CASSA: A Context-Aware Synonym Simplification Algorithm

Ricardo Baeza-Yates
Yahoo Labs &
Web Research Group
Universitat Pompeu Fabra,
Barcelona, Spain
rbaeza@acm.org

Luz Rello
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA, USA
luzrello@acm.org

Julia Dembowski
Department of
Computational Linguistics
Saarland University, Germany
juliad@coli.uni-saarland.de

Abstract

We present a new context-aware method for lexical simplification that uses two free language resources and real web frequencies. We compare it with the state-of-the-art method for lexical simplification in Spanish and the established simplification baseline, that is, the most frequent synonym. Our method improves upon the other methods in the detection of complex words, in meaning preservation, and in simplicity. Although we use Spanish, the method can be extended to other languages since it does not require alignment of parallel corpora.

1 Introduction

Simplified text is crucial for some populations to read effectively, especially for cognitively impaired people such as people with autism spectrum disorder (Evans et al., 2014; Orasan et al., 2013), aphasia (Carroll et al., 1999), dyslexia (Rello et al., 2013a), Down syndrome (Saggion et al., 2015; Saggion et al., 2011), or other intellectual disabilities (Huenerfauth et al., 2009).

In fact, the United Nations (1994) proposed a set of standard rules to leverage document accessibility for persons with disabilities. Text simplification attempts to solve this problem automatically by reducing the complexity of the lexicon, syntax, or semantics while attempting to preserve its meaning and information content (Siddharthan, 2006). Among all the types of text simplification this paper focuses on lexical simplification.

Lexical simplification methods require language resources, such as simplified corpora or synonyms dictionaries. For languages with less resources than English, *e.g.* no Simple Wikipedia (Biran et al., 2011; Yatskar et al., 2010) or less representation in WordNet, such as Spanish,[1] the creation of lexical simplification methods is more challenging.

Our approach makes use of two free resources, Google Books Ngram Corpus and the Spanish OpenThesaurus, as well as real web frequencies to create a lexical simplification system. Our system improves upon the state of the art for lexical simplification in Spanish and the established simplification baseline, *i.e.*, the most frequent synonym, in several aspects: complex word detection, meaning preservation, and simplicity. We also show the coverage of our technique in a collection of books in Spanish, as this is another relevant measure of a simplification algorithm. The method is language independent and given that these resources are available in other languages, it could be easily extended to other languages with similar language resources.

The rest of the paper is organized as follows. The next section presents related work. Then in Section 3 we present the simplification algorithm while in Sections 4 and 5 we present our experimental evaluation. Finally, in Section 6 we discuss our results, extensions to other languages, and outline future work.

[1]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. (Vossen, 2004).

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1380–1385,
Denver, Colorado, May 31 – June 5, 2015.

2 Related Work

Lexical simplification is a kind of text simplification that aims at the word level. It can be performed through the substitution of words by simpler synonyms, by adding a definition, or by showing simpler synonyms. Most of the approaches aim at the substitution of complex words.

To find appropriate synonyms, many approaches use WordNet (Burstein et al., 2007; Carroll et al., 1999; Lal and Ruger, 2002). De Belder *et al.* (2010) apply explicit word sense disambiguation with a latent words language model. Devlin and Unthank (2006) use dictionaries. Aluisio and Gasperin (2010) use a thesaurus and lexical ontologies.

More recently, Biran *et al.* (2011) and Yatskar *et al.* (2010) used Simple English Wikipedia, in combination with the standard English Wikipedia for their lexical simplification algorithms using machine learning.

There are also machine translation based approaches (Coster and Kauchak, 2011; Specia, 2010) as well as hybrid approaches (Narayan and Gardent, 2014; Siddharthan and Angrosh, 2014) that are also able to handle lexical simplification, since the translation model maps words from the non-simplified language to words of the simplified language.

The closest algorithm to ours is LexSiS (Bott et al., 2012; Saggion et al., 2013), that uses the Spanish OpenThesaurus and a corpus that contains 6,595 words of original and 3,912 words of manually simplified news articles. To the best of our knowledge this is the first and only lexical simplification algorithm for Spanish. Hence, we use it here as the state-of-the-art in our evaluation.

To the best of our knowledge, our approach is novel in using the Google Books Ngram corpus for the word context, Open Thesaurus for the synonyms, and real web frequencies for disambiguating synonym candidates. However, Google Ngram have been previously used to find synonyms, for instance to expand user queries by including synonyms (Baker and Lamping, 2011).

3 Method

CASSA (Context-Aware Synonym Simplification Algorithm) is a method that generates simpler synonyms of a word. It takes into consideration the context and the web frequency of the complex word for disambiguation.

3.1 Resources

Our method uses the following two resources:

– **Spanish OpenThesaurus** (version 2): The thesaurus is freely available[2] to be used with OpenOffice.org. This thesaurus provides 21,378 target words (lemmas) with a total of 44,348 different word senses for them. The following is a part of the thesaurus entry for *farol*, which is ambiguous, as it could mean *'lie'*, *'lamp'*, or the adjective *'flashy'*, among others.

```
farol
- embuste|mentira ('lie')
- luz|lámpara|fuego|bombilla ('lamp')
- ostentoso|jactancioso|farolero ('flashy')
```

– **Google Books Ngram Corpus for Spanish** (2012 edition): The corpus consists of n-grams and their usage frequency over time,[3] and is derived from 8,116,746 books, over 6% of all books ever published. The corpus has 854,649 volumes and 83,967,471,303 tokens (Lin et al., 2012).

3.2 Algorithm Description

First, we modified and enriched the Spanish OpenThesaurus and created our List of Senses. Instead of having a target word with different senses, we included the target word in each sense, and we kept a list of unique senses, including for each word its frequency in the Web using a large search engine index. The Spanish OpenThesaurus contains single-word and multi-word expressions. We only treated single-word units, which represent 98% of the cases, leaving out only 399 multi-word expressions, such as *de esta forma ('in this manner')*.

We lemmatized the words because the frequencies were all for inflected word forms as they appear in the Web while we were interested in the lemma frequencies for the synonyms, adding all the

[2] http://openthes-es.berlios.de

[3] http://books.google.com/ngrams

frequencies for each lemma. We take into account the frequency of the words, because previous studies have shown that less frequent words were found to be more challenging for people with and without the most frequent reading disorder, that is, dyslexia (Rello et al., 2013b).

Second, we use the 5-grams in the Google Books Ngram Corpus, where we use the third token of each 5-gram as our target words. The other tokens are the context of the target word. A context is considered valid if all words, including the target word, consist only of lowercase alphabetic characters, to filter for proper names, and is not a stop word, using a standard list of stop words in Spanish.

The lemmatized token is included in the list of target words only if it appears in our List of Senses. The remaining four tokens are the context, kept in a context list. We count the frequency of the target word appearing with that context in the corpus, as well as the frequency of the same context appearing with different target words. See two possible contexts for *noche* and *fortuna* in the examples below:

`era una` *noche* `oscura de` *('it was a dark night of')*
`de probar` *fortuna* `en el` *('to try fortune in the')*

Third, we define the complexity of a word using the relative frequency of the synonyms within the same sense in the List of Senses.

That is, our definition is tailored to web text. For this we use a parameter k such that if a word is k or more times less frequent than one or more of its synonyms, is considered a complex word. We used $k = 10$ as the default threshold to get that 27% of the words have simpler synonyms. We later show how this percentage changes with smaller k.

Finally, for each complex word and the contexts it appears in, we select as simpler synonym the most frequent synonym of the sense that appears most frequently for the n-gram corresponding to that $(word, context)$ pair. That is, to disambiguate the sense, our method uses the context where the target word appears. If the context is not found, we use the most frequent sense (baseline below).

4 Quality Evaluation

4.1 Comparison Points

Baseline: replaces a word with its most frequent synonym (presumed to be the simplest). This baseline has been broadly used in previous lexical simplification studies (Burstein et al., 2007; Carroll et al., 1999; Devlin and Unthank, 2006; Lal and Ruger, 2002), with the exception of (Bott et al., 2012) that used word frequency and length. It is very hard to beat this baseline for simpler synonyms generation. For instance, in SemEval task for English lexical simplification (Specia et al., 2012), only one system out of nine outperformed the frequency baseline. For the complexity part we use the same as our new method. That is, both algorithms consider the same words as complex.

Original	Él **contemplaba** en silencio aquella cruz.
	He was contemplating in silence that cross.
Baseline	Él **veía** en silencio aquella cruz.
	He was seeing in silence that cross.
LexSis	Él **consideraba** en silencio aquella cruz.
	He was considering in silence that cross.
CASSA	Él **miraba** en silencio aquella cruz.
	He was looking in silence that cross.

Figure 1: Example of substitutions performed by the three algorithms.

LexSiS: replaces a word with the output of the state-of-the-art method for Spanish lexical simplification (Bott et al., 2012).

4.2 Frequency Bands

We divided the selected complex words methods into two groups: [LOW], that includes very low frequency complex words, and [HIGH], that contains high frequency complex words. The word frequency ranges from 40 to 2,000 occurrences in Google Books Ngram Corpus for the [LOW] group, and from 2,001 to 1,300,000 for the [HIGH] group.

4.3 Evaluation Datasets

Main dataset: From a set of texts of scientific and literature genres (37,876 words), we randomly selected 20 [LOW] and 20 [HIGH] complex words within the sentence they appear, together with their corresponding candidate for substitution generated by the Baseline, LexSiS, and ours (a valid sentence must had at least 2 different substitutions). We had in total 120 simplification examples (composed by an original and a simplified sentence). Figure 1

shows a set of substitutions along with the original sentence.

Similar studies had smaller or slightly larger evaluation data sets. In Yatskar *et al.* (2010), 200 simplification examples were rated by six annotators (three native, three non-native speakers of English), although only the native speakers annotations were used for the results because they yielded higher inter-annotator agreement. Biran *et al.* (2011) used 130 examples that were judged by three annotators (native English speakers). In Bott *et al.* (2012), three annotators (native speakers of Spanish) rated 69 sentences each of the Spanish lexical simplification performed by LexSiS.

Complexity dataset: This dataset was created to evaluate the degree of complexity of the words selected by the algorithms. Using the same texts as before we extracted 40 random complex words according to LexSiS and 40 according to our method (recall that those also are complex for the baseline).

4.4 Judgment Guidelines

We presented the 200 examples in two different online tests, one for each evaluation dataset. The examples were presented in random order to three native Spanish speakers, frequent readers and non-authors of this paper. For the Main Dataset each annotator rated the simplification examples on two scales: Meaning Preservation –does the transformation preserve the original meaning of the sentence (yes/no); and Simplification –does the transformation result in a simpler sentence (more complex, same complexity or simpler). For the Complexity Dataset the annotators rated the examples on a three point scale (complex, neither complex or simple and simple).

We used Fleiss Kappa (Fleiss, 1971) to measure the inter-annotator agreement for multiple raters. We obtained a reasonable agreement: 0.46 for meaning preservation, 0.54 for simplicity ratings, and 0.41 for complexity. Hence, we have a moderate agreement (Landis and Koch, 1977), comparable with agreements in related literature (Biran et al., 2011; Bott et al., 2012; Yatskar et al., 2010).

4.5 Results

In Table 1 we show the results for the Main Dataset, where in the last column we consider only the simpler synonym substitutions (21 for Baseline, 16 for LexSiS, and 32 for ours) that preserved their meaning (agreement of 2 annotators). In this case, the simplicity performance improves for all the methods. In Table 2 we give the results for the two band frequencies for meaning preservation and simplicity. In the dataset our method overlaps in 15.79% with LexSiS candidates and in 65.79% with the baseline.

Type	**Mean.** (%)	**Simp.** (%)	**SimpSyn.** (%)
Baseline	49.17	60.00	65.08
LexSiS	42.50	35.83	45.83
CASSA	**74.17**	**70.83**	**77.08**

Table 1: Average percentage scores in meaning preservation (Mean.), simplification (Simp.), and simplification among the synonyms (SimpSyn.).

The results of LexSiS are consistent with the ones presented in Bott *et al.* (2012) for the news genre. In that study only for one dataset among three improved upon the frequency baseline in some measures (meaning preservation and global simplicity). As it can be observed from the frequency band results and the complexity measure, LexSiS offers better synonyms for high frequency and not for low frequency words. On the other hand, our method improves with low frequency complex words.

In the complexity evaluation, the prediction accuracy for complex words was only 13.33% for LexSis while was more than double, 34.17%, for ours (idem for the baseline as it used the same complexity criteria). The percentages for the complexity are low as the annotators were regular readers and non-impaired native speakers. For people with language difficulties or cognitive disabilities the accuracy should be higher because people which cognitive disabilities are more sensitive to text simplifications, such as people with Down Syndrome (Saggion et al., 2015), dyslexia (Rello and Baeza-Yates, 2014), or mild intellectual disabilities (Huenerfauth et al., 2009).

5 Coverage Evaluation

As not all possible contexts appear in Google Books Ngrams, we created a corpus made of 195 classic literature books from the 15th century to the 20th century of over 100Mb, to check the coverage of our

Type	Freq.	Meaning (%)	Simp. (%)
Baseline	[HIGH]	40.00	58.33
LexSiS	[HIGH]	41.67	36.67
CASSA	[HIGH]	**73.33**	**70.00**
Baseline	[LOW]	58.33	61.67
LexSiS	[LOW]	43.33	35.00
CASSA	[LOW]	**75.00**	**71.67**

Table 2: Average percentage scores by frequency band.

Case	$k = 10$	$k = 5$	$k = 2$	No k
Comp. words	27.16	38.80	54.24	100.00
Baseline (abs.)	24.07	35.32	50.14	84.43
Baseline (rel.)	88.62	91.03	92.04	84.43
Comp. contexts	27.95	40.03	55.84	100.00
CASSA (abs.)	2.67	4.14	6.44	12.14
CASSA (rel.)	9.55	10.34	11.53	12.14

Table 3: Coverage of the baseline and our method.

method. We included the books that are compulsory readings for secondary and high school in Spain. This corpus is composed by 16,495,885 tokens and 5,886,366 lexical words (without stop words, proper names and punctuation marks).[4]

The coverage of the Spanish Open Thesaurus in our corpus is 88.34%.[5] This is the maximum that any simplification algorithm that uses this resource can obtain. In Table 3 we present the coverage of the baseline and our method depending on the threshold k used to decide what a complex word is and hence a complex content, including the absolute percentages as well as the relative percentages with respect to the complex words or contexts.

For smaller k, the coverage of the baseline increases significantly being the maximum possible 84.43% when all words are considered complex (more than three times the default coverage). On the other hand, our method does not increase much the coverage as that is limited by the context coverage reaching a maximum of 12.14%, only 27% more than the default case ($k = 10$). This maximum, compared with the baseline is a bit more than 14% of the cases, implying that our method is equal to the baseline around 85% of the time.

[4]All the book titles used for this corpus are given in the Appendix of Rello (2014).

[5]Note that this only applies to the corpus used in the coverage, not for the evaluation dataset.

Considering the maximum possible coverage of the baseline and assuming that all non covered sentences contain complex words (most probable case), the simplicity performance of the baseline drops to 53.1% while for ours would be 54.2% (that is, a 2.1% improvement). This should improve if any of the resources used grow.

6 Conclusions and Future Work

Our method improves upon LexSiS and the baseline for all the measures. As we mentioned earlier, even beating the baseline is very hard and we improve upon both other methods by more than 50% in meaning preservation and is 11.8% better than the baseline, the second best, for simplicity. Compared to the results for English of Biran *et al.* (2011) using WordNet, our method has better simplicity scores for low frequency words as well as is in meaning preservation, although they are not directly comparable as different resources are used.

Although Open Thesaurus is available in nine languages and Google Books Ngrams in seven, there are only two languages in both sets: Spanish and German. Hence our method should be easily extended to German. Other languages are also possible with language resources, in particular English.

Acknowledgments

We thank Diego Saez-Trumper, Eduardo Graells and Miguel Ballesteros for their suggestions in the implementation of CASSA.

References

S. M. Aluísio and C. Gasperin. 2010. Fostering digital inclusion and accessibility: the PorSimples project for simplification of Portuguese texts. In *Proc. NAACL HLT '10 Workshop YIWCALA '10*, pages 46–53, Stroudsburg, PA, USA.

S. D. Baker and J. O. Lamping. 2011. Identifying a synonym with n-gram agreement for a query phrase. US Patent 7,925,498.

O. Biran, S. Brody, and N. Elhadad. 2011. Putting it simply: a context-aware approach to lexical simplification. In *Proc. ACL'11*, pages 496–501, Portland, Oregon, USA.

S. Bott, L. Rello, B. Drndarevic, and H. Saggion. 2012. Can Spanish be simpler? LexSiS: Lexical simplification for Spanish. In *Proc. Coling '12*, Mumbay, India.

J. Burstein, J. Shore, J. Sabatini, Yong-Won Lee, and M. Ventura. 2007. The automated text adaptation tool. (demo). In *Proc. NAACL'07*, pages 3–4.

J. Carroll, G. Minnen, D. Pearce, Y. Canning, S. Devlin, and J. Tait. 1999. Simplifying text for language-impaired readers. In *Proc. EACL '09*, pages 269–270.

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

J. De Belder, K. Deschacht, and Marie-Francine Moens. 2010. Lexical simplification. In *Proceedings of LTEC 2010: 1st International Conference on Interdisciplinary Research on Technology, Education and Communication*.

S. Devlin and G. Unthank. 2006. Helping aphasic people process online information. In *Proc. ASSETS '06*, pages 225–226. ACM.

R. Evans, C. Orasan, and I. Dornescu. 2014. An evaluation of syntactic simplification rules for people with autism. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) at EACL*, pages 131–140.

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

M. Huenerfauth, L. Feng, and N. Elhadad. 2009. Comparing evaluation techniques for text readability software for adults with intellectual disabilities. In *Proc. ASSETS '09*, pages 3–10. ACM.

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

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

Y. Lin, J-B. Michel, A. E. Lieberman, J. Orwant, W. Brockman, and S. Petrov. 2012. Syntactic annotations for the Google books ngram corpus. (demo). In *Proc. ACL'12*, pages 169–174.

S. Narayan and C. Gardent. 2014. Hybrid simplification using deep semantics and machine translation. In *ACL'14*, pages 435–445.

United Nations. 1994. Standard Rules on the Equalization of Opportunities for Persons with Disabilities.

C. Orasan, R. Evans, and I. Dornescu, 2013. *Towards Multilingual Europe 2020: A Romanian Perspective*, chapter Text Simplification for People with Autistic Spectrum Disorders, pages 287–312. Romanian Academy Publishing House, Bucharest.

L. Rello and R. Baeza-Yates. 2014. Evaluation of Dyswebxia: A reading app designed for people with dyslexia. In *Proc. W4A '14*, Seoul, Korea.

L. Rello, R. Baeza-Yates, S. Bott, and H. Saggion. 2013a. Simplify or help? Text simplification strategies for people with dyslexia. In *Proc. W4A '13*, Rio de Janeiro, Brazil.

L. Rello, R. Baeza-Yates, L. Dempere, and H. Saggion. 2013b. Frequent words improve readability and short words improve understandability for people with dyslexia. In *Proc. INTERACT '13*, Cape Town, South Africa.

L. Rello. 2014. *DysWebxia. A Text Accessibility Model for People with Dyslexia*. Ph.D. thesis, Universitat Pompeu Fabra.

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

H. Saggion, S. Bott, and L. Rello. 2013. Comparing resources for Spanish lexical simplification. In *SLSP 2013: Statistical Language and Speech Processing*, volume 7978 of *Lecture Notes in Computer Science*, pages 236–247.

H. Saggion, S. Štajner, S. Bott, S. Mille, L. Rello, and B. Drndarevic. 2015. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, In Press.

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

A. Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109.

L. Specia, S. K. Jauhar, and R. Mihalcea. 2012. Semeval-2012 task 1: English lexical simplification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 347–355.

L. Specia. 2010. Translating from Complex to Simplified Sentences. In *PROPOR*, pages 30–39.

P. Vossen. 2004. EuroWordNet: A multilingual database with lexical semantic networks. *International Journal of Lexicography*, 17(2):161–173.

M. Yatskar, B. Pang, C. Danescu-Niculescu-Mizil, and L. Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *Proc. ACL'10*, pages 365–368, Uppsala, Sweden.