CASSAurus: A Resource of Simpler Spanish Synonyms

Ricardo Baeza-Yates,[1] **Luz Rello,**[1*] **Julia Dembowski**[2]

Web Research Group, DTIC[1] Computational Linguistics Dept.[2]
Universitat Pompeu Fabra[1] Saarland University[2]
Barcelona, Spain[1] Saarland, Germany[2]
{rbaeza,luzrello}@acm.org juliad@coli.uni-saarland.de

Abstract

In this work we introduce and describe a language resource composed of lists of simpler synonyms for Spanish. The synonyms are divided in different senses taken from the Spanish OpenThesaurus, where context disambiguation was performed by using statistical information from the Web and Google Books Ngrams. This resource is freely available online and can be used for different NLP tasks such as lexical simplification. Indeed, so far it has been already integrated into four tools.

Keywords: Lexical simplification, synonyms, Spanish OpenThesaurus, Google Books Ngrams, Spanish.

1. Introduction

There are few online free resources regarding synonyms in most languages. That is certainly the case in Spanish, where the OpenThesaurus is the main free resource. However, OpenThesaurus does not have information of what synonyms are simpler or more complex, something crucial to perform lexical simplification. CASSAurus is, to the best of our knowledge, the first freely available resource that ranks synonyms depending on their complexity.

The resource consists of over 40 thousand complex words with their corresponding synonyms depending on the sense of the word. The resource is freely available online.[1]

The rest of the paper is organized as follows. Related work is covered in Section 2.. In Section 3. we summarize the algorithm used to generate the simpler synonyms while Section 4. details the resource. In Section 5. we give an example of the use of the resource and its evaluation. We end with some concluding remarks and future work in Section 6..

2. Related Work

One of the main limitations that previous lexical simplification studies in Spanish point out is the lack of resources in this language.

For instance, there is no Simple Wikipedia in Spanish, while there is a Simple English Wikipedia (Coster and Kauchak, 2011) that has lead to new approaches for lexical simplification in English (Yatskar et al., 2010; Biran et al., 2011).

As far as we know, the existing resources previously used for lexical simplification in Spanish are the following.

Simplext Corpus (Bott and Saggion, 2012) is used in the first lexical simplification system for Spanish (Bott et al., 2012). Simplext Corpus is a set of 200 news articles of which 40 have been manually simplified. The parallel part of this corpus contains 6,595 words of original and 3,912 words of simplified text. All texts have been annotated using Freeling, including part-of-speech tagging, named entity recognition and parsing (Padró et al., 2010).

Another language resource used for Spanish lexical simplification (Bott and Saggion, 2012) is the Spanish OpenThesaurus (SpOT).[2] SpOT is freely available under the GNU Lesser General Public License, to be used with OpenOffice.org. This thesaurus provides 21,378 target words (lemmas) and a total of 44,348 different word senses for them.

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) and this resource was also used for lexical simplification (Saggion et al., 2013). 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. While SpOT is freely available, EuroWordNet is not.

To the best of our knowledge there is no other resource like CASSAurus containing lists of synonyms ranked by their complexity.

3. The CASSA Algorithm

In this section we present a summary of the CASSA algorithm, from which we have generated the language resource that we present in this paper.

CASSA (Context Aware Synonym Simplification Algorithm) (Baeza-Yates et al., 2015) is a method that generates simpler synonyms of a word. Words can be polysemic,[3] that is, they can have different meanings or senses depending on their context. For instance, the Spanish verb *acostar* can mean either 'to go to bed' or 'to reach coast'. In this example, the most common sense is the first. CASSA takes into consideration the context of the complex word for disambiguation in order to find the correct simpler synonyms to show.

[*] Currently at HCI Institute, Carnegie Mellon University, Pittsburgh, USA.

[1] `grupoweb.upf.es/WRG/` & `www.luzrello.com`

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

[3] Polysemy refers to the coexistence of many possible meanings for a word or phrase.

Resources The method is language independent although it was implemented and evaluated for Spanish. It only needs the following two usually freely available resources: (a) a dictionary of synonyms, where we used the Spanish OpenThesaurus and (b) a large n-gram corpus with frequencies, where we used Google Books Ngram Corpus (Michel et al., 2011). Next we detail them:

– **Spanish OpenThesaurus** (version 2): This thesaurus provides 21,378 target words (lemmas) and a total of 44,348 different word senses for them. The following is the thesaurus entry for *mono*, which is ambiguous, as it could mean 'ape', 'overall', or the adjective 'cute'.

(a) mono| 3
- simio|chimpancé|mandril|mico|macaco|gorila| antropoide
- overol|traje de faena
- llamativo|vistoso|atractivo|provocativo| sugerente|resultón|bonito

– **Google Books Ngram Corpus** (2012 edition): The corpus consists of words and phrases (that is, n-grams) and their usage frequency over time. The data is available for download[4] and is derived from 8,116,746 books, over 6% of all books ever published. For Spanish the corpus has 854,649 volumes and 83,967,471,303 tokens (Lin et al., 2012).

Algorithm First, we modified the Spanish OpenThesaurus and created our List of Senses. Instead of having a target word with different senses, we kept only the list of senses and included the target word in each one.

Then, for each of the words we included their frequency on the Web using a large search engine frequency index. As a result we had a set of lists of synonyms with their word frequencies, where each list corresponds to one sense. 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')*.

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. This token is lemmatized and it is included in the Synonyms List as a target word only if it appears in our List of Senses, filtering proper names and stop words (*and, of, at, etc.*). The other four tokens are the context of the target word, as well as its frequency in the corpus and the number of times that the contexts appear having different target words. For example, below we give a context for the target word *noche ('night')*:

era una **noche** *oscura de* (it was a dark night of)

Third, we define complexity of a word taking into account the frequency of the words in the Web, because previous studies have shown that less frequent words were found to be more challenging for people without reading impairments (Inhoff and Rayner, 1986; Just and Carpenter, 1980; Raney and Rayner, 1995; Rayner and Duffy, 1986; Rayner and Raney, 1996) as well as for readers with learning disabilities such as dyslexia (Hyönä and Olson, 1995; Rello et al., 2013; Rüsseler et al., 2003). Next, to determine the word complexity we use the relative frequency of the synonyms with the same sense in the List of Senses. If a word is ten or more times less frequent than one or more of its synonyms is considered a complex word. We used ten times as a *popularity threshold* because worked well in practice (31% of the words have simpler synonyms in this way), but this is an adjustable parameter of the algorithm (complexity depends on different factors such as age or education).

Finally, for each complex word and the contexts where it appears, we select as simpler synonyms the three most frequent ones that belong to the sense that appears most frequently for the n-gram corresponding to that (word,context) pair. That is, to disambiguate the sense, CASSA uses the context where the target word appears in the n-grams corpus.

4. CASSAurus

4.1. Description

The CASSAurus resource is composed of two files: one that includes the context of the target synonym for disambiguation and a second one that gives the most common senses for the case where the context is not found in the first file. As shown in previous work (see (Specia et al., 2012), this baseline, based in the most frequent senses, is very hard to beat by other approaches to lexical simplification, due to the dominance of the most frequent senses in written text.

CASSAurus: This file includes 9,345 lists of simpler synonyms for 41,106 complex words in Spanish. Depending on the sense of each complex word, a (potentially different) list of synonyms is provided. Each line is composed by six elements: **frequency, complex word, context, simpler synonyms** and **lemma**. Following there is an example of a line and its decoding.

```
48, fortuna, golpe de que le,
[gracia,suerte,dicha], fortuna
```

Frequency: It is the first element of the lines in the resource (*e.g. 48*), and it corresponds with the **frequency** of the **complex** inflected word (*fortuna, 'fortune'*, second element) in its **context** (*e.g. golpe de * que le, 'stroke of that his'*, third element), where * indicates the location of the target word in the context.

Complex word: The inflected complex word in the same form it is found in the n-grams.

Context: The context of the inflected complex word, the first, second, fourth and fifth elements of the n-gram.

Simpler synonyms: The three most frequent synonyms of the disambiguated senses of the complex word, (*e.g. gracia, suerte, dicha, 'grace, luck, happiness'*),

Lemma: The lemma of the complex word, (*e.g. fortune, 'fortune'*).

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

124	fortuna	de probar en el	[sino, estrella, destino]	fortuna
	fortune	**to try ∼ in*	*[destiny, fate, luck]*	*fortune*
48	fortuna	la mala de cruzarse	[sino, estrella, destino]	fortuna
	fortune	**the mis∼ of crossing*	*[destiny, fate, luck]*	*fortune*
49	fortuna	desigualdad de en el	[recursos, medios, capital]	fortuna
	fortune	**inequality of ∼ in the*	*[resources, means, capital]*	*fortune*
64	fortunas	acumularon las de las	[recursos, medios, capital]	fortuna
	fortunes	**accumulated the ∼ of the*	*[resources, means, capital]*	*fortune*
56	fortunas	de las amasadas por	[recursos, medios, capital]	fortuna
	fortunes	**of the ∼ amassed by*	*[resources, means, capital]*	*fortune*
52	fortuna	dueño de y de	[capital, dinero, patrimonio]	fortuna
	fortune	**owner of ∼ and of*	*[capital, money, heritage]*	*fortune*
63	fortuna	fue una para el	[gracia, favor, suerte]	fortuna
	fortune	**it was a ∼ for him*	*[grace, favor, luck]*	*fortune*
60	fortuna	golpe de que le	[gracia, suerte, dicha]	fortuna
	fortune	**a stroke of ∼ that*	*[grace, luck, bliss]*	*fortune*
46	fortuna	la inmensa de haber	[regalo, paz, suerte]	fortuna
	fortune	**extremely ∼ to have had*	*[gift, peace, luck]*	*fortune*

Figure 1: Example extracted from CASSAurus.

952414	fortuna	[recursos, medios, capital]	fortuna
	fortune	*[resources, means, capital]*	*fortune*
100797	fortunas	[recursos, medios, capital]	fortuna
	fortunes	*[resources, means, capital]*	*fortune*

Figure 2: Example extracted from the Frequency baseline.

In Figure 1 we find different simpler synonyms for the word *fortuna, 'fortune'*. For instance, in *a stroke of fortune (golpe de fortuna que le)*, *fortune* means *luck (gracia,suerte,dicha)* while in *owner of fortunes and (dueño fortuna de y de)*, *fortune* means *money (capital money, property)*.

Frequency baseline: This file includes 7,562 lists of simpler synonyms for 40,825 complex words in Spanish. In this case the different senses of the complex words are not taken into consideration and then can be used independently of the context if needed.

Each line is composed by six elements: **frequency, complex word, simpler synonyms** and **lemma**:

```
952414
fortuna
[recursos, medios, capital]
fortuna
```

Frequency: The absolute frequency of the complex word in the Web (large sample from 2013).

Complex word: The inflected complex word in the same form it is found in the n-grams.

Simpler synonyms: The three most frequent synonyms taking within all the senses where the complex word appear in SpOT.

Lemma: The lemma of the complex word.

Resource	**CASSAurus**	**Baseline**
Complex Words	41,106	40,825
Contexts	1,817,069	–
Complex Lemmas	9,928	9,732
Simpler Synonyms	9,345	7,562

Table 1: Number of complex inflected words, complex lemmas and simpler synonyms for each file.

Table 1 gives the number of complex inflected words, complex lemmas and simpler synonyms of both CASSAurus files.

5. Use Case and Evaluation

CASSAurus was successfully integrated in an ebook reader for iOS tailored to Spanish speaking people with dyslexia called Dyswebxia (Rello, 2014). As evaluation we analyze the coverage of CASSAurus and the baseline taking into consideration the compulsory reading of secondary and high school in Spain.

To check the coverage of the synonyms resource, we create a corpus made of 196 classic literature books from the 15th century to the 20th century. We included the books that are compulsory readings in secondary and high school in Spain.

The corpus is composed by 16,495,885 tokens and 5,886,366 lexical words (without stop words, proper names and punctuation marks).

Using this corpus we calculated three coverages: (1) the coverage of SpOT, (2) the coverage of the Frequency baseline, and (3) the coverage of CASSAurus.

First, the coverage of all the synonyms present in SpOT – including simple and complex synonyms– is 88.34%, that is, words in the corpus which are covered by the thesaurus. This is the maximum that any simplification algorithm that uses SpOT can achieve.

The percentage of complex words found by CASSA is 27.2%. From those 24.1% are covered by the baseline. On the other hand, of the 28.0% complex contexts found, 2.7% are covered by CASSAurus. These are absolutes percentages, while the relative percentages are 88.6% for the baseline case and 9.6% for the disambiguated case. Hence, together, our resources cover a large majority of the simplification cases. In (Baeza-Yates et al., 2015) we give results for smaller values of the popularity threshold, which makes more words complex but increase the coverage.

This application was also evaluated regarding synonymy and simplicity through a user study that included 32 participants with dyslexia and 38 without dyslexia (Rello and Baeza-Yates, 2014). In this study CASSA and CASSAurus were statistically better than the baseline for both measures. Indeed, for the control group without dyslexia, they found CASSA 29% better on average for synonymy and 23% better on average for simplicity.

6. Concluding Remarks

Open Thesaurus is available in nine languages and Google Books Ngrams in seven. However, there are only two languages in both sets: Spanish and German. Hence our method can be easily extended to German as there are German lemmatizers freely available (Perera and Witte, 2005). Other languages are also possible with proprietary language resources, either a thesaurus or an n-gram corpus. For example, for English just a thesaurus, a lemmatizer and the n-grams corpus would be needed. This resource has been integrated in a plug-in for Chrome that presents definitions and simpler synonyms on demand for the selected web text (Rello et al., 2015).

Further work is needed to expand the coverage of our resource as well as to study other metrics for word complexity in the context of word simplification for people with dyslexia. We will consider the orthographic and phonetic similarity of words, because those language features makes words more difficult to recognize for people with and also without dyslexia (Mitkov et al., 2009).

Acknowledgements

The research of the second author was partially funded by the Catalonian FI scholarship program and for the first two authors by Grant TIN2012-38741 from the Ministry of Economy and Competitiveness of Spain. We also thank the anonymous reviewers for their valuable comments on the paper.

7. Bibliographical References

Baeza-Yates, R., Rello, L., and Dembowski, J. (2015). CASSA: A simple context aware synonym simplification method. In *NAACL 2015*, Boulder, USA.

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

Bott, S. and Saggion, H. (2012). Text simplification tools for Spanish. In *Proc. LREC'12*, Istanbul, Turkey, May. ELRA.

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

Coster, W. and Kauchak, D. (2011). Simple english wikipedia: A new text simplification task. In *ACL (Short Papers)*, pages 665–669.

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

Inhoff, A. W. and Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics*, 40(6):431–439.

Just, M. A. and Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological review*, 87:329–354.

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

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., The Google Books Team, Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., and Nowak, M. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

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

Mitkov, R., Ha, L. A., Varga, A., and Rello, L. (2009). Semantic similarity of distractors in multiple-choice tests: extrinsic evaluation. In *Proc. EACL Workshop GeMS '09*, pages 49–56.

Padró, L., Collado, M., Reese, S., Lloberes, M., and Castellón, I. (2010). Freeling 2.1: Five years of open-source language processing tools. In *Proc. LREC'10*, Valletta, Malta, May.

Perera, P. and Witte, R. (2005). A self-learning context-aware lemmatizer for German. In *Proc. of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pages 636–643, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics (ACL).

Raney, G. E. and Rayner, K. (1995). Word frequency effects and eye movements during two readings of a text. *Canadian Journal of Experimental Psychology*, 49(2):151.

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

Rayner, K. and Raney, G. E. (1996). Eye movement control in reading and visual search: Effects of word frequency. *Psychonomic Bulletin & Review*, 3(2):245–248.

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

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

Rello, L., Carlini, R., Baeza-Yates, R., and Bigham, J. (2015). A plug-in to aid online reading in spanish. In *Proc. W4A '15*, Florence, Italy. ACM.

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

Rüsseler, J., Probst, S., Johannes, S., and Münte, T. F. (2003). Recognition memory for high-and low-frequency words in adult normal and dyslexic readers: an event-related brain potential study. *Journal of clinical and experimental neuropsychology*, 25(6):815–829.

Saggion, H., Bott, S., and Rello, L. (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.

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

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

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