Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 35 (2016) 200–218

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Simplifying words in context. Experiments with two lexical resources in Spanish☆

Horacio Saggion[a,*], Stefan Bott[b], Luz Rello[c]

[a] *Universitat Pompeu Fabra, c/ Tanger 122, Barcelona 08018, Spain*
[b] *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Pfaffenwaldring 5b, 70569 Stuttgart, Germany*
[c] *HCI Institute, School of Computer Science, Carnegie Mellon University, USA*

Received 18 November 2013; received in revised form 10 July 2014; accepted 6 February 2015
Available online 14 February 2015

Abstract

In this paper we study the effect of different lexical resources for selecting synonyms and strategies for word sense disambiguation in a lexical simplification system for the Spanish language. The resources used for the experiments are the Spanish EuroWordNet, the Spanish Open Thesaurus and a combination of both. As for the synonym selection strategies, we have used both local and global contexts for word sense disambiguation. We present a novel evaluation framework in lexical simplification that takes into account the level of ambiguity of the word to be simplified. The evaluation compares various instances of the lexical simplification system, a gold standard, and a baseline. The paper presents an in-depth qualitative error analysis of the results.

Keywords: Lexical simplification; Spanish; Text simplification; Evaluation

1. Introduction

Automatic text simplification is a technology to adapt the content of a text to the specific needs of particular individuals or target populations so 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, or to be used in specific information access tasks such as information extraction. 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. 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; Carroll et al., 1998), Brazilian Portuguese

☆ This paper has been recommended for acceptance by R.K. Moore.

* Corresponding author. Tel.: +34 93 542 1119; fax: +34 93 542 2517.
E-mail addresses: horacio.saggion@upf.edu (H. Saggion), stefan.bott@upf.edu (S. Bott), luz.rello@upf.edu (L. Rello).

http://dx.doi.org/10.1016/j.csl.2015.02.001
0885-2308/

(Aluísio and Gasperin, 2010), Japanese (Inui et al., 2003), French (Seretan, 2012), Italian (Dell'Orletta et al., 2011; Barlacchi and Tonelli, 2013), and Basque (Aranzabe et al., 2012). Text simplification, as a general task, is similar to other NLP tasks, such as machine translation, paraphrasing, text summarization or sentence compression. The mixed nature of the task and the specific requirements of each aspect, however, make text simplification not fully comparable to the mentioned NLP problems. While text summarization, for example, tries to select the most relevant information from input texts, text simplification rather concentrates on the elimination of superfluous details, by either deleting them or re-phrasing them in a more general way. Text summarization techniques for simplification can be found in Drndarevic and Saggion (2012a) and Stajner et al. (2013).

Our research, concerned with simplification in the Spanish language (Saggion et al., 2011), has produced a text simplification system made up of components for reducing the syntactic complexity of sentences, deleting unnecessary information, rewriting numbers, normalizing reporting verbs, and substituting difficult words by their simpler synonyms (Bott et al., 2012; Drndarevic et al., 2013; Bott and Saggion, 2014). It is this last technology, lexical simplification, which is the focus of the present work. Lexical Simplification aims at replacing difficult words with easier synonyms, while preserving the meaning of the original text segments. Lexical simplification requires the solution of at least two problems: First, the finding of a set of synonymic candidates for a given word, generally relying on a dictionary or a lexical ontology and, second, replacing the target word by a synonym which is easier to read and understand in the given context. For the first task, lexical resources such as WordNet (Miller et al., 1990) can be used. For the second task, different strategies of word sense disambiguation (WSD) and simplicity computation are requited.

Even if there is a considerable number of approaches to lexical simplification in different languages, an estimation of how different lexical resources and WSD strategies impact the task have not yet been studied. There is also no previous work which addresses the question in how far the level of ambiguity of a word influences the degree of success in automatic lexical simplification. The goal of this paper is to address these gaps presenting LexSiS (Bott et al., 2012), a system for Spanish lexical simplification which is parametrized in a way it can use different lexical resources which provide word senses and lists of synonyms. Hence, the main contributions of this paper are:

- A description of a lexical simplification procedure for the Spanish language.
- A comparison of the performance of our lexical simplification system with two different lexical resources (Open Thesaurus and EuroWordNet), in addition to a combined version of the two.
- A comparison of two different strategies for word sense disambiguation, one which only considers the local context of a target word and another which assumes that each target word has only one meaning per text and takes all local contexts for a given target into account.
- An evaluation that assesses the performance of the system depending on different levels of the ambiguity of target words.
- A quantitative and qualitative analysis of the results.

The rest of the paper is organized as follows: In Section 2 we discuss the related work and the context in which our proposal has to be seen. In Section 3 we present a corpus study which provided insights in the human production of lexical simplifications and guided the development of our system. In Section 4 we describe our system, including the alternative resources it can work with and alternative strategies to perform word sense disambiguation. Section 5 explains the evaluation framework and presents the experimental results, while Section 6 draws some conclusions on the use of different lexical resources and disambiguation methods. In Section 7 we present an in-depth error analysis, which complements the quantitative analysis. Section 8 concludes the paper with a summary of the main results and an outlook on future work.

2. Related work on lexical 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). Many approaches to lexical simplification (Carroll et al., 1998; Lal and Ruger, 2002; Burstein et al., 2007) used WordNet in order to find appropriate word substitutions. Bautista et al. (2011) instead use a dictionary of synonyms. As a measure of lexical simplicity most of the cited approaches (Carroll et al., 1998; Lal and Ruger, 2002; Burstein et al., 2007) have relied on word frequency, with the exception of Bautista et al. (2011), who use word length as a predictor for lexical simplicity. Since both word frequency and word length

have been shown to correlate to the cognitive effort in reading (Rayner and Duffy, 1986) they are combined in several works (Keskisärkkä, 2012).

Concerning the measurement of word complexity, recent studies show that both frequency and length influence readability and understandability at least for specific conditions such as dyslexia. For instance, in an eye-tracking study with 46 participants (23 with dyslexia) (Rello et al., 2013), the authors found that more frequent words improved the readability of text for people with dyslexia and the presence of shorter words improved their comprehension score. However, it is worth mentioning that it is not possible to fully dissociate word length from word frequency in language because both parameters are naturally related. In natural language longer words tend to be less frequent (Jurafsky et al., 2001).

As for the overall simplification method, 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. Given a word and a set of possible substitutes, their language model is used to identify which synonyms best fit the context of the word to be replaced. Additionally, a second factor is taken into account: the probability that the word is "easy" – which the authors claim can be calculated based on factors such as word length, word frequency, or information on word use from a psycholinguistic database.

In Bautista et al. (2011) a method is proposed where a readability measure (e.g. Flesch Reading Ease, (Flesch, 1949)) is used to guide the selection and replacement of words in the text. Words which are considered complex either in length or in number of syllables are looked up in a lexical database (e.g. WordNet) in order to find a list of possible replacements. A replacement which is shorter or has a lower number of syllables is chosen as a substitute.

More recently, the availability of the Simple English Wikipedia (SEW) (Coster and Kauchak, 2011), 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, 2011; Coster and Kauchak, 2011; Wubben et al., 2012). This includes some new approaches to lexical simplification, which are the most important points of reference for our work.

Yatskar et al. (2010) use edit histories from the SEW to identify pairs of possible synonyms and the combination of SEW and EW in order to create a set of lexical substitution rules of the form $x \rightarrow y$. Given a pair of edit histories eh_1 and eh_2, they identify which words from eh_1 have been replaced in order to make eh_2 "simpler" than eh_1. A probabilistic approach is used to model the likelihood of a replacement of word x by word y being made because y is "simpler". In order to estimate the parameters of the model, various assumptions are made, such as considering that word replacement in SEW is due to simplification or normal edition and that the frequency of editions in SEW is proportional to that of editions in EW. In their evaluation, subjects are presented with substitution pairs ($x \rightarrow y$) obtained using different methods including a man-made dictionary of substitutions and random and frequency-based baselines obtained from all possible substitution pairs. Subjects are asked to evaluate how x is when compared to y – more simple, more complex, equally simple/complex, not a synonym or "?". Although the proposed approach performs worse than the dictionary, it is better than the two baselines.

Biran et al. (2011) also rely on the SEW/EW combination without paying attention to the edit history of the SEW. They use context vectors to identify pairs of words which occur in similar contexts in SEW and EW (using cosine similarity). WordNet is used as filter for possible lexical substitution rules ($x \rightarrow y$). In their approach a word complexity measure is defined which takes into account word length and word frequency. Given a pair of "synonym" words (w_1,w_2), their raw frequencies are computed on SEW ($\text{freq}_{\text{sew}}(w_i)$) and EW ($\text{freq}_{\text{ew}}(w_i)$). A complexity score for each word is then computed as the ratio between its EW and SEW frequencies (i.e. $\text{complexity}(w) = \text{freq}_{\text{ew}}(w)/\text{freq}_{\text{sew}}(w)$). The final word complexity combines the frequency complexity factor with the length factor in the following formula: $\text{final\_complexity}(w) = \text{complexity}(w) * \text{len}(w)$. As an example, for the word pair (canine, dog) the following inequality will hold: $\text{final\_complexity}(canine) > \text{final\_complexity}(dog)$. This indicates that *canine* could be replaced by *dog* but not vice versa. During text simplification, they use a *context-aware* method comparing how well the substitute word fits the context, filtering out possibly harmful rule applications which would select word substitutes with the wrong word sense. Their work is interesting because they use a Vector Space Model to capture lexical semantics and, with that, context preferences. In our work we also rely on a formula that combines frequency and length, however for Spanish there is no such a parallel dataset as the SEW and EW.

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 (2011) and Specia (2010), drawing on work by Caseli et al. (2009), use standard statistical machine translation machinery for text simplification. In this case, lexical simplification

Table 1
Corpus examples.

Original sentence	Simplified sentence	Alignment
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 already opened 203 branches in a total of 28 countries worldwide, also in Spain, starting this Monday.*)	El Banco de China tiene oficinas en muchos países del mundo. Ahora, también tiene una oficina en España. (*The Bank of China has offices in many countries of the world. Now, also has an office in Spain.*)	1–2
La sucursal, que ocupa una superficie de 1.100 metros cuadrados, se ubica en el número 3 del Paseo de Recoletos de Madrid. (*The branch, which occupies an area of 1,100 square meters, is located in number 3 of Paseo de Recoletos in Madrid.*)	La oficina del Banco de China en Madrid está en el Paseo de Recoletos. (*The Bank of China office in Madrid is at Paseo de Recoletos.*)	1–1
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. (*To show its importance, according to 2009 data, ICBC had a total of 386 723 employees, only in China, with a total of 16,232 branches.*)		1–0

is treated as an implicit part of the machine translation problem. 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, although for a limited range of phenomena, with a surprisingly small bi-dataset of only 4483 sentences.

It is worth noting that a lexical simplification task was recently proposed in the SemEval evaluation (Specia et al., 2012) where given a word and a set of possible substitutes, systems have to identify the simpler synonym(s). The task was in fact rather complex as demonstrated by the evaluation results: only one of the participating systems was able to perform better than a baseline.

3. Corpus analysis

In order to study the problem at hand, we have gathered a corpus of text simplification in Spanish, consisting of 200 informative texts obtained from a Spanish news agency Servimedia. The articles have been classified into four categories: national news, international news, society and culture. We then obtained simplified versions of the said texts, courtesy of the DILES (Discurso y Lengua Española) group of the Autonomous University of Madrid. Simplifications have been created manually, by trained human editors, following easy-to-read guidelines and methodology suggested by Anula (2008, 2009). The corpus has been automatically annotated using the FreeLing toolkit for part-of-speech tagging and named entity recognition (Padró et al., 2010). Furthermore, a text aligning algorithm based on Hidden Markov Models has been developed to obtain sentence-level alignments (Bott and Saggion, 2011). The automatic alignments have then been manually corrected through a graphical editing tool within the GATE framework (Cunningham et al., 2002). Any of the following correlations between original and simplified sentences is possible: *one to one*, *one to many* or *many to one*, as well as cases where there is no correlation (cases of information compression or expansion). Examples of alignments in the corpus are shown in Table 1. The corpus has been used to carry out several studies in text simplification (Rello et al., 2013; Stajner and Saggion, 2013; Bautista and Saggion, 2014) and although it has not been formally released, it can be obtained directly by contacting the first author of this paper.

We have conducted an empirical analysis of 40 document pairs from the corpus (590 sentences with 246 and 324 in the original (O) and simplified (S) sets respectively). Our methodology, explained more in depth in Drndarevic and Saggion (2012b), consists in observing lexical changes applied by trained human editors. In addition to that, we carried out quantitative analyses 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 (see Section 2). 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

Table 2
Lexical Substitution Table for common words and named entities.

Original	Simple
impartir/give	*pronunciar*/pronounce
informar/inform	*decir*/say
inmigrante/immigrant	*extranjero*/foreigner
letras/literature	*literatura*/literature
Comité Español de Representates de Personas con Discapacidad/Spanish Committee of Representatives of People with Disability	*CERMI*
el PSOE/the PSOE	*el Partido Socialista Obrero Español*/Party of the Spanish Socialist Workers
Gonzales-Sinde	*Angeles Gonzales-Sinde*
el artista/the artist	*el artista Pablo Picasso*/the artist Pablo Picasso
la ciudad andaluza/the Andalusian city	*Granada*

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 analyze the totality of the words in the parallel corpus. Second, we analyze 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. A small sample of the LST can be appreciated in Table 2.

3.1. Word length

Analyzing the total of 10,507 words (6595 and 3912 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 toward 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 analyze the frequency, a dictionary based on the Referential Corpus of Contemporary Spanish (Corpus de Referencia del Español Actual, CREA)[1] has been compiled for our research. 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 categorization of words according to their frequency and elegant presentation and interpretation of results. However, in Section 4 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*

[1] http://corpus.rae.es/creanet.html.

Table 3
The distribution of *n*-frequency words in original and simplified texts.

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%

(*'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 3.

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 analyzed 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 4.2 we propose criteria for determining word simplicity exploiting these findings.

4. Lexical simplification in Spanish: LexSiS

LexSiS tries to find the best substitution candidate (a word lemma) for every word which has an entry in a lexical database, such as the Spanish Open Thesaurus or the Spanish WordNet. The substitution operates in two steps: first the system tries to find the most appropriate substitution set (a SWN synset or its equivalent in SOT) 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 the lexical database. 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.

4.1. Lexical resources

As already mentioned, 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. Spanish is one of the languages represented in EuroWordNet (Vossen, 2004), although its scope is more modest.[2] We have tried three lexical

[2] The Spanish part of EuroWordNet contains only 50,526 word meanings and 23,370 synsets, in comparison to 187,602 meanings and 94,515 synsets in the English WordNet 1.5.

resources in LexSiS: the **Spanish Open Thesaurus**[3] (SOT), the **Spanish EuroWordNet** (SWN), and **combination of SWN and SOT** (SWN + SOT). We describe each of them below.

The Spanish Open 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 entry in SOT for the word *hoja* is as in (a).

(a) hoja | 3
- | acero | espada | puñal | arma_blanca
- | bráctea | hojilla | hojuela | bractéola
- | lámina | plancha | placa | tabla | rodaja | película | chapa | lata | viruta | loncha | lonja | capa | . . .

The first line of the entry represents the target word and states that there are three different meanings. The three lines that follow list synonyms for the three word meanings (*blade*, *leaf* and *sheet* in English).

A second resource we use is the Spanish EuroWordNet. However, for its use with LexSiS we extracted a dictionary from WordNet which represented synset of SWN in the same format as the Spanish Open Thesaurus. We additionally enriched each entry with hyperonyms (e.g. *organo_de_una_planta/plant organ* in the last sense below) of the word.[4] The SWN entry for *hoja* is given in (b).[5]

(b) hoja | 4
- | instrumento_cortante
- | folio | cuartilla | pliego | hoja_de_papel | papel
- | folio | folio | cuartilla | pliego | hoja_de_papel
- | follaje | órgano | órgano_de_una_planta | órgano_vegetal

The word *hoja* is also semantically ambiguous here and can mean *blade*, *leaf* or *sheet of paper*. The sense for *sheet of paper* is represented by two synsets (second and third lines), a distinction which only captures a subtle difference in meaning.

Finally, we are interested in whether a combination of SWN and SOT is able to produce better substitutions since this combination provides more substitution candidates to choose from. For this end we used a union of SWN synsets and SOT substitution sets and let LexSiS choose freely from the alternative lists of synonym words stemming from the two resources. The combined (SOT + SWN) representation for *hoja* contains all the lines contained in (a) and in (b).

4.2. 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 to a linear distribution.

[3] We used version 2. The Spanish Open Thesaurus is included in the distribution of OpenOfficeOrg and can be found at http://openoffice-es.sourceforge.net/thesaurus/.

[4] We expected the inclusion of hyperonyms to be beneficial. In the Open Thesaurus we observed that hyperonyms were often listed as possible substitutes, alongside synonyms. We also observed that LexSiS in combination with the Open Thesaurus could often produce good simplifications which were hyperonyms. In the error analysis based on the human evaluation (cf. Section 7) we found that, on the other hand, the inclusion of hyperonyms can also lead to over-simplifications in cases where WordNet only provides very general top-level categories as immediate hyperonyms.

[5] It can be seen in this example that SWN lists many multi-word expressions. At the moment we do not have a module that can detect the same kind of multi-word expressions in the linguistic pre-process, so we have to ignore these entries. In the future we plan to include the treatment of such expressions, which we expect to lead to a better coverage of the system.

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, assuming that words which are shorter than 5 characters do not display a real difference in simplicity[6]:

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

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:

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

Now the combination of the two values is

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

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 Open Thesaurus for given contexts taken from a corpus of simple non-parallel text.[7] 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 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 $\alpha_1 = -0.39$[8] and $\alpha_2 = 1.11$.

With this formula, the simplicity score is mainly determined by the word's frequency, but there are also examples where the word length determines the final decision of one candidate over another. An example can be seen in (1). The two most frequent substitution candidates from SOT for the word "*especie*" ("species") are "*tipo*" ("type") and "*grupo*" ("group"). The frequencies of the two candidates in the training corpus are 2499 and 5437, respectively, while the frequency of "*especie*" is 427. With this, "*grupo*" is more frequent than "tipo", but since the latter is shorter, it receives a final simplicity score of 3.77, as opposed to 3.76 for the former. As a result, the metric prefers "*tipo*", which is also the more adequate choice for the given example.

(1) Descubren en Valencia un nuevo ESPECIE de pez prehistórico.
'A new SPECIES of prehistoric fish is discovered in Valencia.'

4.3. Word Vector Space Model

In order to measure lexical similarity between words and contexts, we used a Word Vector Space Model (Sahlgren, 2006), a type of Vector Space Model (Salton et al., 1975) which represent words in very local contexts. Word Vector Space Models are a good way of modeling lexical semantics (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

[6] 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. Section 5). 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.

[7] It should be noted that for the estimation of these weights we did not use parallel corpus data. The use of parallel data for this purpose is problematic because the parallel corpus we used for the corpus study described in Section 3 only contained one substitution solution for any lexically simplified word, where other alternatives might have been correct, too. Often these human generated alternatives were also not listed in the lexical resources.

[8] 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.

Table 4
The 15 most dominant dimensions in two sample vectors for the words "*museo*" ("museum") "*jurisdicción*" ("jurisdiction")and "*competencia*" ("competence").

museo		jurisdicción		competencia	
arte	59	aquel	4	comisión	26
prado	47	constitucional	4	nacional	23
director	38	cesión	3	sobre	19
nacional	30	caso	3	defensa	19
nuevo	23	contra	3	entre	15
picasso	19	español	3	nuevo	14
historia	18	ejercicio	2	comunidad	12
bello	16	delito	2	problema	12
contemporáneo	16	bajo	2	autoridad	12
guggenheim	16	deber	2	según	11
reina	16	cuestión	2	estado	10
sofía	15	falta	2	transferir	9
parir	13	representar	1	europeo	9
colección	12	dado_que	1	asumir	9
visitar	12	condición	1	desleal	8
madrid	12	dictar	1	feroz	8
sala	10	equivaler	1	exclusivo	8
natural	11	entender	1	deber	8
bilbao	11	reconocer	1	libre	8
obra	10	desvincular	1	invadir	7
arqueología	11	adecuar	1	reconocer	7

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 able to represent word senses and to discriminate between them. For example, vectors for word senses can be built as the sum of word vectors which share one meaning. Standard measures, such as cosine distance, can then be used to determine the distances between a given context and different word senses.

We trained a vector model on a 8M word corpus of Spanish on-line news. We lemmatized the corpus with FreeLing, version 3 (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). Stop words, such as articles, pronouns and auxiliary verbs and conjunctions were excluded. The vector model has *n* dimensions, where *n* is the number of lemmas 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.

The 20 most dominant dimensions of the vectors for three sample words can be found in Table 4. The dimensions with the strongest extension for "*museo*" ("museum") correspond to words like "Prado" (a famous museum in Madrid), "director", "national" and "Picasso". "*Jurisdicción*" ("Jurisdiction") is a low frequency word, as can be seen from its vector representation and "*competencia*" ('competence") is a possible simplification substitute for it. The words that can be found frequently in the context of "*jurisdicción*" and "*competencia*" center around entities that can have competence, such as states ("*estado*") and communities ("*comunidad*"), and the adjectives that express such entities, like national ("*nacional*"), Spanish ("*español*"), European ("*europeo*") or constitutional ("*constitutional*") competence. Jurisdiction and competence can also be recognized, respected or transferred and the corresponding Spanish verbs ("*reconocer*", "*respetar*" and "*transferir*") can be found here. The overlap among the context words of the two target concepts is higher than appreciated by looking at only the most dominant dimensions in Table 4. Out of the 84 dimensions that have a non-zero count in the vector for "*jurisdicción*", 49 also have a non-zero count in the vector for "*competencia*".

4.4. Word sense disambiguation in LexSiS

We implemented two different methods to carry out word sense disambiguation, which we call the *local* and the *global* method. The local method only looks at the local context of a target word assuming that the local context provides enough information for disambiguation (Krovetz, 1998), while the global method takes all the occurrences of each target word within a text and constructs a combined representation of the contexts in which they are found, assuming the one sense per discourse hypothesis (Gale et al., 1992).

For the **local method** we extract a vector from the surrounding 9-word window. As already mentioned, 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 of 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 to all candidate words and applying several thresholds in order to eliminate candidates which are either not much simpler or seem to differ too much from the context.

The **global method** works largely like the local method, with one difference. We assume that each target word has only one meaning in each text it appears. So, instead of extracting a local context vector for each target instance of a word w, we extract all of the local vectors for w found in the text. Then we sum over all of these local vectors, and obtain a global vector for w and compare it to the vectors representing word senses.

4.5. 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 Open Thesaurus 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 0.001% (calculated on the corpus we used to train the vector model in Section 4.2). 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 by comparing the simplicity scores of pairs of target words and proposed substitutions in which we could not observe a clear difference in word complexity.

Many of the alternatives proposed by the lexical resource 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 in a similar way in which we estimated the frequency threshold: we applied LexSiS to a series of non-parallel input texts and manually annotated those substitutions which clearly did not fit into the context because they did not preserve the original word sense. We also annotated cases of good and clearly meaning-preserving substitutions. Then we set the cosine filter to a value which minimized the cases which did not match the context while still not filtering out too many clearly synonymous substitution candidates.

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), and second, there are also cases where the highest scoring substitute is the same as the original lemma. In both cases the original word is preserved.

5. Evaluation

In this section we present the experimental set-up employed to evaluate the different resources and word sense disambiguation strategies for LexSiS. The evaluation was conducted thoroughly, rating the degree of simplification and the preservation of meaning of the substitutions.

5.1. Baseline

As baseline we use the method of Devlin and Unthank (2006). It replaces a word with its most frequent synonym, presumed to be the simplest. This frequency baseline was also used in SemEval-2012 shared task for lexical simplification (Specia et al., 2012). We also implemented two baselines which we did not use in the final evaluation described below. Both of these baselines were intended to assess the contribution made by our simplicty metric (cf. Section 4.2), which

implemented findings from our corpus study and from the recent literature (cf. Sections 2 and 3). One of these baselines was derived choosing the alternative with the highest simplicity score $score_{simp}$ among all of the listed alternatives, without the application of word sense disambiguation. We found that in only 14.5% of the cases it produced results different from the frequency baseline. In addition, since no word sense disambiguation was applied, we could hardly observe cases where both baselines produced different acceptable synonyms, which made the assessment in terms of simplicity improvement practically impossible. A further baseline we implemented applied word sense disambiguation but used simple frequency instead of $score_{simp}$ as a measure of simplicity. Again, the difference between this baseline and the final system outputs was so small that we could not hope to derive statistically significant results concerning their difference.

5.2. Evaluation dataset

From the corpus described in Section 3 we extracted a set of manual lexical simplifications in order to create our evaluation dataset. As in Biran et al. (2011), we only selected examples where the system could produce an alternative. In addition to 55 human created lexical simplifications, we included substitutions proposed by LexSiS using the 3 different resources (55 lexical substitutions each). Substitutions produced by the baseline method were also included in the evaluation dataset. Thus, the evaluation consists of baseline substitutions (**FREQ**), **SWN** substitutions, **SOT** substitutions, **SWN + SOT** substitutions, and **Gold** manual lexical substitutions. Concerning the distribution of substitution targets across parts of speech, we found that most of the substitutions affected nouns (73.6%), followed by adjectives (15.09%) and verbs (11.3%).

The two WSD strategies were also evaluated, having as a result 165 simplifications using a local strategy (**Local WSD**) and 165 simplifications using a global strategy (**Global WSD**).

Since we wanted to evaluate the meaning presentation as well as the simplification, each of the substitutions were inserted in their original sentences. In total we had 550 lexical substitutions to be compared with the original target words. We manually corrected the ungrammatical examples[9] and deleted the duplicated lexical substitutes giving a total of 456 unique lexical substitutions. We believe this is a reasonable size for an evaluation dataset; in Yatskar et al. (2010) they use a total of 200 simplification examples, and in Biran et al. (2011) 130 sentences were used. Below, we show two examples of a sentence with its original word (O) and the lexical substitution proposed by our system using SWN + SOT.

(2) (O) Se encuentra a favor de la lucha contra la DESIGUALDAD y la pobreza.
'It is in favor of the fight against INEQUALITY and poverty.'

(SWN + SOT) Se encuentra a favor de la lucha contra la IRREGULARIDAD y la pobreza.
'It is in favor of the fight against IRREGULARITY and poverty.'

5.3. Ambiguity bands

We divided the target words of the dataset in three levels of difficult depending on their degree of ambiguity. For measuring the degree of ambiguity we considered the average number of senses per word given by WordNet and Open Thesaurus. Hence, our dataset has three ambiguity bands: low (from 0.5 to 1.5 senses, 49.08% of the dataset), medium (from 2 to 2.5 senses, 25.09% of the dataset) and high (3 senses or more, 25.84% of the dataset).

5.4. Evaluation design

We created a multiple choice questionnaire, presenting two sentences for each item. The test included all the unique lexical substitutions. Each item contained one sentence with a simplification example and the same sentence with the original word. These sentences were presented in counterbalanced and randomized order to the annotator (i.e., either as Original vs. SYSTEM or SYSTEM vs. Original). For each pair of sentences, the annotators were asked two questions to choose one option in each of them, one regarding the meaning preservation ("the sentences above have the same meaning" vs. "the sentences above do not have the same meaning") and another one regarding the simplicity degree

[9] The correction only affected inflections and agreement errors, since we could not use a morphological generator in the experimental setting.

Table 5
Results for the different resources using local and global WSD.

Method	WSD	Meaning preservation	Simpler
SWN	Local	**63.2**	68.2
SWN	Global	62.2	**69.2**
SOT	Local	62.0	66.7
SOT	Global	62.0	66.7
SWN + SOT	Local	58.6	66.4
SWN + SOT	Global	60.6	68.2
Baseline	–	52.6	**71.1**
Gold	–	**75.9**	69.3

Bold represent the highest values.

Table 6
Results for meaning preservation for different ambiguity levels.

Ambig. band	WSD	Meaning preservation			Simpler synonyms		
		SWN	SOT	SOT + SWN	SWN	SOT	SOT + SWN
Low	Local	60.0	62.5	56.0	70.0	**70.0**	68.7
Low	Global	61.0	62.5	56.9	70.3	**70.0**	68.9
Med.	Local	65.5	51.1	56.9	72.2	60.9	70.3
Med.	Global	61.7	51.1	63.1	**73.0**	60.9	**70.8**
High	Local	**67.4**	**70.4**	**66.7**	60.6	65.8	58.3
High	Global	65.3	**70.4**	66.1	62.5	65.8	64.1

Bold represent the highest values.

("the first of the sentences above is simpler than the second" vs. "the first of the sentences above is not simpler than the second"). Five 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. The five participants annotated all the instances of the datasets, achieving a Fleiss' kappa score of 0.33. Hence, we can assume we have a fair agreement (Fleiss, 1971; Landis and Koch, 1977), comparable with other inter-annotator agreements in related work, where kappa score was between 0.35 and 0.53 (Biran et al., 2011).

5.5. Results

Table 5 shows a direct comparison of the performance of LexSiS with different resources and the two different WSD methods, the baseline and the gold standard. In Table 6 we show the results for WSD and simplicity by ambiguity level. The scores for *meaning preservation* were calculated over all data points and represent the judgments of the participants about whether both variants of the sentence had the same meaning. The scores for simplicity were calculated over those data points which were judged as being synonymous in order to be able to achieve independent scores for synonymity and simplicity. The WSD methods of *local* and *global* correspond to the two ways of constructing context vectors described in Section 4.

Calculated over the whole corpus segment from which the evaluation dataset was extracted, we found that LexSiS produced a simplification for 1.40% of the words with the use of SWN and for 1.27% with the use of SOT. Note that we could not calculate the recall of the system because neither the gold standard nor the human annotation allowed us to make a reliable estimation of the number of simplification targets, i.e. the words which could benefit from lexical simplification. Theoretically, the number of simplification targets could be determined by human annotation, this can be expected to be subject to a hight variation between different annotators.

6. Discussion

We observe (Table 5) that LexSiS shows consistently much higher *meaning preservation* (synonymity) scores than the baseline. For the whole dataset, without distinction of levels of ambiguity LexSiS with SWN achieves a score of 63.16% in comparison to 52.59% produced by the baseline. When LexSiS uses other resources (SOT or SWN + SOT) the scores are only slightly lower. The score for the gold standard is higher (75.92%), but surprisingly it does not even get close to 100%, which shows that human judges are reluctant to accept alternatives as being fully synonymous and gives an idea of the difficulty of the task. A surprise is that the combination of the two resources (SWN + SOT) performs much worse than the two resources on their own. We hoped that the word disambiguation component would perform better with the availability of more synonym sets, because the cosine distance between the vectors for these sets should lead to a selection of the set with the most coherent meaning, penalizing sets which include words that are not coherent with the rest of the set. This expectation was not met. A possible alternative would be to align the resources so that equivalent synsets are merged together providing additional synonym choices for equivalent senses. We will investigate this in future work.

Turning to the question whether the use of SOT shows a worse performance than that of SWN, we can observe a slightly better performance of LexSiS + SWN than LexSiS + SOT in both categories (Table 5), but none of the differences are statistically significant. This is an interesting result, because it suggests that a simpler resource like SOT can lead to nearly the same level of performance than the use of a more sophisticated resource like SWN, whose quality is controlled in a much stricter way. In SOT we often observed incoherent synonyms sets, cases in which more than one synonym sets appear to represent the same word sense and word senses which are not represented by a single synonym set. Nevertheless, we think that the use of the word sense disambiguation component in combination with the threshold that filters out candidates with a high cosine distance to the context can partially remedy the shortcomings of the thesaurus.

The left part of Table 6 (meaning preservation) suggests that words with a higher degree of ambiguity are easier to disambiguate, which might come as another surprise. We suspect that this reveals a possible shortcoming of the resources used: words which are listed with only one word sense are often still ambiguous, while the entries of words with many listed senses tend to be disambiguated better in the dictionary entry.

Turning to the production of synonyms which are perceived as being actually simpler than the original, again SWN outperforms SOT and SWN + SOT (cf. Table 5). In the right part of Table 6 (simpler synonyms) we can observe that the success of producing simpler synonyms depends very much on the level of ambiguity of the target word. Highly ambiguous words are harder to simplify, while words with low ambiguity are easier. Words with a medium level ambiguity show a curious behavior: for SWN and the combination of SWN and SOT these words appear to be easier to simplify.

Table 6 also shows that the global method of choosing synonyms (one synonym for each target per text) systematically outperforms the local method in its ability to produce simpler substitutes. We attribute this to the fact that the summed vectors for target word contexts present much richer context information and are much more reliable than the rather sparse vectors for individual contexts. The assumption that each target word has only one meaning per text proves to be quite helpful. For example, in sentence (3) the original word *noción* (notion) has been substituted with the word *representacíon*(representation) with the local method and the much more acceptable word *idea* with the global method, using SWN + SOT.

(3) . . . vivimos en un mundo en el que se ha perdido la NOCIÓN de autoridad.
'. . . we live in a world where the NOTION of authority has been lost'

In Table 5 the baseline outperforms LexSiS and even the gold standard in the simplification task, but it has to be taken into account that the simplicity scores were calculated only over those data instances that were actually perceived by the annotators as being synonymous. This amounts to saying that the frequency baseline would perform extraordinarily well if all the non-synonym productions were filtered out, which is first impossible and would result in a much lower coverage (i.e. much less substitutions produced) than that of LexSiS. As a curious matter of fact, only 51.85% of the gold standard cases were both judged as being synonymous and being simpler, which illustrates the difficultly of the combined task.

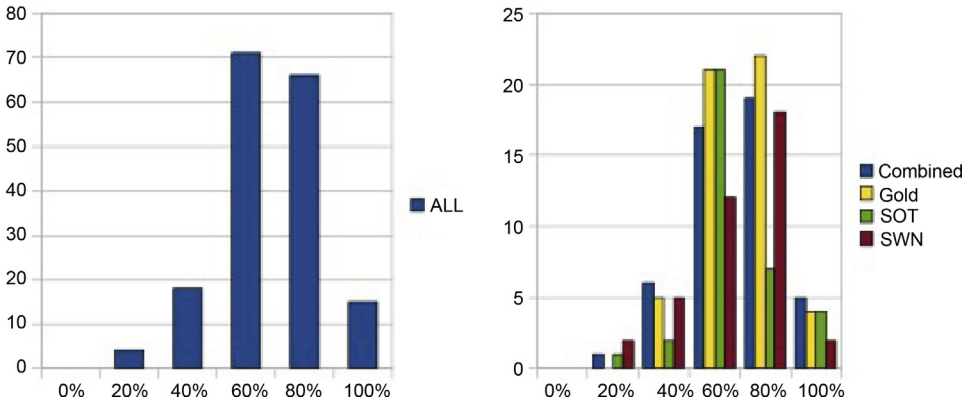

Fig. 1. Distribution of different rating levels for simplicity. The left graphic shows the distribution for all judgments, the right one the judgments for selected modes separately. The level of 100% corresponds to 5 subjects.

Turning to significance, we only found a significant effect between different methods on the meaning preservation; the gold standard preserved significantly more meaning in their substitutions that the rest of the methods ($F(9, 2262) = 4.062$, $p < 0.01$). This finding is hardly surprising, given the difficulty of the word sense disambiguation task. It is probably more interesting to note that, while the gold standard achieves higher scores for simplicity, this score is not much higher than the score for LexSiS with the use of SWN. Also, even if the scores for LexSiS with different configurations are lower, the difference to the gold standard could not be shown to be statistically significant.

7. Error analysis

In order to be able to interpret the numbers obtained in the evaluation, we carried out an in-depth error analysis, looking into different kinds of human judgment configurations which deviated very much from our expectations or from the general tendencies.

We first looked at the worst cases, the ones which were accepted by none of the participants. Interestingly, for the simplicity judgment there were no such cases: All substitutes were judged as being simpler by at least one of the participants. There were cases which were judged as being simpler by only one participant out of five (i.e. 20% acceptance), but even those cases were quite rare, as can be seen in the right side of Fig. 1 we only found 2 data points produced with SWN and one produced with OT (ignoring for the sake of simplicity the baseline and the global strategies). But also total acceptance by all five subjects was rare in this category: 15 data-point. The distribution of the frequency of scores (with discrete values which directly represent the number of participant that judged every example pair positively) can be seen in Fig. 1: the left side shows the distribution of all simplicity judgments, while the right side shows the distribution separately for each modality. The maximum lies at 80% (the cases where 4 out of 5 persons agreeing that the substitute is simpler to understand is the most frequent case) for the gold standard and SWN, while it lies at 60% for SOT. It is also interesting to note that the gold standard does not deviate too much from the general distribution.

This clearly shows two things: First, human judges do not appear to have clear-cut intuitions on simplicity, on which they unanimously agree. This is an interesting finding, even if it can be easily explained: simplicity is a gradual value and if the actual difference in simplicity is not very high, individual preferences start to play a larger role. Second, and as we have already seen in Table 5, an automatic simplicity metric can nearly achieve the level of adequacy of the gold standard. In fact, our simplicity measure proves to be quite successful.

Turning to meaning preservation, represented in Fig. 2, the distribution looks different, with a maximum at the 100% acceptability level (agreement among all 5 subjects that the substitute is actually synonymous). There are also a series of cases where all participants fully agree that the substitute does not preserve the meaning: 1 produced by SOT, 2 by SWN, 4 by the SWN + SOT and – interestingly – 1 taken directly from the gold standard. Human judges thus seem to have crisper, inter-individually more coherent intuitions about synonymity than they have about simplicity. This also influences inter-annotator agreement when it is computed as Fleiss' Kappa score. While the Kappa for meaning preservation is 0.42, the Kappa for simplicity judgments is still fair, but with 0.24 it is much lower.

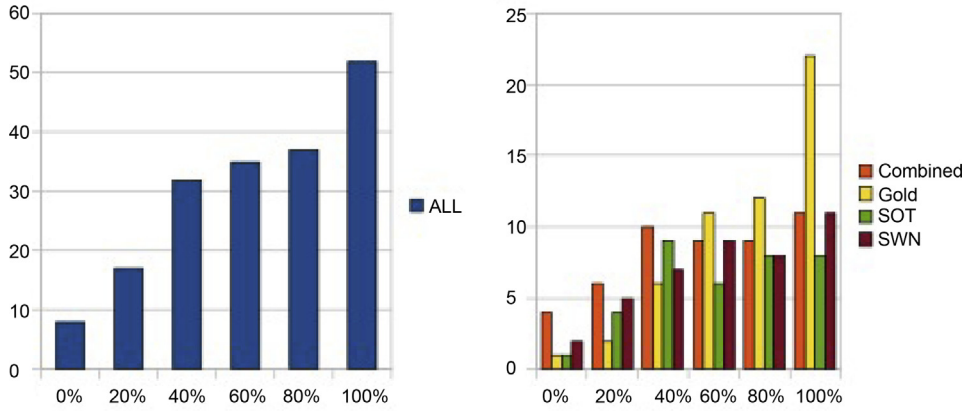

Fig. 2. Distribution of different rating levels for meaning preservation. The left graphic shows the distribution for all judgments, the right one the judgments for selected modes separately. The level of 100% corresponds to 5 subjects.

The reasons for the system producing bad substitutes are many-fold: sometimes the targets for replacements represent very abstract or complicated concepts, sometimes the lexical resources list no acceptable alternatives and sometimes the target words are ambiguous in surprising ways. In Section 6 we noted that lexical simplification shows a curious behavior with respect to the degree of ambiguity of the target words: there was a general tendency toward the system better preserving the meaning in the case of more ambiguous words, while the degree of simplicity showed the inverse effect: words with lower ambiguity had a stronger tendency to be replaced with words that were judged as being simpler.

As for low ambiguity words we often found that the entries in the lexical resources fail to list existing word senses. A good example is the word "*envergadura*", which literally means "wingspan", but also has the more abstract meaning of "importance" or "magnitude". We found that in this case SWN represented only the literal meaning, while SOT only the abstract reading. More importantly, SWN listed no synonyms at all and only one hyperonym: "distance" (which is partially adequate since "*envergadura*" describes a distance, but only in the case of the distance between the tips of two wings or the extremes of a boat). Accordingly, the produced synonym "*distancia*"/"distance" in example in example (4) was rated very poorly (accepted by only 20% of the subjects, i.e. one participant), but the system could not have done better on the basis of the resource used. The use of SOT lead to the synonym "*valor*"("value" or "boldness"), which is not perfect, but rated as synonym by 60% (3) of the subjects. The gold standard solution "*dimensión*" ("dimension"), which had an 100% rate of agreement on meaning preservation (and a 60% score on simplicity), is neither represented in the entry of SWN, nor in that of SOT as can be seen below:

(SWN)	envergadura|1
-	distancia
(OT)	envergadura|1
-	|alcance|importancia|trascendencia|repercusión|resonancia|difusión |calibre|significación|valor|magnitud|gravedad

(4) Como muestra de su ENVERGADURA, (. . .) el ICBC tenía en nómina a un total de 386,723 empleados, (. . .)
'To show its IMPORTANCE, (. . .) ICBC had an total payroll of 386.723 employees, (. . .)'

Also "*mortalidad*" ("mortality") was listed as an unambiguous entry in WN. In this case it is truly unambiguous, but none of the listed synonyms are actually simpler. Since we used hyperonyms for WN entries, LexSiS also found the possible substitute "rate" (because mortality is a rate of deaths) and since this word is much simpler (both more frequent and shorter) it was chosen as a substitute, leading again to a very low acceptability as a synonym. There are also missing entries: For example, the word "*creatividad*" ("creativity") was not represented in SOT at all.

Compared to the low ambiguity band we just discussed, also the highly ambiguous target words presented some difficulty, but of a different kind. We found that such words sometimes represented quite complex concepts, such as

"*jurisdicción*" ("jurisdiction")[10] which has very long entries, both in OT and in WN. These entries include closely related concepts alongside real synonyms. Apparently LexSiS picked out the right word senses (both using OT and WN) and produced quite acceptable synonyms ("*competencia*"/"competence" in the case of SWN and "*término*"/"premises" in that of OT) for the target context given in (5), but since the synonyms are still technical terms, they were not judged as much simpler (only 2 and 3 subjects accepted them as simpler, respectively). The entries for SWN and OT used to simplify (5) are shown below.

(SWN)	jurisdicción | 4
-	| poder | potencia
-	| circunscripción
-	| función | competencia | incumbencia | atribución | campo | órbita | ámbito | terreno | esfera | dominio
(OT)	jurisdicción | 4
-	| autoridad | mando | poder | potestad | atribución | facultad | competencia
-	| bailía | bailiazgo | territorio | bailiaje | baile | bailío | alcaldía | diputación | municipio | localidad | ayuntamiento | consistorio | casa consistorial | casa_de_la_villa
-	| dominio | demarcación | término | territorio | comarca
-	| lugar | emplazamiento | sitio | punto | lado | zona | área | extensión | superficie | espacio | división | circunscripción

(5) (. . .) pecios (. . .) que se encuentran en las aguas de soberanía o JURISDICCIÓN española.
'(. . .) wrecks (. . .) that are found in the waters under the sovereignty or Spanish JURISDICTION.'

The gold standard solution "*autoridad*" ("authority") shared the same fate and was accepted as a good synonym (100% positive ratings), but was not perceived as much simpler (only 60% positive ratings). In general the substitutes of such highly ambiguous words were seldom unanimously judged as being more difficult, but they were never fully and unanimously accepted as being simpler, either. This lead to the quite mediocre average simplicity score for high ambiguity items which can be seen in Table 6.

Another serious and non-trivial problem we could detect has to do with the distinction of true synonyms and hyperonyms. When using LexSiS with SWN, we included hyperonyms as possible substitutes because often a hyperonym is a good simplification by virtue of the fact that it represents a more general concept. This is also motivated by the fact that in the gold standard we could find substitutions by hyperonyms, such as "*autovía*" ("freeway") being changed to "*carretera*" ("road") or "*novelista*" ("novelist") to "*escritor*" ("writer"). Such examples made us believe that the system would benefit from the inclusion of the hyperonyms. We found cases where the system created good hyperonym substitutions, often in the same cases where the gold standard suggested a hyperonym, but we also found that in many cases the hyperonyms were too general to serve as a good substitute. For example, a hyperonym of "*esclavo*" ("slave") in SWN is the very genetic term "*persona*" ("person") and "*porvenir*" ("future") has both "*potenciál*" ("potential", described as "the inherent capacity for coming into being") and "*tiempo*" ("time") as hyperonyms, both very inappropriate as simplification substitutes. We found that the entry for "*tenor*" (referring to the type of male singer) proves to be paradigmatic for our dilemma: "*tenor*" itself has no direct synonyms in WN, but it has two hyperonym synsets: "voice" and "singer". The latter would make a good simplification candidate, which clearly shows that the inclusion of hyperonyms is beneficial. Unfortunately the system chose the former option ("voice"), which resulted in a high score for simplicity (80% acceptance), but a low score for meaning preservation (40%). The ideal solution would be to include only hyperonyms with a low degree of abstraction with respect to the target word (e.g. novelist/writer) and exclude hyperonyms with a very high degree of abstraction. WordNet does not distinguish between different levels of abstraction and gives no reliable clues on when to include and when to exclude a hyperonym.

The inclusion of hyperonyms can also be problematic in the case of two co-hyponyms being coordinated. We found the following specific case in (6):

(6) (. . .) la concesión de la Orden de las Artes y las Letras de España al novelista y ensayista José Luis Sampedro, (. . .) . . .
'(. . .) the concession of the Order of Arts and Literature to the Spanish novelist and essayist José Luis Sampedro, (. . .)'

[10] The vector representation for "*jurisdicción*" and it's simplification "*competencia*" can be found in Table 4.

The problem here is that both "*novelista*" ("novelist") and "*ensayista*" ("essayist") can be simplified to "*escritor*" ("writer", stemming from SWN) or "*autor*" ("author", produced with SOT). Accordingly the fully simplified version reads as "the author and author José Luis Sampedro", which is clearly not adequate. Even if the subjects were not shown the two simplifications in one individual sample, they did not seem to have found sequences like "the author and essayist" very appealing and, surprisingly, rated neither "author" nor "writer" as much simpler (ratings of 40% and 60%, respectively). So, both including and excluding hyperonyms as possible simplification candidates leads to certain problems and we still do not know if they do more good or more harm. The ideal solution would certainly be to find a way in which to include only those hypernonyms which are not very generic. But this is a quite complex problem in its own right and we would like to address this question in future work.

Finally, we noted that the global method did not work equally well for the use of SOT and SWN: We found that, when SWN was used, LexSiS could provide an alternative substitution in 15.5% of the cases, while for SOT this was the true for only 5.7% of the cases. More importantly for SWN in nearly 60% of these mentioned cases the global method could produce a substitute where the local method found no better alternative to the original word. For SOT this percentage was only 31.8%.

8. Conclusions and future work

Lexical simplification, an essential component in a text simplification system, aims at replacing difficult words with easier synonyms, while preserving the meaning of the original text. Most works on lexical simplification have been undertaken for the English language and rely on the WordNet lexical database as a resource to find synonyms. A handful of approaches also rely on the availability of huge parallel or comparable datasets such as the English and the Simple English Wikipedias.

In this paper we have presented our approach for the simplification of vocabulary in Spanish texts. The approach relies on a lexical resource to find possible replacements for a word to be simplified, a vector model to represent "word meaning" and perform word sense disambiguation, and a simplicity criteria for choosing the simplest synonym. Here we have compared the effect of using different lexical resources and disambiguation strategies. In particular we have instantiated experiments with the Spanish WordNet and the Spanish Open Thesaurus as lexical resources. Where disambiguation methods are concerned, we have tried local and global disambiguation strategies.

The comparison of two different lexical resources shows how far the quality of the resource used influences the quality of the lexical simplifications the system produces. Since Open Thesaurus is an open collaborative effort, the quality of the thesaurus entries is not strongly controlled, a factor which we could see reflected in poorly separated word senses and even missing representation for some senses. We could find differences in the performance depending on the lexical resource used, but it was surprisingly low and not statistically significant. This is a good result because the main bottleneck for the most language dependent part of a lexical simplification system like LexSiS is the availability of lexical resources. Our evaluation suggests that thesauri may be a good substitute for more sophisticated lexical ontologies.

Another contribution of this paper is the comparison of two WSD methods: one based on local context and the second global method based on summed local context on the text level. We could show that the global method performs better for the lexical substitution task. The choice of the lexical resource is only one of a list of possible optimizations for the LexSiS system. There are other possibilities we would like to explore in the future, such as the use of TF*IDF weights and the investigation of in how far the size of the window which represents the context influences the system performance. After developing LexSiS, a new lexical resource for Spanish lexical simplification called CASSA (Rello, 2014) has been produced which is based on Google n-grams and Open Thesaurus; we plan to compare LexSiS to CASSA in the near future. Our lexical simplification system could also help to normalize paraphrases to the simplest word choice, which could be useful in plagiarism detection (Barrón-Cedeño et al., 2013).

In our future work we will extend the functionalities of our simplification system to cover other languages starting with English. The availability of lexical resources in English and huge textual datasets to model lexical knowledge will facilitate porting LexSiS to other languages.

Acknowledgments

We are grateful to the reviewers who helped us improve the paper and provided useful insights to continue our work. The first author wish to acknowledge support from programa Ramón y Cajal 2009 (RYC-2009-04291), project number TIN2012-38584-C06-03 from Ministerio de Economía y Competitividad, Secretaría de Estado de Investigación, Desarrollo e Innovación, Spain, and from the project ABLE-TO-INCLUDE (CIP-ICT-PSP-2013-7/621055).

References

Aluísio, S.M., Gasperin, C.,2010. Fostering digital inclusion and accessibility: the PorSimples project for simplification of Portuguese texts. In: NAACL/HLT, Young Investigators Workshop on Computational Approaches to Languages of the Americas, YIWCALA'10. ACL, Stroudsburg, PA, USA, pp. 46–53.

Anula, A., 2008. Lecturas adaptadas a la enseñanza del español como L2: variables lingüísticas para la determinación del nivel de legibilidad. In: Pastor, Roca (Eds.), La evaluación en el aprendizaje y la enseñanza del español como LE/L2. , pp. 162–170.

Anula, A., 2009. Tipos de textos, complejidad lingüística y facilicitación lectora. In: Actas del Sexto Congreso de Hispanistas de Asia, pp. 45–61.

Aranzabe, M., de Ilarraza, A., Gonzalez-Dios, I., 2012. First approach to automatic text simplification in Basque. In: Natural Language Processing for Improving Textual Accessibility (NLP4ITA) Workshop Programme, pp. 1–8.

Barlacchi, G., Tonelli, S., 2013. ERNESTA: a sentence simplification tool for children's stories in Italian. In: CICLing (2), pp. 476–487.

Barrón-Cedeño, A., Vila, M., Martí, M.A., Rosso, P., 2013. Plagiarism meets paraphrasing: insights for the next generation in automatic plagiarism detection. Comput. Linguist. 39 (4).

Bautista, S., Saggion, H., 2014. Can numerical expressions be simpler? Implementation and demonstration of a numerical simplification system for Spanish. In: LREC, pp. 956–962.

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

Biran, O., Brody, S., Elhadad, N.,2011. Putting it simply: a context-aware approach to lexical simplificaion. In: ACL/HLT. ACL, Portland, OR, USA, pp. 496–501.

Bott, S., Saggion, H.,2011. An unsupervised alignment algorithm for text simplification corpus construction. In: Proceedings of the Workshop on Monolingual Text-to-Text Generation. ACL, Portland, OR, pp. 20–26.

Bott, S., Saggion, H., 2014. Text simplification resources for Spanish. J. Lang. Resour. Eval. 48 (1), 93–120.

Bott, S., Rello, L., Drndarević, B., Saggion, H., 2012. Can Spanish be simpler? LexSiS: lexical simplification for Spanish. In: CoLing.

Burstein, J., Shore, J., Sabatini, J., Lee, Y.-W., Ventura, M., 2007. The automated text adaptation tool. In: NAACL/HLT (Demonstrations), pp. 3–4.

Carroll, J., Minnen, G., Canning, Y., Devlin, S., Tait, J., 1998. Practical simplification of English newspaper text to assist aphasic reader. In: Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology, pp. 7–10.

Caseli, H.M., Pereira, T.F., Specia, L., Pardo, T.A.S., Gasperin, C., Aluísio, S.M., 2009. Building a Brazilian Portuguese parallel corpus of original and simplified texts. In: CICLing.

Chandrasekar, R., Doran, D., Srinivas, B., 1996. Motivations and methods for text simplification. In: CoLing, pp. 1041–1044.

Coster, W., Kauchak, D.,2011. Learning to simplify sentences using Wikipedia. In: Proceedings of the Workshop on Monolingual Text-to-Text Generation. ACL.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., 2002. GATE: a framework and graphical development environment for robust NLP tools and applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics.

De Belder, J., Deschacht, K., Moens, M.-F., 2010. Lexical simplification. In: Proceedings of Itec2010: 1st International Conference on Interdisciplinary Research on Technology, Education and Communication, https://lirias.kuleuven.be/handle/123456789/268437

Dell'Orletta, F., Montemagni, S., Venturi, G.,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, pp. 73–83.

Devlin, S., Unthank, G., 2006. Helping aphasic people process online information. In: The International ACM SIGACCESS Conference on Computers and Accessibility, pp. 225–226.

Drndarevic, B., Saggion, H., 2012a. Reducing text complexity through automatic lexical simplification: an empirical study for Spanish. Proces. Leng. Nat. 49, 13–20.

Drndarevic, B., Saggion, H., 2012b. 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).

Drndarevic, B., Stajner, S., Bott, S., Bautista, S., Saggion, H., 2013. Automatic text simplification in Spanish: a comparative evaluation of complementing modules. In: Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CiCLing 2013).

Fleiss, J.L., 1971. Measuring nominal scale agreement among many raters. Psychol. Bull. 76 (5), 378–382.

Flesch, R., 1949. The Art of Readable Writing. Harper, New York.

Gale, W.A., Church, K.W., Yarowsky, D.,1992. One sense per discourse. In: Proceedings of the Workshop on Speech and Natural Language. HLT, http://dx.doi.org/10.3115/1075527.1075579, ISBN 1-55860-272-0, 233-237.

Inui, K., Fujita, A., Takahashi, T., Iida, R., Iwakura, T., 2003. Text simplification for reading assistance: a project note. In: Proceedings of the 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP), pp. 9–16.

Jurafsky, D., Bell, A., Gregory, M., Raymond, W.D., 2001. Evidence from reduction in lexical production. Frequency and the Emergence of Linguistic Structure, vol. 45., pp. 229.
Keskisärkkä, R., 2012. Automatic Text Simplification via Synonym Replacement. Linköping University (Master's thesis).
Krovetz, R., 1998. More than one sense per discourse. In: NEC Princeton NJ Labs., Research Memorandum.
Lal, P., Ruger, S., 2002. Extract-based summarization with simplification. In: Proceedings of the ACL 2002 Automatic Summarization/DUC 2002 Workshop.
Landis, J., Koch, G., 1977. The measurement of observer agreement for categorical data. Biometrics, 159–174.
Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K., 1990. Introduction to WordNet: an on-line lexical database. Int. J. Lexicogr. 3 (4), 235–244.
Padró, L., Collado, M., Reese, S., Lloberes, M., Castellón, I., 2010. FreeLing 2.1: five years of open-source language processing tools. In: LREC, Valletta, Malta.
Rayner, K., Duffy, S.A., 1986. Lexical complexity and fixation times in reading: effects of word frequency, verb complexity, and lexical ambiguity. Mem. Cogn. 14 (3), 191–201.
Rello, L., Baeza-Yates, R., Dempere, L., Saggion, H., 2013. Frequent words improve readability and short words improve understandability for people with dyslexia. In: INTERACT'13, Cape Town, South Africa.
Rello, L., 2014. DysWebxia. A Text Accessibility Model for People with Dyslexia. Universitat Pompeu Fabra, Barcelona, Spain (Ph.D. thesis).
Saggion, H., Gómez-Martínez, E., Etayo, E., Anula, A., Bourg, L., 2011. Text simplification in Simplext: making text more accessible. Rev. Soc. Esp. Proces. Leng. Nat. 47.
Sahlgren, M., 2006. The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations Between Words in High-Dimensional Vector Spaces.
Salton, G., Wong, A., Yang, C., 1975. A vector space model for automatic indexing. Commun. ACM 18 (11), 613–620.
Seretan, V., 2012. Acquisition of syntactic simplification rules for French. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey.
Siddharthan, A., 2002. An architecture for a text simplification system. In: Proceedings of the Language Engineering Conference, pp. 64–71.
Specia, L., Jauhar, S., Mihalcea, R., 2012. Semeval-2012 task 1: English lexical simplification. In: SemEval 2012.
Specia, L., 2010. Translating from complex to simplified sentences. In: The International Conference on Computational Processing of the Portuguese Language, Berlin/Heidelberg, pp. 30–39, http://dx.doi.org/10.1007/978-3-642-12320-7_5, ISBN 3-642-12319-8, 978-3-642-12319-1.
Stajner, S., Saggion, H., 2013. Adapting text simplification decisions to different text genres and target users. Proces. Leng. Nat. 51, 135–142.
Stajner, S., Drndarevic, B., Saggion, H., 2013. Eliminación de frases y decisiones de división basadas en corpus para simplificación de textos en español. Comput. Sist. 17 (2).
Turney, P., Pantel, P., 2010. From frequency to meaning: vector space models of semantics. J. Artif. Intell. Res. 37 (1), 141–188.
Vossen, P. (Ed.), 2004. EuroWordNet: A Multilingual Database with Lexical Semantic Networks, vol. 17. Oxford University Press.
Woodsend, K., Lapata, M., 2011. WikiSimple: automatic simplification of wikipedia articles. In: Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI), pp. 927–932.
Woodsend, K., Feng, Y., Lapata, M.,2010. Generation with quasi-synchronous grammar. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. ACL, pp. 513–523.
Wubben, S., van den Bosch, A., Krahmer, E., 2012. Sentence simplification by monolingual machine translation. In: ACL.
Yatskar, M., Pang, B., Danescu-Niculescu-Mizil, C., Lee, L., 2010. For the sake of simplicity: unsupervised extraction of lexical simplifications from Wikipedia. In: ACL, pp. 365–368.
Zhu, Z., Bernhard, D., Gurevych, I., 2010. A monolingual tree-based translation model for sentence simplification. In: CoLing, pp. 1353–1361.
Zipf, G., 1935. The Psychobiology of Language. Houghton Mifflin, Boston, MA.