therefore the simplification output.

Where the lexical simplification is concerned we have noticed that, on the one hand, many difficult words are not simplified due to the lack of coverage of the lexical resource and, on the other hand, that sometimes the replacement of a word by a synonym harms the simplification because of a bad word sense disambiguation. The first problem could be addressed by implementing a module of word sense induction providing appropriate synonyms for unknown words. The second problem could be addressed by further refining the filters that block those words which do not appear to fit in the context. We already integrated one such filter in our system, but more refined techniques from distributional semantics, such as a more appropriate term-weighting or the use of larger context windows, may lead to further improvements.

Acknowledgements

We are grateful to several anonymous reviewers for their very constructive comments and insights which helped us improve the final version of the paper. We acknowledge support from the Simplext project (Avanza Competitiveness TSI-020302-2010-84). We are grateful to the fellowship RYC-2009-04291 from Programa Ramón y Cajal 2009, to the project SKATER-UPF-TALN (TIN2012-38584-C06-03), Ministerio de Economía y Competitividad, Secretaría de Estado de Investigación, Desarrollo e Innovación, Spain, and to the project ABLE-TO-INCLUDE (CIP-ICT-PSP-2013-7/621055).

REFERENCES

Sandra Aluísio and Caroline Gasperin. 2010. Fostering digital inclusion and accessibility: the PorSimples project for simplification of Portuguese texts. In *Young Investigators NAACL Workshop on Computational Approaches to Languages of the Americas (YIWCALA '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 46–53.

Sandra Aluísio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability Assessment for Text Simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications (IUNLPBEA '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1–9.

Sandra Aluísio, Lucia Specia, Thiago A.S. Pardo, Erick G. Maziero, and Renata P.M. Fortes. 2008. Towards Brazilian Portuguese Automatic Text Simplification Systems. In *Proceedings of the Eighth ACM Symposium on Document Engineering (DocEng '08)*. ACM, New York, NY, USA, 240–248.

Mandya Angrosh, Tadashi Nomoto, and Advaith Siddharthan. 2014. Lexico-syntactic text simplification and compression with typed dependencies. In *COLING'14*. 1996–2006.

Mandya Angrosh and Advaith Siddharthan. 2014. Text simplification using synchronous dependency grammars: Generalising automatically harvested rules. In *Proceedings of the 8th International Natural Language Generation Conference (INGL)*. 16–25.

Alberto Anula. 2007. 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.)*. 162–170.

Alberto Anula. 2011. *Pautas básicas de simplificación textual y Diseño del corpus SIMPLEXT*. Technical Report. Grupo DILES. Universidad Autónoma de Madrid.

María Jesús Aranzabe, Arantza. Díaz de Ilarraza, and Itziar Gonzalez-Dios. 2012. First Approach to Automatic Text Simplification in Basque. In *Natural Language Processing for Improving Textual Accessibility (NLP4ITA) Workshop Programme*. 1–8.

María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Itziar Gonzalez-Dios. 2013. Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque. *Procesamiento del Lenguaje Natural* 50 (2013), 61–68.

Gianni Barlacchi and Sara Tonelli. 2013. ERNESTA: A sentence simplification tool for childrens stories in italian. In *Computational Linguistics and Intelligent Text Processing*. 476–487.

Kathy Barthe, Claire Juaneda, Dominique Leseigneur, Jean-Claude Loquet, Claude Morin, Jean Escande, and Annick Vayrette. 1999. GIFAS Rationalized French: A Controlled Language for Aerospace Documentation in French. *Technical Communication* 46, 2 (1999), 220–229.

Regiba Barzilay and Lillian Lee. 2004. Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization. In *NAACL HLT'04*. 113–120.

Susana Bautista, Carlos León, Raquel Hervás, and Pablo Gervás. 2011. Empirical Identification of Text Simplification Strategies for Reading-Impaired People. In *European Conference for the Advancement of Assistive Technology*. Maastricht, the Netherlands, 567–574.

Or Biran, Samuel Brody, and Noémie Elhadad. 2011. Putting it Simply: a Context-Aware Approach to Lexical Simplificaion. In *ACL' HLT11*. Association for Computational Linguistics, Portland, Oregon, USA, 496–501.

Bernd Bohnet. 2009. Efficient Parsing of Syntactic and Semantic Dependency Structures. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task (CoNLL '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 67–72.

Bernd Bohnet, Andreas Langjahr, and Leo Wanner. 2000. A Development Environment for an MTT-based Sentence Generator. In *Proceedings of the First International Conference on Natural Language Generation - Volume 14 (INLG '00)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 260–263.

Ignacio Bosque Muñoz and Violeta Demonte Barreto. 1999. *Gramática Descriptiva de la Lengua Española*. Real Academia Española.

Stefan Bott, Luz Rello, Biljana Drndarevic, and Horacio Saggion. 2012a. Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish. In *COLING'12*. 357–374.

Stefan Bott and Horacio Saggion. 2011. Spanish Text Simplification: An Exploratory Study. *Procesamiento del Lenguaje Natural* 47 (2011), 87–95.

Stefan Bott and Horacio Saggion. 2014. Text simplification resources for Spanish. *Language Resources and Evaluation* 48, 1 (2014), 93–120.

Stefan Bott, Horacio Saggion, and Simon Mille. 2012b. Text Simplification Tools for Spanish. In *LREC'12*. 1665–1671.

Nadjet Bouayad-Agha, Gerard Casamayor, Gabriela Ferraro, and Leo Wanner. 2009. Simplification of Patent Claim Sentences for their Paraphrasing and Summarization. In *FLAIRS Conference*. 302–303.

Laetitia Brouwers, Delphine Bernhard, Anne-Laure Ligozat, and Thomas François. 2014. 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*. 47–56.

Jill Burstein, Jane Shore, John Sabatini, Yong-Won Lee, and Matthew Ventura. 2007. The Automated Text Adaptation Tool.. In *NAACL HLT'07 (Demonstrations)*. 3–4.

Yvonne Canning, John Tait, Jackie Archibald, and Ros Crawley. 2000. Cohesive Generation of Syntactically Simplified Newspaper Text. In *Proceedings of the Third International Workshop on Text, Speech and Dialogue (TDS '00)*. Springer-Verlag, London, UK, UK, 145–150.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In *Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*. 7–10.

Raman Chandrasekar. 1994. *A Hybrid Approach to Machine Translation using Man Machine Communication*. Ph.D. Dissertation. Tata Institute of Fundamental Research/University of Bombay, Bombay.

Raman Chandrasekar, D. Doran, and B. Srinivas. 1996. Motivations and Methods for Text Simplification. In *COLING'96*. 1041–1044.

Raman Chandrasekar and Bangalore Srinivas. 1997. Automatic induction of rules for text simplification. *Knowledge-Based Systems* 10, 3 (1997), 183–190.

James Clarke and Mirella Lapata. 2006. Models for Sentence Compression: A Comparison Across Domains, Training Requirements and Evaluation Measures. In *ACL'06*. Association for Computational Linguistics, Stroudsburg, PA, USA, 377–384.

Kevyn Collins-Thompson and Jamie Callan. 2005. Predicting Reading Difficulty with Statistical Language Models. *J. Am. Soc. Inf. Sci. Technol.* 56, 13 (Nov. 2005), 1448–1462.

William Coster and David Kauchak. 2011a. Learning to simplify sentences using Wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*. Association for Computational Linguistics, 1–9.

William Coster and David Kauchak. 2011b. Simple English Wikipedia: a new text simplification task. In *ACL HLT'11*. 665–669.

Scott A. Crossley and Danielle S. McNamara. 2008. Assessing L2 reading texts at the intermediate level: An approximate replication of Crossley, Louwerse, McCarthy & McNamara (2007). *Language Teaching* 41 (7 2008), 409–429. Issue 03.

Hamish Cunningham, Diana Maynard, and Valentin Tablan. 2000. *JAPE: a Java Annotation Patterns Engine (Second Edition)*. Research Memorandum CS–00–10. Department of Computer Science, University of Sheffield.

Jan De Belder, K. Deschacht, and Marie-Francine Moens. 2010. Lexical simplification. In *ITEC'10*. https://lirias.kuleuven.be/handle/123456789/268437

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Proceedings of the SIGIR workshop on accessible search systems*. 19–26.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. READ-IT: Assessing Readability of Italian Texts with a View to Text Simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies (SLPAT '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 73–83.

Siobhan Devlin. 1999. *Simplifying natural language text for aphasic readers*. Ph.D. Dissertation. University of Sunderland, UK.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In *Proceedings of the 2nd Conference on Human Language Technology Research, San Diego*. 138–145.

Biljana Drndarevic and Horacio Saggion. 2012. Reducing Text Complexity through Automatic Lexical Simplification: an Empirical Study for Spanish. *Procesamiento del Lenguaje Natural* 49 (2012), 13–20.

Biljana Drndarevic, Sanja Štajner, Stefan Bott, Susana Bautista, and Horacio Saggion. 2013. Automatic Text Simplification in Spanish: A Comparative Evaluation of Complementing Modules. In *CICLing (2)*. 488–500.

Biljana Drndarevic, Sanja Štajner, and Horacio Saggion. 2012. Reporting Simply: A Lexical Simplification Strategy for Enhancing Text Accessibility. In *Proceedings of the Easy-to-read on the Web Symposium*.

William H. DuBay. 2004. The principles of readability. *Impact Information* (2004), 1–76.

Inmaculada Fajardo, Vicenta vila, Antonio Ferrer, Gema Tavares, Marcos Gmez, and Ana Hernndez. 2014. Easy-to-read Texts for Students with Intellectual Disability: Linguistic Factors Affecting Comprehension. *Journal of Applied Research in Intellectual Disabilities* 27, 3 (2014), 212–225.

Dan Feblowitz and David Kauchak. 2013. Sentence Simplification as Tree Transduction. In *Proceedings of the 2n Workshop on Predicting and Improving Text Readability for Targe Reader Populations (PITR), Sofia, Bulgaria*. 1–10.

Lijun Feng. 2009. Automatic readability assessment for people with intellectual disabilities. In *SIGACCESS Access. Comput.* Number 93. ACM, New York, NY, USA, 84–91. DOI:http://dx.doi.org/10.1145/1531930.1531940

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively Motivated Features for Readability Assessment. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 229–237.

John R. Firth. 1957. *Papers in Linguistics 1934-51*. Longmans, London, UK.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology* 32, 3 (1948), 221–233.

Thomas François and Patrick Watrin. 2011. On the Contribution of MWE-based Features to a Readability Formula for French as a Foreign Language. In *RANLP*. 441–447.

Geert Freyhoff, Gerhard Hess, Linda Kerr, Bror Tronbacke, and Kathy Van Der Veken. 1998. *Make it Simple, European Guidelines for the Production of Easy-toRead Information for People with Learning Disability*. ILSMH European Association, Brussels.

Morton A. Gernsbacher and Mark E. Faust. 1991. The mechanism of suppression: A component of general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 17 (1991), 245–262.

Goran Glavaš and Sanja Štajner. 2013. Event-Centered Simplication of News Stories. In *Proceedings of the Student Workshop held in conjunction with RANLP 2013, Hissar, Bulgaria*. 71–78.

P. Gómez. 2011. Identificación de dificultades de comprensión lectora en el uso de Internet en jóvenes con discapacidad intelectual para el desarrollo de un periódico digital.. In *Poster presented at the XV Congreso Nacional y I Internacional de Modelos de Investigación Educativa, Madrid, 2011, September*. 1–13.

Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza, and H. Salaberri. 2014. Simple or Complex? Assessing the readability of Basque Texts. In *COLING'14*. 334–344.

Zellig Harris. 1968. Distributional Structure. In *The Philosophy of Linguistics*, Jerold J. Katz (Ed.). Oxford University Press, 26–47.

Matt Huenerfauth, Lijun Feng, and Noémie Elhadad. 2009. Comparing evaluation techniques for text readability software for adults with intellectual disabilities. In *ASSETS '09*. ACM, 3–10.

Kentaro Inui, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, and Tomoya Iwakura. 2003. Text simplification for reading assistance: A project note. In *Proceedings of the second international workshop on Paraphrasing - Volume 16 (PARAPHRASE '03)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 9–16.

Joyce Karreman, Thea van der Geest, and Esmee Buursink. 2007. Accessible Website Content Guidelines for Users with Intellectual Disabilities. *Journal of Applied Research in Intellectual Disabilities* 20 (2007), 510–518.

J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel.* Technical Report.

Beata Beigman Klebanov, Kevin Knight, and Daniel Marcu. 2004. Text Simplification for Information-Seeking Applications. In *On the Move to Meaningful Internet Systems, Lecture Notes in Computer Science*. Springer Verlag, 735–747.

Partha Lal and Stefan Rüger. 2002. Extract-based Summarization with Simplification. In *Proceedings of the ACL 2002 Automatic Summarization / DUC 2002 Workshop*.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. Association for Computational Linguistics, Barcelona, Spain, 74–81.

Adam Lopez. 2008. Statistical Machine Translation. *ACM Comput. Surv.* 40, 3, Article 8 (Aug. 2008), 49 pages.

J. Martos, S. Freire, A. González, D. Gil, and M. Sebastian. 2012. *D2.1: Functional requirements specifications and user preference survey*. Technical Report. FIRST technical report.

Aurélien Max. 2006. Writing for Language-Impaired Readers. In *Computational Linguistics and Intelligent Text Processing: 7th International Conference, CICLing 2006, Mexico City, Mexico, February 19-25, 2006, Proceedings*, Alexander Gelbukh (Ed.). 567–570.

Harry G. McLaughlin. 1969. SMOG grading - a new readability formula. *Journal of Reading* (May 1969), 639–646.

Mencap. 2002. *Am I making myself clear? Mencaps guidelines for accessible writing*.

Simon Mille and Leo Wanner. 2008. Making Text Resources Accessible to the Reader: the Case of Patent Claims. In *LREC'08*. 1393–1400.

Michelle F. Morgan and Karen B. Moni. 2008. Meeting the challenge of limited literacy resources for adolescents and adults with intellectual disabilities. *British Journal of Special Education* 35, 2 (2008), 92–101.

Shashi Narayan and Claire Gardent. 2014. Hybrid Simplification using Deep Semantics and Machine Translation. In *ACL'14*. 435–445.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating Content Selection in Summarization: The Pyramid Method. In *NAACL-HLT'04*. 145–152.

Misako Nomura, Gyda Skat Nielsen, and Bror Tronbacke. 1997. *Guidelines for Easy-to-Read Materials*. Technical Report. IFLA, Library Services to People with Special Needs Section. http://www.ifla.org/files/assets/hq/publications/professional-report/120.pdf

Courtenay Frazier Norbury. 2005. Barking up the wrong tree? Lexical ambiguity resolution in children with language impairments and autistic spectrum disorders. *Journal of experimental child psychology* 90 (2005), 142–171.

Charles Kay Ogden. 1937. *Basic English: A General Introduction with Rules and Grammar*. Paul Treber, London.

Ethel Ong, Jerwin Damay, Gerard Lojico, Kimberly Lu, and Dex Tarantan. 2007. Simplifying Text in Medical Literature. *Journal of Research in Science, Computing and Engineering* 4, 1 (2007), 37–47.

Constantin Orasan, Richard Evans, and Iustin Dornescu. 2013. *Towards Multilingual Europe 2020: A Romanian Perspective*. Romanian Academy Publishing House, Bucharest, Chapter Text Simplification for People with Autistic Spectrum Disorders, 287–312.

Lluís Padró, Miquel Collado, Samuel Reese, Marina Lloberes, and Irene Castellón. 2010. FreeLing 2.1: Five Years of Open-source Language Processing Tools. In *LREC'10*. Valletta, Malta, 931–936.

Gustavo H. Paetzold and Lucia Specia. 2013. Text Simplification as Tree Transduction. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, Fortaleza, CE, Brazil*. 116–125.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL'02*. 311–318.

Sarah E. Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *Proc. of Workshop on Speech and Language Technology for Education*. 69–72.

Sarah E. Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer Speech & Language* 23, 1 (2009), 89–106.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman Inc. New York.

Luz Rello and Ricardo Baeza-Yates. 2014. Evaluation of DysWebxia: A Reading App Designed for People with Dyslexia. In *W4A'14*. ACM, Seoul, Korea, 10.

Luz Rello, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion. 2013a. Simplify or help?: text simplification strategies for people with dyslexia. In *W4A'13*. ACM, Rio de Janeiro, Brazil, 15.

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. 2013b. Frequent Words Improve Readability and Short Words Improve Understandability for People with Dyslexia. In *INTERACT'13*. Springer, Cape Town, South Africa, 229–245.

Luz Rello, Ricardo Baeza-Yates, Horacio Saggion, and Eduardo Graells. 2012. Graphical Schemes May Improve Readability but Not Understandability for People with Dyslexia. In *Proceedings of the NAACL HLT 2012 Workshop Predicting and improving text readability for target reader populations (PITR 2012)*. Montreal, Canada.

Luz Rello, Susana Bautista, Ricardo Baeza-Yates, Pablo Gervás, Raquel Hervás, and Horacio Saggion. 2013. One Half or 50%? An Eye-Tracking Study of Number Representation Readability. In *INTERACT'13*. Springer, Cape Town, South Africa, 203–219.

Olga Rodriguez and Lola Izuzquiza. 2013. *Informe iterativo de los resultados de nivel de lecto-comprensión*. Technical Report. PRODIS.

Mikael Roll, Johan Frid, and Merle Horne. 2007. Measuring Syntactic Complexity in Spontaneous Spoken Swedish. *Language and Speech* 50, 2 (2007), 227–245.

Marina B. Ruiter, Toni C. M. Rietveld, Cucchiarini Catina, Krahmer Emiel J., and Helmer Strik. 2010. Human Language Technology and communicative disabilities: Requirements and possibilities for the future. In *LREC'10*. 143–151.

Horacio Saggion, Stefan Bott, and Luz Rello. 2013. Comparing Resources for Spanish Lexical Simplification. In *SLSP*. 236–247.

Horacio Saggion, Elena Gómez-Martínez, Esteban Etayo, Alberto Anula, and Lorena Bourg. 2011. Text Simplification in Simplext. Making Text More Accessible. *Procesamiento del Lenguaje Natural* 47 (2011), 341–342.

Horacio Saggion and Thierry Poibeau. 2013. Automatic Text Summarization: Past, Present, and Future. In *Multi-source, Multilingual Information Extraction and Summarization*, T. Poibeau, H. Saggion, J. Piskorski, and R. Yangarber (Eds.). Springer.

Horacio Saggion, Juan-Manuel Torres-Moreno, Iria da Cunha, Eric SanJuan, and Patricia Velázquez-Morales. 2010. Multilingual Summarization Evaluation without Human Models. In *COLING'10*. 1059–1067.

Magnus Sahlgren. 2006. *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces*. Ph.D. Dissertation. Stockholm University.

Sarah E. Schwarm and Mari Ostendorf. 2005. Reading Level Assessment Using Support Vector Machines and Statistical Language Models. In *ACL'05*. 523–530.

Violeta Seretan. 2012. Acquisition of Syntactic Simplification Rules for French. In *LREC'12* (23-25), Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ug(ur Dog(an, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Istanbul, Turkey, 4019–4026.

Luo Si and Jamie Callan. 2001. A Statistical Model for Scientific Readability. In *Proceedings of the Tenth International Conference on Information and Knowledge Management (CIKM '01)*. ACM, New York, NY, USA, 574–576.

Advaith Siddharthan. 2002. An Architecture for a Text Simplification System. In *In LEC02: Proceedings of the Language Engineering Conference (LEC'02)*. 64–71.

Advaith Siddharthan. 2011. Text Simplification using Typed Dependencies: A Comparison of the Robustness of Different Generation Strategies. In *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG)*. 2–11.

Advaith Siddharthan and M.A. Angrosh. 2014. Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Gothenburg, Sweden,*. 722–731.

E. A. Smith and R. J. Senter. 1967. *Automated Readability Index*. Technical Report. Aerospace Medical Research Laboratories, Wright-Patterson Air Force Base, Ohio.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas*. 223–231.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation, Athens, Greece*. 259–268.

Seth Spaulding. 1956. A Spanish Readability Formula. *The Modern Language Journal* 40 (1956), 433–441.

Lucia Specia. 2010. Translating from Complex to Simplified Sentences. In *PROPOR*. 30–39.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. Semeval-2012 task 1: English lexical simplification. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*. 347–355.

David Vickrey and Daphne Koller. 2008. Sentence simplification for semantic role labeling. In *ACL HLT'08*. 344–352.

Tim Vor der Brück, Sven Hartrumpf, and Hermann Helbig. 2008. A readability checker with supervised learning using deep syntactic and semantic indicators. *Informatica* 32, 4 (2008), 429–435.

Sanja Štajner. 2014. Translating sentences from 'original' to 'simplified' Spanish. *Procesamiento del Lenguaje Natural* 53 (2014), 61–68.

Sanja Štajner, Biljana Drndarević, and Horacio Saggion. 2013a. Corpus-based Sentence Deletion and Split Decisions for Spanish Text Simplification. *Computacion y Systemas* 17, 2 (2013), 251–262.

Sanja Štajner, Biljana Drndarevic, and Horacio Saggion. 2013b. Eliminación de frases y decisiones de división basadas en corpus para simplificación de textos en español. *Computación y Sistemas* 17, 2 (2013), 251–262.

Sanja Štajner, Richard Evans, Constantin Orasan, and Ruslan Mitkov. 2012. What can readability measures really tell us about text complexity?. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*. Istanbul, Turkey, 1–8.

Sanja Štajner, Ruslan Mitkov, and Gloria Corpas Pastor. 2014. *Recent Advances in Language Production, Cognition and the Lexicon*. Springer, Chapter Simple or not simple? A readability question, 379–398.

Tu Thanh Vu, Giang Binh Tran, and Son Bao Pham. 2014. Learning to Simplify Children Stories with Limited Data. In *ACIIDS 2014*. Springer International Publishing Switzerland, 31–41.

W3C. 2008. *Web Content Accessibility Guidelines (WCAG) 2.0*. http://www.w3.org/TR/WCAG20/

Willian Massami Watanabe. 2010. Facilita: Reading Assistance to the Functionally Illiterate. In *W4A'10*. ACM, New York, NY, USA, Article 7, 2 pages.

Kristian Woodsend and Mirella Lapata. 2011a. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 409–420.

Kristian Woodsend and Mirella Lapata. 2011b. WikiSimple: Automatic simplification of wikipedia articles. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*. 927–932.

Sander Wubben, Antal van den Bosch, and Emil Krahmer. 2012. Sentence simplification by monolingual machine translation. In *ACL'12*. 1015–1024.

Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lee Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *NAACL'10*. 365–368.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of The 23rd International Conference on Computational Linguistics*. 1353–1361.