Can Spanish Be Simpler?
LexSiS: Lexical Simplification for Spanish

Stefan BOTT Luz RELLO Biljana DRNDAREVIC Horacio SAGGION
TALN / DTIC
Universitat Pompeu Fabra
Barcelona, Spain
{stefan.bott,luz.rello,biljana.drndarevic,horacio.saggion}@upf.edu

ABSTRACT
Lexical simplification is the task of replacing a word in a given context by an easier-to-understand synonym. Although a number of lexical simplification approaches have been developed in recent years, most of them have been applied to English, with recent work taking advantage of parallel monolingual datasets for training. Here we present LexSiS, a lexical simplification system for Spanish that does not require a parallel corpus, but instead relies on freely available resources, such as an on-line dictionary and the Web as a corpus. LexSiS uses three techniques for finding a suitable word substitute: a word vector model, word frequency, and word length. In experiments with human informants, we have verified that LexSiS performs better than a hard-to-beat baseline based on synonym frequency.

TITLE AND ABSTRACT IN SPANISH

¿Puede ser el Español más simple?
LexSiS: Simplificación Léxica en Español

La tarea de simplificación léxica consiste en sustituir una palabra en un contexto determinado por un sinónimo que sea más sencillo de comprender. Aunque en los últimos años han aparecido algunos sistemas para desempeñar esta tarea, la mayoría de ellos se han desarrollado para el inglés y hacen uso de corpus paralelos. En este artículo presentamos LexSiS, un sistema de simplificación léxica en español que utiliza recursos libremente disponibles tales como un diccionario en línea o la Web como corpus, sin la necesidad de acudir a la creación de corpus paralelos. LexSiS utiliza tres técnicas para encontrar un sustituto léxico más simple: un modelo vectorial basado en palabras, la frecuencia de las palabras y la longitud de la palabras. Una evaluación realizada con tres anotadores demuestra que para algunos conjuntos de datos LexSiS propone sinónimos más simples que el sinónimo más frecuente.

KEYWORDS: Lexical Simplification, Text Simplification, Textual Accessibility, Word Sense Disambiguation, Spanish.

KEYWORDS IN SPANISH: Simplificación Léxica, Simplificación Textual, Accesibilidad Textual, Desambiguación, Español.

Proceedings of COLING 2012: Technical Papers, pages 357–374,
COLING 2012, Mumbai, December 2012.

1 Introduction

Automatic text simplification is an NLP task that has received growing attention in recent years (Chandrasekar et al., 1996; Carroll et al., 1998; Siddharthan, 2002; Aluísio et al., 2008; Zhu et al., 2010). Text simplification is the process of transforming a text into an equivalent which is easier to read and to understand than the original, preserving, in essence, the original content. This process may include the manipulation of several linguistic layers, and consists of sub-tasks such as syntactic simplification, lexical simplification, content reduction and the introduction of clarifications and definitions. Historically, text simplification started as a task mainly intended as a preprocessing stage in order to make other NLP tasks easier (Chandrasekar et al., 1996; Siddharthan, 2002). However, the task of simplifying a text also has a high potential to help people with various types of reading comprehension problems (Carroll et al., 1998; Aluísio and Gasperin, 2010). For example, lexical simplification by itself, without syntactic simplification, can be helpful for users with some cognitive conditions, such as aphasic readers or people with dyslexia (Hyönä and Olson, 1995). This second context in which text simplification is carried out is closely related to social initiatives which promote easy-to-read material, such as the Simple English section of the Wikipedia.[1] There are also various national and international organizations dedicated to the (mostly human) production of simple and simplified text.

Lexical simplification, an indispensable component of a text simplification system, aims at the substitution of words by simpler synonyms, where the evident question is: "What is a simpler synonym?". The lion's share of the work on lexical simplification has been carried out for English. In this paper, we present LexSiS, the first system for the lexical simplification of Spanish text, which proposes and evaluates a solution to the previous question. LexSiS is being developed in the context of the Simplext project (Saggion et al., 2011), which aims at improving text accessibility for people with cognitive impairments. Until now text simplification in Spanish has concentrated mainly on syntactic simplification (Bott and Saggion, 2012). Lexical and syntactic simplification are tasks which are very different in nature. Working with Spanish presents particular challenges, most notably dealing with the lack of large-scale resources which could be used for our purposes.

LexSiS uses (i) a word vector model to find possible substitutes for a target word and (ii) a simplicity computation procedure grounded on a corpus study and implemented as a function of word length and word frequency. LexSiS uses available resources such as the free thesaurus OpenThesaurus and a corpus of Spanish documents from the Web. The approach we take here serves to test how well relatively simple open domain resources can be used for lexical simplification. Since comparable resources can be found for many other languages, our approach is, in principle, language independent. As will be shown in this paper, by using contextual information and a well-grounded simplicity criterion, LexSiS is able to outperform a hard-to-beat frequency-based lexical replacement procedure.

Next section discusses related work on text simplification with particular emphasis on lexical simplification. Section 3 presents the analysis of a sample of original and simplified texts to design a word simplicity criteria. In Section 4 we present the resources we use for the development of LexSiS, while in Section 5 we describe our lexical simplification approach. We present the evaluation design in Section 6 and discuss the obtained results in Section 7. Finally, in Section 8 we summarize our findings and indicate possible ways to improve our results.

[1] `http://simple.wikipedia.org/`

2 Related Work

Text simplification has by now become a well-established paradigm in NLP, combining a number of rather heterogeneous sub-tasks, such as syntactic simplification, content reduction, lexical simplification and the insertion of clarification material. In this paper, we are only interested in lexical simplification as one of the various aspects of text simplification. Lexical simplification requires, at least, two things: a way of finding synonyms (or, in some cases, hyperonyms), and a way of measuring lexical *complexity* (or simplicity, see Section 3). Note that applying word sense disambiguation can improve the accuracy of the simplification. Consider trying to simplify the word *hogar* in the following sentence: *La madera ardía en el hogar ('The wood was burning in the fireplace')*. The most frequent synonym of *hogar* is *casa ('house')*; however, choosing this word for simplification would produce the sentence *La madera ardía en la casa ('The wood was burning in the house')*, which does not preserve the meaning of the original sentence. Choosing the correct meaning of *hogar*, in this case *'fireplace'*, is important for lexical simplification.

Early approaches to lexical simplification (Carroll et al., 1998; Lal and Ruger, 2002; Burstein et al., 2007) often used WordNet in order to find appropriate word substitutions, in combination with word frequency as a measure of lexical simplicity. Bautista et al. (2011) use a dictionary of synonyms in combination with a simplicity criterion based on word length. De Belder et al. (2010) apply explicit word sense disambiguation, with a Latent Words Language Model, in order to tackle the problem that many of the target words to be substituted are polysemic.

More recently, the availability of the Simple English Wikipedia (SEW) (Coster and Kauchak, 2011b), in combination with the "ordinary" English Wikipedia (EW), made a new generation of text simplification approaches possible, which use primarily machine learning techniques (Zhu et al., 2010; Woodsend et al., 2010; Woodsend and Lapata, 2011b; Coster and Kauchak, 2011a; Wubben et al., 2012). This includes some new approaches to lexical simplification. Yatskar et al. (2010) use edit histories for the SEW and the combination of SEW and EW in order to create a set of lexical substitution rules. Biran et al. (2011) also use the SEW/EW combination (without the edit history of the SEW), in addition to the explicit sentence alignment between SEW and EW. They use WordNet as a filter for possible lexical substitution rules. Although they do not apply explicit word sense disambiguation, their approach is *context-aware*, since they use a cosine-measure of similarity between a lexical item and a given context, in order to filter out possibly harmful rule applications which would select word substitutes with the wrong word sense. Their work is also interesting because they use a Vector Space Model to capture the lexical semantics and, with that, their context preferences.

Finally, there is a recent tendency to use statistical machine translation techniques for text simplification (defined as a monolingual machine translation task). Coster and Kauchak (2011a) and Specia (2010), drawing on work by Caseli et al. (2009), use standard statistical machine translation machinery for text simplification. The former uses a dataset extracted from the SEW/EW combination, while the latter is noteworthy for two reasons: first, it is one of the few statistical approaches that targets a language different from English (namely Brazilian Portuguese); and second, it is able to achieve good results with a surprisingly small bi-data-set of only 4,483 sentences. Specia's work is closely related to the PorSimples project, described in Aluísio and Gasperin (2010). In this project a dedicated lexical simplification module was developed, and it uses a thesaurus and a lexical ontology for Portuguese. They use word frequency as a measure for simplicity, but apply no word sense disambiguation.

3 Corpus Analysis

As the basis for the development of LexSiS, we have conducted an empirical analysis of a small corpus of news articles in Spanish, the Simplext Corpus (Bott and Saggion, 2011). It consists of 200 news articles, 40 of which have been manually simplified. Original texts and their corresponding simplifications have been aligned at the sentence level, thus producing a parallel corpus of a total of 590 sentences (246 and 324 in the original and simplified sets respectively). All texts have been annotated using Freeling, including part-of-speech tagging, named entity recognition and parsing (Padró et al., 2010).

Our methodology, explained more in depth in Drndarevic and Saggion (2012), consists in observing lexical changes applied by trained human editors and preparing their computational implementation accordingly. In addition to that, we conduct quantitative analysis on the word level in order to compare frequency and length distributions in the sets of original and simplified texts. Earlier work on lexical substitution has largely concentrated on word frequency, with occasional interest for word length as well (Bautista et al., 2009). It has also been shown that lexical complexity correlates with word frequency: more frequent words present less cognitive effort for the reader (Rayner and Duffy, 1986). Our analysis is motivated by the desire to test the relevance of these factors in the text genre we treat and the possibility of their combined influence on the choice of the simplest out of a set of synonyms to replace a difficult input word.

We observe a high percentage of named entities (NE) and numerical expressions (NumExp) in our corpus, due to the fact that it is composed of news articles, which naturally abound in this kind of expressions. NEs and NumExps have been discarded from the frequency and length analysis because they are tagged as a whole by Freeling, and this presents us with two difficulties. First, some expressions, such as *30 millones de dólares ('30 million dolars')* or *Programa Conjunto de las Naciones Unidas sobre el VIH/sida ('Joint United Nations Programme on HIV/AIDS')*, are extremely long words (some exceed 40 characters in length) and are not found in the dictionary; thus, we cannot assign them a frequency index. Second, such expressions are not replaceable by synonyms, but require a different simplification approach.

We conduct word length and frequency analysis from two angles. First, we analyse the totality of the words in the parallel corpus. Second, we analyse all lexical units (including multi-word expressions, e.g. complex prepositions) that have been substituted with a simpler synonym. These pairs of lexical substitutions (O-S) have been included in the so-called Lexical Substitution Table (LST) and are used for evaluation purposes (see Section 6).

3.1 Word Length

Analysing the total of 10,507 words (6,595 and 3,912 in the original and simplified sets respectively), we have observed that the most prolific words in both sets are two character words, the majority of which are function words (97.61% in O and 88.97% in S). Two to seven-character words are more abundant in the S set, while longer words are slightly more common in the O set. The S set contains no words with more than 15 characters. Analysis of the pairs in the LST has given us similar results: almost 70% of simple words are shorter than their original counterparts.

On the whole, we can conclude that in S texts there is a tendency towards using shorter words of up to ten characters, with one to five-character words taking up 64.10% of the set and one to ten-character words accounting 95.54% of the content.

3.2 Word Frequency

To analyse the frequency, a dictionary based on the Referential Corpus of Contemporary Spanish (Corpus de Referencia del Español Actual, CREA)[2] has been compiled for the purposes of the Simplext project. Every word in the dictionary is assigned a frequency index (FI) from 1 to 6, where 1 represents the lowest frequency and 6 the highest. We use this resource for the corpus analysis because it allows easy categorisation of words according to their frequency and elegant presentation and interpretation of results. However, in Section 5 this method is abandoned and relative frequencies are calculated based on occurrences of given words in the training corpus, so as to ensure that words not found in the above mentioned dictionary are also covered.

In the parallel corpus, we have documented words with FI 3, 4, 5 and 6, as well as words not found in the dictionary. The latter are assigned FI 0 and termed *rare words*. This category consists of infrequent words such as *intransigencia* (*'intransigence'*), terms of foreign origin, like *e-book*, and a small number of multi-word expressions, such as *a lo largo de* (*'during'*). The latter are recognized as multi-word expressions by Freeling, but are not included in the dictionary as such. The ratio of these expressions with respect to total is rather small (1.08% in O and 0.59% in S), so it should not significantly influence the overall results, presented in Table 1.

Frequency index	Original	Simplified
Freq. 0	10,53%	4,71%
Freq. 3	1,36%	0,74%
Freq. 4	1,35%	1,00%
Freq. 5	6,68%	5,67%
Freq. 6	80,08%	87,88%

Table 1: The distribution of n-frequency words in original and simplified texts.

We observe that lower frequency words (FI 3 and FI 0) are around 50% more common in O texts than in S texts, while the latter are somewhat more saturated in highest frequency words. As a general conclusion we observe that simple texts (S set) make use of more frequent words from CREA than their original counterparts (O set).

In order to combine the factors of word length and frequency, we have additionally analysed the length of all the words in the category of *rare words*. We have found that rare words are largely (72.44% in O and 77.44% in S) made up of seven to nine-character words, followed by longer words of up to twenty characters in O texts (39.42%) and fourteen characters in S texts (29.88%).

We are, therefore, lead to believe that there is a degree of connection between the factors of word length and word frequency, and that these are to be combined when scores are assigned to synonym candidates. In Section 5.1 we propose criteria for determining word simplicity exploiting these findings.

4 Resources

As we already mentioned in Section 2, most attempts to resolve the problem of lexical simplification are concentrated on English and, in recent years, Simple English Wikipedia in combination with the "ordinary" English Wikipedia has become a valuable resource for the study of text

[2] `http://corpus.rae.es/creanet.html`

simplification in general, and lexical simplification in particular. For Spanish, like for most other languages, no comparably large parallel corpora are available.

Some approaches to lexical simplification make use of WordNet (Miller et al., 1990) in order to measure the semantic similarity between lexical items and to find an appropriate substitute. While Spanish is one of the languages represented in EuroWordNet (Vossen, 2004), its scope is much more modest. The Spanish part of EuroWordNet contains only 50,526 word meanings and 23,370 synsets, in comparison to 187,602 meanings and 187,602 synsets in the English WordNet 1.5.

4.1 Corpora

The most valuable resources for lexical simplification are comparable corpora which represent the "normal" and a simplified variant of the target language. Although the corpus described in Section 3 served us as a basis for the corpus study and provided us with gold standard examples for the evaluation presented in Section 6, it is not large enough to train a simplification model. We, therefore, made use of an 8M word corpus of Spanish text extracted from the Web to train the vector models in Section 4.3.

4.2 Thesaurus

We use the Spanish OpenThesaurus (version 2),[3] which is freely available under the GNU Lesser General Public License, for the use with OpenOffice.org. This thesaurus lists 21,831 target words (lemmas) and provides a list of word senses for each word. Each word sense is, in turn, a list of substitute words (and we shall refer to them as *substitution sets* hereafter). There is a total of 44,353 such word senses. The substitution candidate words may be contained in more than one of the substitution sets for a target word. The following is the Thesaurus entry for *mono*, which is ambiguous between the nouns *'ape'*, *'monkey'* and *'overall'*, as well as the adjective *'cute'*.

(a) mono|4
- |gorila|simio|antropoide
- |simio|chimpancé|mandril|mico|macaco
- |overol|traje de faena
- |llamativo|vistoso|atractivo|sugerente|provocativo|resultón|bonito

OpenThesaurus lists simple one-word and multi-word expressions, both as target and substitution units. In the current version of LexSiS, we only treat single-word units, but we plan to include the treatment of multi-word expressions in future versions. We counted 436 expressions of the kind, such as *arma blanca* (*'stabbing or cutting weapon'*) or *de esta forma* (*'in this manner'*). Some of those expressions are very frequent and are used as tag phrases. The treatment of multi-word expressions only requires a multi-word detection module as an additional resource.

4.3 Word Vector Model

In order to measure lexical similarity between words and contexts, we used a Word Vector Model (Salton et al., 1975). Word Vector Models are a good way of modelling lexical semantics

[3] `http://openthes-es.berlios.de`

(Turney and Pantel, 2010), since they are robust, conceptually simple and mathematically well defined. The 'meaning' of a word is represented as the contexts in which it can be found. A word vector can be extracted from contexts observed in a corpus, where the dimensions represent the words in the context, and the component values represent their frequencies. The context itself can be defined in different ways, such as an n-word window surrounding the target word. Whether two words are similar in meaning can be measured as the cosine distance between the two corresponding vectors. Moreover, vector models are sensitive to word senses. For example, vectors for word senses can be built as the sum of word vectors which share one meaning.

We trained this vector model on the 8M word corpus mentioned in 4.1. We lemmatized the corpus with FreeLing (Padró et al., 2010) and for each lemma type in the corpus we constructed a vector, which represents co-occurring lemmas in a 9-word (actually 9-lemma) window (4 lemmas to the left and to the right). The vector model has n dimensions, where n is the number of lemmata in the lexicon. The dimensions of each vector in the model (i.e. the vector corresponding to a target lemma) represent the lemmas found in the contexts, and the value for each component represents to number of times the corresponding lemma has been found in the 9-word context. In the same process, we also calculated the absolute and relative frequencies of all lemmas observed in this training corpus.

5 LexSiS Method

LexSiS tries to find the best substitution candidate (a word lemma) for every word which has an entry in the Spanish OpenThesaurus. The substitution operates in two steps: first the system tries to find the most appropriate substitution set for a given word, and then it tries to find the best substitution candidate within this set. Here the *best* candidate is defined as the simplest and most appropriate candidate word for the given context. As for the simplicity criterion, we apply a combination of word length and word frequency, and for the determination of appropriateness we perform a simple form of word sense disambiguation in combination with a filter that blocks words which do not seem to fit in the context.

In the first step, we check for each lemma if it has alternatives in OpenThesaurus. If this is the case, we extract a vector from the surrounding 9-word window, as described in Section 4.3. Since each word is a synonym to itself (and might actually be the simplest word among all alternatives), we include the original word lemma in the list of words that represent the word sense. We construct a common vector for each of the word senses listed in the thesaurus by adding all the vectors (resulting from Section 4.3) to the words listed in each word sense. Then, we select the word sense with the lowest cosine distance to the context vector. In the second step, we select the best candidate within the selected word sense, assigning a simplicity score and applying several thresholds in order to eliminate candidates which are either not much simpler or seem to differ too much from the context.

5.1 Simplicity

According to our discussion in Section 3, we calculate simplicity as a combination of word length and word frequency. The task of combining them, however, is not entirely trivial, considering the underlying distribution of lengths and frequencies. In both cases simplicity is clearly not linearly correlated to the observable values. We know that simplicity monotonically decreases with length and monotonically increases with frequency, but a linear combination of the two factors not necessarily behaves monotonically as well. What we need is a score for simplicity, such that for all possible combinations of word lengths and frequencies of two words, w_1 and

w_2, $score(w_1) > score(w_2)$ iff w_1 is simpler than w_2. For this reason, we try to approximate the correlation between simplicity and the observable values at least to some degree.

In the case of length, our corpus study showed that a word with length *wl* is simpler than a word with length $wl + 1$. But the degree to which it is simpler depends on the value of *wl*. The corresponding difference decreases with longer values for *wl*. For words with a very high *wl* value, a difference in simplicity between *wl* words and *wl* - 1 words is not perceived any more. In our corpus, we found that very long words (10 characters and longer) were always substituted with much shorter words with an average length difference of 4.35 characters. In medium length range (from 5 to 9 characters), the average difference was only 0.36 characters, and very short original words (4 characters or shorter) did not tend to be shortened in the simplified version at all. For this reason we use the following formula:[4]

$$score_{wl} = \begin{cases} \sqrt{wl - 4} & \text{if } wl \geq 5\,, \\ 0 & \text{otherwise.} \end{cases}$$

In the case of frequency, we make the standard assumption that word frequency is distributed according to Zipf's law (Zipf, 1935); therefore, simplicity must be similarly distributed (when we abstract away from the influence of word length). In order to get a score which associates simplicity to frequency in a way which comes closer to linearity, we calculate the simplicity score for frequency as the logarithm of the frequency count c_w for a given word:

$$score_{\text{freq}} = \log c_w$$

Now the combination of the two values is

$$score_{\text{simp}} = \alpha_1 score_{wl} + \alpha_2 score_{\text{freq}}$$

where α_1 and α_2 are weights. We determined values for α_1 and α_2 in the following way: we manually selected 100 good simplification candidates proposed by OpenThesaurus for given contexts. We only considered cases which were both indisputable synonyms and clearly perceived as being simpler than the original. Then we calculated the average difference between the scores for word length and word frequency between the the original lemma and the simplified lemma, and took these averaged differences as being the average contribution of length and frequency to the receivable *simplicity* of the lemma. This resulted in $\alpha_1 = -0.39$[5] and $\alpha_2 = 1.11$.

[4]The formula for $score_{wl}$ resulted in quite a stable average value for $score_{wl}(w_{\text{original}}) - score_{wl}(w_{\text{simplified}})$ for the different values of wl in the range of word lengths from 7 to 12, when tested on the gold standard (cf 6.3 below). For longer and shorter words this value was still over-proportionally high or low, respectively, but the difference is less pronounced than with alternative formulas we tried, and much smoother than the direct use of wl counts. In addition, 74% of all observed substitutions fell into that range.

[5]Note that word length is a penalizing factor, since longer words are generally less simple. For this reason, the value for α_1 is negative.

5.2 Thresholds

There are several cases in which we do not want to accept an alternative for a target word, even if it has a high simplicity score. First of all, we do not want to simplify frequent words, even if OpenThesaurus lists them. So we set a cutoff point for frequent words, such that LexSiS does not try to simplify words with a frequency higher than 1% (calculated on the training corpus in 4.3). We also discard substitutes where the difference in the simplicity score with respect to the original word is lower than 0.5, because such words can be expected not to be significantly simpler. We achieved this latter value through experimentation.

Many of the alternatives proposed by OpenThesaurus are in reality not acceptable substitutes. We try to filter out words that do not fit into the context by discarding all candidates whose word vector has a distance with a cosine inferior to 0.013, another value achieved through experimentation.

Finally, there are two cases in which the system does not propose a substitute. First, there are cases where none of the substitution candidates have low enough cosine distance to the context vector (with the threshold of 0.013). There are also cases where the highest scoring substitute is the same as the original lemma. In both cases the original word is preserved.

6 Evaluation

In this section we present the experimental set-up employed to evaluate LexSiS by comparing it with two baselines and a gold standard. The evaluation was conducted thoroughly, rating the degree of simplification and the preservation of meaning of the substitutions.

6.1 Baselines

We employ two baselines:

(a) Random: it replaces the target word with a synonym selected randomly from our resource.

(b) Frequency: it replaces a word with its most frequent synonym provided by the thesaurus, presumed to be the simplest, similar to Devlin and Unthank (2006).

6.2 Gold Standard

As the gold standard we used the synonym pairs composed by the manual lexical substitutions extracted from the corpus described in Section 3. Since current lexical simplification systems for English handle only one-word cases (Yatskar et al., 2010; Biran et al., 2011), only single lexical substitutions were taken intro consideration. For instance, the substitution of the word *pinacoteca (‘art gallery’)* by *museo (‘museum’)* in the following sentence:

(b) (O) El visitante puede contemplar los óleos y esculturas que se exponen en la PINACOTECA.
‘The visitor can appreciate the paintings and sculptures showed in the ART GALLERY.’
(S) El visitante puede contemplar los óleos y esculturas que se exponen en el MUSEO.
‘The visitor can appreciate the paintings and sculptures showed in the MUSEUM.’

With that restriction, we found a total of 26 lexical substitutions in the corpus. We discarded collocations and multi-word expressions, as well as single lexical substitutions which supposed a mayor transformation in the sentence structure. For instance, the word *pese a (‘despite’)*, substituted with *sin embargo (‘however’)*, has changed the structure of the sentence itself:

(c) (O) Amnistía subrayó que Manning se encuentra detenido en "custodia máxima", PESE A carecer de antecedentes por violencia o infracciones disciplinarias durante la detención, lo que significa que está atado de pies y manos durante todas las visitas y se le niega la oportunidad de trabajar y salir de su celda.
'Amnesty underlined that Manning is being held in "maximum custody", DESPITE having no history of violence or disciplinary offenses during detention, which means his hands and feet are tied during all visits and he is denied the opportunity to work and leave his cell.'
(S) Bradley Manning ha tenido un buen comportamiento en la cárcel. SIN EMBARGO, Bradley Manning está atado durante las visitas. Tampoco puede trabajar ni salir de su celda.
'Bradley Manning has exhibited good behavior in prison. HOWEVER, Bradley Manning is tied during visits. He cannot work or leave his cell.'

6.3 Evaluation Dataset

We have three different evaluation datasets.

(T-S/B) Target *vs.* **System and Baselines:** This dataset is composed of 50 unique target words, together with their synonyms generated by LexSiS and the baselines. To create this dataset, we first ran our system for lexical simplification through the original texts from the Spanish Simplext Corpus (Bott and Saggion, 2011). This gave us 739 automatic lexical substitutions. We randomly selected 50 sentences that included one lexical substitution. Subsequently, for each of the examples, we generated two baselines and inserted them in the sentences, obtaining a total of 200 sentences. We manually corrected the ungrammatical examples resulting from lexical substitution.

(T-G/S/B) Target *vs.* **Gold, System and Baselines:** This dataset is composed of the examples which were manually simplified in the gold standard and which were also simplified by our system (12 cases out of the 26 cases of single lexical substitution in the Simplext Corpus). For each of these examples, we also generated two baselines, obtaining a total or 48 sentences to evaluate.

(G-T) Gold *vs.* **Target:** This dataset is composed of the remaining lexical simplification examples of our gold standard (14 instances) and their target words. For these examples our system did not propose a simplified example.

We could set up more evaluation data to compare the system and the baselines. However, we considered 50 sentences (a total of 200 sentences for dataset T-S/B, and 276 different sentences if we consider all data sets) to be reasonable because: (i) It was feasible for the annotators to perform the task; and (ii) previous works in lexical simplification used slightly smaller data sets for evaluation - in Yatskar et al. (2010) they use a total of 200 simplification examples, and in Biran et al. (2011) 130 simplification examples are used. Below, we show two examples of a sentence with its original word (O) and the lexical substitution proposed by our system (S).

(d) (O) De la Iglesia merece este GALARDÓN.
'De la Iglesia deserves this AWARD.'
(S) De la Iglesia merece este PREMIO.
'De la Iglesia deserves this PRIZE.'

(e) (O) Cuenta con una AMPLIA oferta de títulos.
'It has a WIDE range of titles.'
(S) Cuenta con una GRAN oferta de títulos.
'It has a GREAT range of titles.'

6.4 Design

We created a multiple choice questionnaire presenting two sentences for each item. Each item contained one sentence with a simplification example and the same sentence with the target word. These sentences were presented in random order to the annotator (i.e., either as Target *vs.* LexSiS/Gold/Baseline or LexSiS/Gold/Baseline *vs.* Target). The labels for each pair of sentences in comparison with the other were: "simpler", "more complex", "equally simple or complex", "they do not have the same meaning", and "I do not know or I do not understand at least one of the sentences". In Table 2 we show an example of one target word and its corresponding substitutions presented in the sentences.

Type	Sentence
Target	Descubren en Valencia una nueva ESPECIE de pez prehistórico. *A new SPECIES of prehistoric fish is discovered in Valencia.*
LexSIS	Descubren en Valencia un nuevo TIPO de pez prehistórico. *A new TYPE of prehistoric fish is discovered in Valencia.*
Frequency Baseline	Descubren en Valencia un nuevo GRUPO de pez prehistórico. *A new GROUP of prehistoric fish is discovered in Valencia.*
Random Baseline	Descubren en Valencia un nuevo LINAJE de pez prehistórico. *A new LINEAGE of prehistoric fish is discovered in Valencia.*

Table 2: Example extracted from dataset (T-S/B) Target *vs.* LexSIS and Baselines.

6.5 Annotators

Three annotators with no previous annotation experience performed the tests using an on-line form. They were all Spanish native speakers, frequent readers and were not the authors of this paper. To measure the inter-annotator agreement of multiple annotators and multiple classes (five classes, one per label), we used the Fleiss' kappa measure (Fleiss, 1971). The three participants annotated all the instances of the three datasets, achieving a Fleiss' kappa score of 0.330. Hence, we can assume we have a fair agreement (Landis and Koch, 1977), comparable with other inter-annotator agreements in related literature (see Section 7).

7 Results and Discussion

In this section we present the results and discuss the performance of LexSiS. We calculate (in percentage) how well the meaning has been preserved, taking into consideration all the instances (Synonym column). The percentage of simpler, equal and more complex synonyms is calculated among the synonymous instances (Simpler Syn., Equal Syn. and Complex Syn. columns, respectively) and among all the instances, taking into consideration the global performance of LexSiS (Simpler Syn. Glob., Equal Syn. Glob. and Complex Syn. Glob. columns, respectively).[6]

System	Synonym (%)	Simpler Syn. (%)	Equal Syn. (%)	Complex Syn. (%)	Simpler Syn. Glob. (%)	Equal Syn. Glob. (%)	Complex Syn. Glob. (%)
Random	65.03	15.13	24.37	58.82	9.94	**16.02**	38.67
Frequency	66.12	**42.98**	19.01	**38.02**	28.42	12.57	**25.14**
LexSiS	**72.49**	40.88	**17.52**	41.61	**29.63**	12.70	30.16
Gold	97.44	71.05	17.11	11.84	69.23	16.67	11.54

Table 3: Evaluation of all the datasets.

[6]The sum of Simpler Syn. Glob., Equal Syn. Glob. and Complex Syn. Glob. columns equals the number of synonyms (Synonym column) and the sum of Simpler Syn., Equal Syn. and Complex Syn. columns equals 100.

In Table 3 we present the results for all the datasets. The percentage of meaning preservation is higher in LexSiS than in the baselines, and LexSiS offers simpler synonyms (29.63%) than the frequency baseline (28.42%). However, 30.16% of the lexical substitutions suggested by our system are found to be more complex by the annotators. One possible reason for this is that the texts used for this study were already simple from a lexical point of view. For instance, we found only 26 manual lexical substitutions in the Simplext corpus, and our system only offered a lexical substitute for 12 of these manual substitutions. Consequently, 53.84% of the manually substituted target words were not simplified by LexSiS, and 78.57% of these manual substitutions were found to be simpler synonyms by the annotators (see Table 4). Even though our gold standard is small, to the best of our knowledge, this is the only existing resource from which we could generate a gold standard for lexical simplification in Spanish.

System	Synonym (%)	Simpler Syn. (%)	Equal Syn. (%)	Complex Syn. (%)	Simpler Syn. Glob. (%)	Equal Syn. Glob. (%)	Complex Syn. Glob. (%)
Gold	100	78.57	16.67	4.76	78.57	16.67	4.76

Table 4: Evaluation of dataset (G-T): Gold *vs.* Target.

We performed an error analysis to try and discover why LexSiS did not account for the remaining instances from the gold standard, and we found five possible reasons: (1) the target word does not occur in the LexSiS dictionary; (2) the target word is already too frequent to be substituted by LexSiS; (3) the target word and its lexical substitution are the same; (4) the lexical substitution proposed by LexSiS has a very similar ranking to the target word, hence the substitution is not performed; and (5) the target word is discarded because its vectorial distance taking into account the original context is too large.

System	Synonym (%)	Simpler Syn. (%)	Equal Syn. (%)	Complex Syn. (%)	Simpler Syn. Glob. (%)	Equal Syn. Glob. (%)	Complex Syn. Glob. (%)
Random	**57.14**	4.76	42.86	47.62	2.86	25.71	28.57
Frequency	47.22	47.06	17.65	35.29	22.22	**8.33**	16.67
LexSiS	55.56	**65.00**	**15.00**	**20.00**	**36.11**	**8.33**	**11.11**
Gold	95.83	58.70	17.39	21.74	61.11	16.67	16.67

Table 5: Evaluation of dataset (T-G/S/B): Target *vs.* Gold, LexSiS and Baselines.

Among the instances from the gold standard not accounted for by LexSiS, our system presents more lexical substitutions synonymous with the target word than the baselines, and a greater amount of simpler synonyms (65.00%) than the frequency (47.06%) and random (4.76%) baselines (see results for dataset (T-G/S/B) in Table 5). We have found that the dataset (T-S/B) is the only case where the frequency baseline presents a higher percentage of suggesting simpler synonyms (31.29%) than LexSiS (27.33%). Nevertheless, LexSiS shows a higher meaning preservation percentage.

System	Synonym (%)	Simpler Syn. (%)	Equal Syn. (%)	Complex Syn. (%)	Simpler Syn. Glob. (%)	Equal Syn. Glob. (%)	Complex Syn. Glob. (%)
Random	66.00	18.18	20.20	60.61	12.08	**13.42**	40.27
Frequency	72.11	**43.40**	18.87	**37.74**	**31.29**	13.61	**27.21**
LexSiS	**76.00**	35.96	**18.42**	45.61	27.33	14.00	34.67

Table 6: Evaluation of dataset (T-S/B): Target *vs.* LexSiS and Baselines.

These results are consistent with the state of the art of lexical simplification for English. In SemEval-2012 shared task for lexical simplification, only one lexical simplification system

among nine systems presented a higher ranking (0.496) than the frequency baseline (0.471), according to a pairwise kappa metric (Specia et al., 2012).

Our results are comparable with the lexical simplification system for English presented in (Biran et al., 2011), whose frequency baseline for English target words (medium frequency) is very similar to ours (42.86%). The results of their system are higher for English and they make use of a comparable corpus. Our inter-annotator agreement rate (Fleiss' kappa score of 0.330) is also consistent with their pairwise inter-annotator agreement, where kappa score was between 0.35 and 0.53.

Our evaluation methodology includes five possible answers for each pair of instances, as in SIMPL method (Yatskar et al., 2010), although we present the annotators with the words in their original context. SIMPL reaches 66% precision for the 100 top pairs using a probabilistic model based on the edit history in simple Wikipedia.

Other methods, which include syntactic operations as well, approach lexical simplification as a monolingual machine translation problem and use machine translation measures to evaluate them (Zhu et al., 2010), together with human judges (Woodsend and Lapata, 2011a).

In summary, LexSiS is the first approach to lexical simplification in Spanish, a language where only syntactic simplification has been previously performed (Bott and Saggion, 2012). Our system uses free resources, such as a dictionary and the Web as corpus. LexSiS is based on the analysis of a small sample of data and it does not require a parallel corpus to function.

8 Conclusion and Future work

In the past few years, automatic text simplification has rapidly become an important technology for the information society, due to its potential application in information access and the benefits it may bring to people with special needs. One important aspect of text simplification is lexical simplification: the selection of simpler/easier substitutes for difficult words. Automatic lexical simplification has, in most cases, been applied to English datasets, with many recent approaches relying on the availability of large parallel simplified and non-simplified documents to train statistical methods.

In this paper we have presented LexSiS, the first implemented lexical simplification system for Spanish, which does not require a parallel corpus for its implementation. LexSiS, rather, uses available texts found on the Web and a freely available Thesaurus. LexSiS uses a word vector model to identify the correct sense of the word to be replaced, and then selects the simplest synonym using simplicity criteria based on word frequency and word length.

We have evaluated three aspects of LexSiS: (i) its ability to identify a correct synonym for the target to be simplified; (ii) its ability to provide a simpler word substitute; and (iii) its performance as a whole. We have compared LexSiS with two baselines, one of which is a hard-to-beat procedure based on word frequency.

For each of the datasets used in the evaluation, LexSiS shows a higher meaning preservation percentage, offering more synonymous lexical substitutions than the baseline. For the overall of the datasets, LexSiS proposes simpler synonyms than the frequency baseline.

We believe our research makes the following contributions: we present a well-grounded procedure for the simplification of Spanish text, which could possibly be a language independent method, due to its small-scale resources that can be found for many other languages. It repre-

sents the first computational implementation of a lexical simplification algorithm for Spanish. Finally, we use a well-designed and viable evaluation methodology for lexical simplification.

Based on the evaluation, we have found various aspects of the method which need to be taken into further consideration. First, we are facing the problem of out-of-dictionary words. One way to overcome this problem is by inducing possible word-substitutes with methods similar to the word vector model we use here, as well as larger datasets. Second, frequency computation in our procedure is not carried out considering word sense disambiguation, but by performing counts disregarding senses. It is possible that taking the word sense into consideration before computing its frequency could improve the results. Finally, in this work we only consider substitution of single lexical units – the identification of multi-word expressions and collocations, such as *poco a poco* (‘*little by little*’), substituted with its simpler synonym *gradualmente* (‘*gradually*’), will be addressed in future work.

Acknowledgments

We would like to thank Ricardo Baeza-Yates for his wise comments and Joan Codina for mathematical advice on the modeling of simplicity. We are also grateful for the annotators for their effort and patience: Toni Carbajo, Roberto Juan Carlini and Alfonso Rello. We also thank the anonymous reviewers for their constructive comments.

The research described in this paper arises from a Spanish research project called Simplext: An automatic system for text simplification (`http://www.simplext.es`). Simplext is led by Technosite and partially funded by the Ministry of Industry, Tourism and Trade of the Government of Spain, by means of the National Plan of Scientific Research, Development and Technological Innovation (I+D+i), within strategic Action of Telecommunications and Information Society (Avanza Competitiveness, with file number TSI-020302-2010-84). We are grateful to fellowship RYC-2009-04291 from Programa Ramón y Cajal 2009, Ministerio de Economía y Competitividad, Secretaría de Estado de Investigación, Desarrollo e Innovación, Spain and to the doctoral fellowship FI-DGR by Generalitat de Catalunya.

References

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

Aluísio, S. M., Specia, L., Pardo, T. A. S., Maziero, E. G., and de Mattos Fortes, R. P. (2008). Towards Brazilian Portuguese automatic text simplification systems. In *ACM Symposium on Document Engineering*, pages 240–248.

Bautista, S., Gervás, P., and Madrid, R. (2009). Feasibility analysis for semiautomatic conversion of text to improve readability. In *The Second International Conference on Information and Communication Technologies and Accessibility*.

Bautista, S., León, C., Hervás, R., and Gervás, P. (2011). Empirical identification of text simplification strategies for reading-impaired people. In *European Conference for the Advancement of Assistive Technology*, Maastricht, the Netherlands.

Biran, O., Brody, S., and Elhadad, N. (2011). Putting it simply: a context-aware approach to lexical simplificaion. In *Proceedings of the 49th Annual Meeting of the Association for*

Computational Linguistics: Human Language Technologies, pages 496–501, Portland, Oregon, USA. Association for Computational Linguistics.

Bott, S. and Saggion, H. (2011). An unsupervised alignment algorithm for text simplification corpus construction. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 20–26, Portland, Oregon. Association for Computational Linguistics.

Bott, S. and Saggion, H. (2012). Text Simplification Tools for Spanish. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey. ELRA.

Burstein, J., Shore, J., Sabatini, J., Lee, Y.-W., and Ventura, M. (2007). The automated text adaptation tool. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. (Demonstrations)*, pages 3–4.

Carroll, J., Minnen, G., Canning, Y., Devlin, S., and Tait, J. (1998). Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In *Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10.

Caseli, H. M., Pereira, T. F., Specia, L., Pardo, T. A. S., Gasperin, C., and Aluísio, S. M. (2009). Building a brazilian portuguese parallel corpus of original and simplified texts. In *10th Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2009)*.

Chandrasekar, R., Doran, D., and Srinivas, B. (1996). Motivations and methods for text simplification. In *16th International Conference on Computational Linguistics*, pages 1041–1044.

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

Coster, W. and Kauchak, D. (2011b). Simple English Wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA: Association for Computational Linguistics*, pages 665–669.

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

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

Drndarevic, B. and Saggion, H. (2012). Towards automatic lexical simplification in Spanish: an empirical study. In *Proceedings of the NAACL HLT 2012 Workshop Predicting and improving text readability for target reader populations (PITR 2012)*.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Hyönä, J. and Olson, R. (1995). Eye fixation patterns among dyslexic and normal readers: Effects of word length and word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(6):1430.

Lal, P. and Ruger, S. (2002). Extract-based summarization with simplification. In *Proceedings of the ACL 2002 Automatic Summarization / DUC 2002 Workshop*.

Landis, J. and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. (1990). Introduction to WordNet: An on-line lexical database*. *International journal of lexicography*, 3(4):235–244.

Padró, L., Collado, M., Reese, S., Lloberes, M., and Castellón, I. (2010). FreeLing 2.1: Five years of open-source language processing tools. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Rayner, K. and Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, 14(3):191–201.

Saggion, H., Gómez-Martínez, E., Etayo, E., Anula, A., and Bourg, L. (2011). Text Simplification in Simplext: Making Text More Accessible. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 47.

Salton, G., Wong, A., and Yang, C. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Siddharthan, A. (2002). An architecture for a text simplification system. In *In LEC02: Proceedings of the Language Engineering Conference (LEC02*, pages 64–71.

Specia, L. (2010). Translating from complex to simplified sentences. In *Proceedings of the 9th international conference on Computational Processing of the Portuguese Language*, pages 30–39, Berlin, Heidelberg.

Specia, L., Jauhar, S., and Mihalcea, R. (2012). Semeval-2012 task 1: English lexical simplification. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*.

Turney, P. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

Vossen, P., editor (2004). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, volume 17. Oxford Univ Press.

Woodsend, K., Feng, Y., and Lapata, M. (2010). Generation with quasi-synchronous grammar. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 513–523. Association for Computational Linguistics.

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

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

Wubben, S., van den Bosch, A., and Krahmer, E. (2012). Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*.

Yatskar, M., Pang, B., Danescu-Niculescu-Mizil, C., and Lee, L. (2010). For the sake of simplicity: Unsupervised extraction of lexical simplifications from wikipedia. In *Proceedings of the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 365–368.

Zhu, Z., Bernhard, D., and Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. In *Proceedings of The 23rd International Conference on Computational Linguistics*, pages 1353–1361.

Zipf, G. (1935). *The Psychobiology of Language*. Houghton Mifflin, Boston, MA.