An Adaptable Lexical Simplification Architecture for Major Ibero-Romance Languages

Daniel Ferrés, Horacio Saggion
Large Scale Text Understanding Systems Lab
TALN - DTIC
Universitat Pompeu Fabra
08018 Barcelona, Spain
daniel.ferres@upf.edu
horacio.saggion@upf.edu

Xavier Gómez Guinovart
TALG Group
Universidade de Vigo
E-36310 Vigo, Spain
xgg@uvigo.es

Abstract

Lexical Simplification is the task of reducing the lexical complexity of textual documents by replacing difficult words with easier to read (or understand) expressions while preserving the original meaning. The development of robust pipelined multilingual architectures able to adapt to new languages is of paramount importance in lexical simplification. This paper describes and evaluates a modular hybrid linguistic-statistical Lexical Simplifier that deals with the four major Ibero-Romance Languages: Spanish, Portuguese, Catalan, and Galician. The architecture of the system is the same for the four languages addressed, only the language resources used during simplification are language specific.

1 Introduction

Text Simplification (Saggion, 2017) should facilitate the adaptation of available and future textual material making texts more accessible. Although there are many characteristics which can be modified in order to make information more readable or understandable, automatic text simplification has usually be concerned with two different tasks: lexical simplification and syntactic simplification. Lexical Simplification, the focus of the present work, aims at replacing difficult words with easier synonyms, while preserving the meaning of the original text. Lexical simplifiers can be potentially useful for different target groups with specific accessibility issues ranging from children, second language (L2) learners (Petersen and Ostendorf, 2007), low literacy readers (Aluísio and Gasperin, 2010), people with cognitive disabilities (Saggion et al., 2015), among others. Moreover, different natural languages have been object of automatic text simplification studies including English (Biran et al., 2011; Ferrés et al., 2016), Spanish (Bott et al., 2012), and Portuguese (Specia, 2010) just to name a few. To the best of our knowledge no previous research has addressed the issue of language adaptation of lexical simplification systems. We here present an approach to Lexical Simplification in the four major Ibero-Romance Languages: Catalan (ca), Galician (gl), Portuguese (pt), and Spanish (es) using the same underlying architecture. The Ibero-Romance languages (also known as Iberian Languages) are the ones that developed on the Iberian Peninsula and in southern France. These languages, that share high lexical similarities, are currently spoken by more than 750 million people around the world. The research and development of Textual Simplification systems for languages with high lexical similarities among them, such as Ibero-Romance languages with about and above 85% of lexical similarities (see Table 1), has the advantage of producing processing and lexical resources that can be easily adapted semi-automatically.

	ca	es	pt
ca	-	85%	85%
es	85%	-	89%
pt	85%	89%	-

Table 1: Lexical similarity between the 3 major Ibero-Romance languages according to Ethnologue[1]. Data for Galician were not available.

The lexical simplifier presented in this paper has been developed following current robust, corpus-based approaches (Biran et al., 2011; Bott et al., 2012; Ferrés et al., 2016) combined with a hybrid Morphological Generator that uses both a wide-coverage lexicon freely available and a

[1] www.ethnologue.com

Decision-Trees based algorithm, and an easy to adapt rule-based context re-writting module. The availability of such a robust multilingual generator is key for inflecting words, which in the rich morphological languages addressed is extremely important.

The contributions of this paper can be summarized as follows:

- The first multilingual lexical simplification architecture.[2]
- The first system to address lexical simplification for Catalan and Galician.
- A well-established evaluation of the adequacy and simplicity of the simplifications based on native speakers' assessment.

The rest of the paper is organized as follows: in Section 2 we describe the related work. The architecture of the lexical simplifier and its evaluation are described in Sections 3 and 4. After a detailed discussion in Section 5, the paper is concluded at Section 6 with some conclusions and further work.

2 Related Work

Work on Lexical Simplification for English began in the PSET project (Devlin and Tait, 1998). The authors used WordNet to identify synonyms and calculated their relative difficulty using Kucera-Francis frequencies in the Oxford Psycholinguistic Database. De Belder and Moens (De Belder and Moens, 2010) combined this methodology with a latent words language model which modeled both language in terms of word sequences and the contextual meaning of words. Wikipedia has also been used in lexical simplification studies. Biran et al. (Biran et al., 2011) used word frequencies in English Wikipedia and Simple English Wikipedia (SEW) to calculate their difficulty while Yatskar et al. (Yatskar et al., 2010) used SEW edit histories to identify the simplify operations. More recently, (Glavaš and Štajner, 2015) proposed a simplification method based on current distributional lexical semantics approaches for languages for which lexical resources are scarce. The same line of research is followed by (Paetzold, 2016) who additionally includes a retrofitting mechanism to better distinguish between synonyms and antonyms (Faruqui et al., 2015).

[2]Not based on parallel or comparable corpora.

Regarding Lexical Simplification in Ibero-Romance languages, there are five systems reported in the literature for Spanish and Portuguese:

- LexSiS (Bott et al., 2012) is a lexical simplifier for Spanish. LexSiS uses a word vector model derived from a 8M word corpus of Spanish text extracted from the Web for Word Sense Disambiguation with the Spanish OpenThesaurus as a source for finding candidate synonyms of complex words. Lexical realization is carried out using a dictionary and hand-crafted rules.

- PorSimples is a lexical simplifier for Portuguese (Aluísio and Gasperin, 2010). PorSimples uses the Unitex-PB dictionary and the MXPOST POS tagger for lemmatization and PoS tagging. Complex word detection is performed with a dictionary of simple words. The TeP 2.0 thesaurus and PAPEL lexical ontology were used to find a set of synonyms without the use of Word Sense Disambiguation. The lexical simplicity order of synonyms is determined with word frequencies obtained through Google API.

- Specia (2010) used the Moses toolkit for phrase-based Statistical Machine Translation (SMT) and a corpus of about 4,483 sentences (3,383 for training, 500 for tuning, and 500 for test) in order to learn how to simplify sentences in Brazilian Portuguese.

- Stajner (2014) also used phrase-based SMT for lexical simplification in Spanish. She built language models derived from the Spanish Europarl corpus and used 700 sentence pairs for training, 100 sentence pairs for development, and three test sets for testing (of 50, 50, and 100 sentences).

- Baeza-Yates et al. (2015) presented CASSA a lexical simplifier for Spanish. CASSA uses the Google Books Ngram Corpus to find the frequency of target words and its contexts and uses this information for disambiguation. The Spanish OpenThesaurus (version 2) is used to obtain synonyms and web frequencies are used for disambiguation and lexical simplicity. No morphological realization is performed in this system.

3 Lexical Simplifier

The Lexical Simplification architecture allows to simplify words (common nouns, verbs, adjectives, and adverbs) in context. The architecture follows an approach similar to the YATS lexical simplifier (Ferrés et al., 2016). The simplifier has the following phases (executed sequentially): (i) Document Analysis, (ii) Complex Words Detection, (iii) WSD, (iv) Synonyms Ranking, and (v) Language Realization (see the architecture of the system in Figure 1). The Document Analysis phase uses the FreeLing 4.0[3] system (Padró and Stanilovsky, 2012) to perform tokenization, sentence splitting, part-of-speech (PoS) tagging, lemmatization, and Named Entity Recognition.

3.1 Complex Word Detection

The Complex Word Detection (CWD) phase is carried out to identify target words to be substituted. The procedure identifies a word as complex when the frequency count of word forms or lemmas in a given frequency list extracted from a corpus is below a certain threshold value (i.e. w is complex if $w_{frequency} \leq theshold$).

The frequency lists that can be used separately by this phase are: 1) the Wikipedia forms counts, 2) the Wikipedia extracted lemmas with associated PoS tags[4] (only common nouns, verbs, adjectives and adverbs are extracted), and 3) the OpenSubtitles 2016 words full frequency list[5].

lang	Wikipedia		OpenSubtitles2016
	#lemmas&PoS	#forms	#forms
ca	2,571,667	1,306,344	65,687
es	6,844,698	2,645,049	1,882,198
gl	1,130,788	630,318	73,808
pt	4,829,021	1,975,973	477,456

Table 2: Statistics of the frequency lists.

For example, Table 3 shows how commonly used noun lemmas such as *hand* (having the forms "mà" in Catalan (ca) ,"mano" in Spanish (es) ,"man" in Galician (gl),"mão" in Potuguese (pt)) and *lawyer* ("advocat" (ca), "abogado" (es), "avogado" (gl), "advogado" (pt)) have much more counts in Wikipedia than less common lemmas such as *democracy* ("democràcia" (ca), "democracia" (es,gl,pt)) and *gastronomy* ("gastronomia" (ca), "gastronomía" (es,gl,pt)).

lang	#counts			
	hand	lawyer	democracy	gastronomy
ca	24,936	4,994	3,055	1,163
es	60,271	20,432	11,485	6,850
gl	4,878	2,003	1,084	457
pt	29,556	12,267	4,443	1,172

Table 3: Example of some word lemmas counts in Wikipedia.

In order to obtain a threshold for each language for the Complex Word Detection phase the following procedure has been applied: 1) A set of pairs <complex word, simpler synonym> (such as <novelist,writer> or <tenor,singer>) has been extracted from the LexSiS Gold (Bott et al., 2012) (Spanish) and the PorSimples FSP (Aluísio et al., 2008) (Portuguese) corpora: 102 pairs have been extracted from the LexSiS Gold corpora and 279 from the PorSimples FSP. 2) The 102 pairs in Spanish from LexSiS Gold have been automatically translated to Catalan and manually revised. In order to create a set of 100 pairs from Galician some pairs have been extracted from the 279 pairs in Portuguese and some new pairs have been manually added. 3) A measure of complex word detection accuracy that involves the use of both the complex word and the simpler synonym for each pair has been created. This measure has been called *accuracy complexS* and calculates the ratio of pairs in which its complex word component has been detected as complex word according to the threshold and at the same time the simpler synonym component has been detected as simple word according to the threshold. On the other hand, another measure called *accuracy complex* has been defined as the ratio of pairs in which its complex word component has been detected as complex word according to the threshold. 4) The measure *accuracy complexS* has been used to tune the thresholds of each language: a) a set of thresholds that have been found empirically to maximize the *accuracy complexS* is obtained by automatic testing through intervals of thresholds (the frequency list is divided in a set of 50,000 intervals of thresholds ranging from 0 to the maximum frequency in the corpus) , b) from the selected set of thresholds another subset is obtained by selecting the ones with the best *accuracy complex* measure results, c) finally the higher threshold from the last subset is chosen to be the complex word threshold

[3] http://nlp.cs.upc.edu/freeling

[4] The tools to extract the lemmas and PoS tags from Wikipedia are explained in the Section 3.2.

[5] https://github.com/hermitdave/FrequencyWords

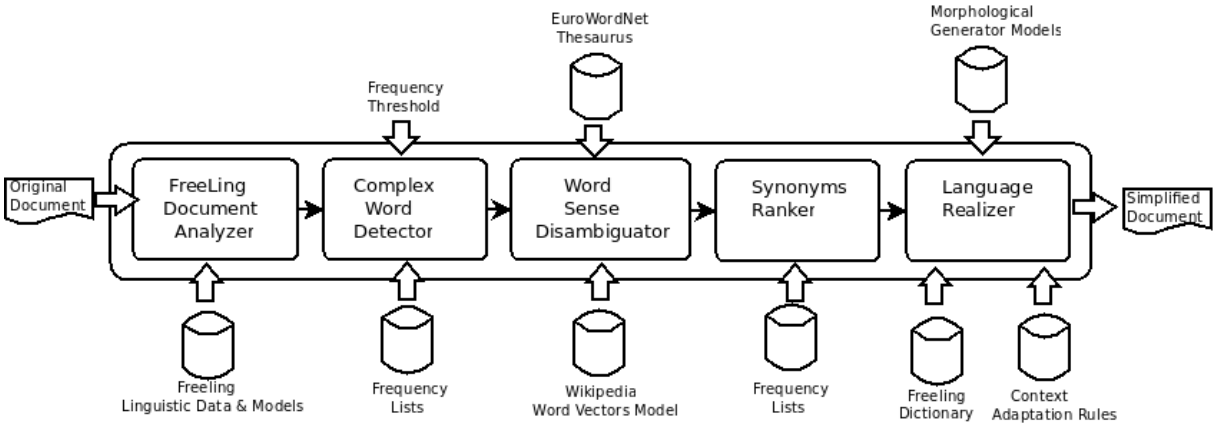

Figure 1: System Architecture.

for the language.

The results of applying this tuning procedure using the 3 frequency lists over the 4 set pairs in each languages are shown in Table 4. The best thresholds for both *accuracy complexS* and *accuracy complex* are obtained by the Wikipedia forms frequency lists (for *ca*, *es*, and *gl*) and with the OpenSubtitles 2016 frequency list for *pt*.

lang	frequency list	accuracy	
		complex	complexS
ca	cawiki (lemma)	0.7500	0.5200
	cawiki (form)	**0.8000**	**0.6200**
	opensubtitles	0.7100	0.5300
es	eswiki (lemma)	0.7524	0.5346
	eswiki (form)	**0.8613**	**0.6237**
	opensubtitles	0.8613	0.5940
gl	glwiki (lemma)	0.5154	0.2371
	glwiki (form)	**0.6082**	**0.4845**
	opensubtitles	0.2886	0.2164
pt	ptwiki (lemma)	0.7562	0.2258
	ptwiki (form)	0.7132	0.4767
	opensubtitles	**0.8530**	**0.6308**

Table 4: Complex word tunning: best accuracies for threshold computation.

3.2 Word Sense Disambiguation

The WSD algorithm used is based on the Vector Space Model (Turney and Pantel, 2010) approach for lexical semantics which has been previously used in Lexical Simplification (Biran et al., 2011; Bott et al., 2012). The set of language-dependent thesaurus used for WSD was extracted from FreeLing 4.0 data which is derived from Multilingual Central Repository (MCR) 3.0[6] (release 2012). Each thesaurus contains a set of synonyms and its associated set of senses with related synonyms (see the number of entries and senses of each language thesaurus in Table 5).

The WSD algorithm uses a word vectors model derived from a large text collection from which a word vector for each word in the thesaurus is created by collecting co-occurring word lemmas of the word in N-window contexts (only nouns, verbs, adjectives, and adverbs). Then, a common vector is computed for each of the word senses of a given target word (lemma and PoS) by adding the vectors of all words in each sense. When a complex word is detected, the WSD algorithm computes the cosine distance between the context vector computed from the words of the complex word context (at sentence level) and the word vectors of each sense from the model. The word sense selected is the one with the lowest cosine distance between its word vector in the model and the context vector of the complex word in the sentence or document to simplify.

lang	EuroWordNet		Wikipedia	
	#entries	#senses	#docs.	#words
ca	46,555	64,095	450,885	124.5M
es	36,571	50,397	1,061,535	349M
gl	23,058	26,009	221,422	36.2M
pt	35,635	45,737	956,553	203M

Table 5: Statistics of the EuroWordNet thesaurus and the Wikipedia collections processed.

The Catalan, Galician, Portuguese and Spanish Wikipedia dumps were used to extract the word vectors model. The plain text of the documents was extracted using the WikiExtractor[7] tool (see in Table 5 the number of documents and words ex-

[6] http://adimen.si.ehu.es/web/MCR/

[7] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

tracted from each Wikipedia dump). The FreeLing 3.1 NLP tool was used to extract the lemmas and PoS tags of each word, from a 11-word window (5 content words to each side of the target word).

3.3 Synonyms Ranking

The Synonyms Ranking phase ranks synonyms by their lexical simplicity and finds the simplest and most appropriate synonym word for the given context (Specia et al., 2012). The simplicity measure implemented is the word form (or lemma) frequency (i.e. more frequent is simpler) (Saggion, 2017). The frequency lists that can be used are the ones described in the CWD phase.

3.4 Language Realization

The Language Realization phase generates the correct inflected forms of the final selected synonyms lemmas and the other lemmas of the context. It has two phases: i) a context-independent Morphological Generator and ii) a rule-based Context Adaptator. The Morphological Generation system combines lexicon-based generation and predictions from Decision-Trees (see (Ferrés et al., 2017) for a more detailed description of this system). The lexicons used are the FreeLing[8] (Padró and Stanilovsky, 2012) morphological dictionaries for *ca*,*es*,*gl* and *pt* (see in Table 6 more details about these dictionaries). The Decision Trees algorithm used to predict the inflected form is the J48 algorithm from the WEKA[9] data mining tool. This algorithm is only used when the lexicon has no inflection for a pair <lemma,PoS>. The J48 model can predict the sequence of edit operations that can transform an unseen pair <lemma,PoS> to an inflected form.

	Freeling Data		Training Data	
lang	#lemmas	#forms	corpus	#tokens
ca	66,168	642,437	CoNLL09	390,302
es	70,150	669,216	CoNLL09	427,442
gl	45,674	570,912	UD_Galician	79,329
pt	94,444	1,214,090	Bosque 8.0	232,600

Table 6: Morphological Generation training data statistics.

The J48 training algorithm uses morphological and lemma based features including the Levenshtein edit distance between lemmas and word forms to create a model for each lexical category. The learning datasets used were: the CoNLL2009 shared Task[10] Catalan and Spanish training datasets, the Bosque 8.0 corpus tagged with EAGLES tagset[11], and the Galician UD treebank[12] based on the CTG corpus[13].

The Morphological Generator was evaluated independently using the following corpora to test: CoNLL2009 Shared task evaluation dataset for Catalan (53,016 tokens) and Spanish (50,635 tokens), the Galician UD test set for Galician (29,748 tokens) and the Portuguese UD test set for Portuguese (5,499 tokens)[14]. The results (see Table 8) show that the Morphological Generator configuration that uses both FreeLing and J48 achieves high performance with accuracies over or close to 99% in almost all cases with the exception of the verbs in Spanish and Portuguese which obtained a 95.77% and 95.49% of accuracy respectively and the adjectives in Portuguese with a 94.34%.

The Context Adaptation phase generates the correct inflected forms of the lemmas in the context of the substituted complex word in case that it is needed an adaptation due to the morphological features of the substitute synonym. In the Ibero-Romance languages treated there are 3 cases of this kind (not all these cases are treated yet by our system):

1) adaptation of articles, pronouns and prepositions due to an ortographic variation of the substituted synonym (only in *ca* and *gl* languages): e.g. apostrophize determiners in *ca* ("el marit/l'home" (husband/man)), pronominal accusative changes in *gl* ("relatouno / díxoo" ("relatou+no" – (s)he related it / "díxo+o" – (s)he said it)).

2) adaptation of determiners (and pronouns) due to a morphological change of noun gender: as an example in the 4 languages the word "sovereignty" ("sobirania" (*ca*), "soberanía" (*es,gl*) "soberania" (*pt*) can be substituted for its synonym "power" ("poder" (*ca,es,gl,pt*)) but if a determiner precedes the word then it has to change its gender ("la" to "el" (*ca,es*) , "a" to "o" (*gl,pt*)).

3) adaptation of verbs (and adjectives) due to the need of gender concordance: e.g the verb "administer" ("administrat/administrada" (*ca*), "ad-

[8] http://nlp.lsi.upc.edu/freeling/

[9] http://www.cs.waikato.ac.nz/~ml/weka/

[10] http://ufal.mff.cuni.cz/conll2009-st/

[11] http://www.linguateca.pt/floresta/corpus.html

[12] http://universaldependencies.org

[13] http://sli.uvigo.gal/CTG/

[14] Both Galician and Portuguese UD datasets taken from http://hdl.handle.net/11234/1-1983

	system	Simplicity scale					Adequacy scale				
		1	2	3	4	5	1	2	3	4	5
ca	MFS	1.35%	15.59%	15.25%	34.91%	32.88%	11.86%	16.27%	8.13%	22.71%	41.01%
	simplifier	2.37%	17.62%	14.57%	27.11%	38.30%	7.79%	21.01%	9.83%	21.01%	40.33%
es	MFS	6.77%	14.57%	15.93%	25.76%	36.94%	15.93%	14.57%	9.49%	17.96%	42.03%
	simplifier	8.13 %	11.52%	23.05%	27.11%	30.16%	18.64%	15.93%	5.76%	15.93%	43.72%
gl	MFS	13.94%	15.30%	23.46%	25.51%	21.76%	16.94%	16.94%	14.23%	31.86%	20.00%
	simplifier	17.62%	16.94%	21.01%	26.44%	17.96%	21.35%	20.33%	10.84%	26.77%	20.67%
pt	MFS	5.6%	17.2%	25.42%	27.79%	23.38%	12.54%	15.93%	12.88%	30.84%	27.79%
	simplifier	8.84%	14.28%	24.48%	31.97%	20.40%	14.96%	18.70%	13.26%	27.55%	25.51%

Table 7: Evaluation of simplicity and adequacy over a subset of 50 randomly selected sentences from the Wikipedia and simplified by the lexical simplifier and the MFS baseline.

lang.	Algorithm	Noun	Verb	Adj	Adv
ca	FreeLing (C)	72.19	96.63	77.63	77.48
	J48	**99.56**	98.42	98.76	**100**
	FreeLing+J48	99.53	**99.39**	**99.47**	**100**
es	FreeLing (C)	72.60	95.03	76.21	72.89
	J48	99.80	94.32	99.24	98.51
	FreeLing+J48	**99.84**	**95.77**	**99.44**	**98.57**
gl	FreeLing (C)	90.31	97.95	94.46	88.82
	J48	99.70	96.95	99.39	97.76
	FreeLing+J48	**99.97**	**99.96**	**99.91**	**98.10**
pt	FreeLing (C)	88.31	91.12	60.00	82.60
	J48	**98.75**	95.21	93.47	**99.56**
	FreeLing+J48	**98.75**	**95.49**	**94.34**	**99.56**

Table 8: Results of the evaluation in accuracy (%) of the Morphological Generator configurations. Note that in the Freeling configuration the accuracy means coverage (C) because the lexicon cannot predict unseen <lemma,PoS> pairs.

ministrado/administrada" (*es,gl,pt*)) in the sentence "the medicine was administered to the patient", has to be conjugated in concordance with the synonym that substitutes the word "medicine".

4 Evaluation

The evaluation has been realized using a lexical simplifier system with the best parameters obtained in the complex word detection tuning phase and these frequency lists have been also used in the Synonyms Ranking phase. We performed manual evaluation of the simplifier relying on 7 different proficient human judges for each language evaluated,[15] who assessed our system with respect to adequacy and simplicity. The evaluation dataset was created from a set of sentences of the Wikipedia which had at least one non-monosemous complex word and 2 synonyms and less than 26 tokens (Named Entities included as tokens). Then this dataset was simplified and the sentences that had only one lexical simplification were selected[16]. A set of 50 sentences with more than 18 tokens was randomly selected from this set of lexically simplified sentences. The participants were presented with the source sentence from the Wikipedia followed by either a sentence simplified by the full system or a by a baseline version of the system that uses the most frequent synonym (MFS) of all senses as WSD. Simplicity was measured using a five point rating scale that indicates how much simpler was the simplified sentence w.r.t the original (high numbers indicate simpler). Adequacy was also measured using a five point rating scale that indicates if the simplified sentences keeps the same meaning (high numbers indicate more adequacy). Table 7 shows the evaluation results in simplicity and adequacy.

5 Discussion

The Complex Word Detection phase presented uses frequency thresholding over frequency lists extracted from corpora. The motivation of using such methodology is to have a generic method to detect complex words for average adult people that can be easily adaptable to several languages and requiring only textual corpora. Obviously this method has some problems: 1) the extraction of frequencies from huge corpora may rely on sets of documents with unbalanced, over-represented or under-represented domains that could suppose to generate high frequencies for real complex words or low frequencies for simple words, 2) the threshold tunning process is sensible to the semantic complexity level of the list word pairs used, and this could led to generate complex word detections useful only for certain groups of people.

In order to test if some simple words could have low frequencies in the corpora (Spanish Wikipedia) with respect to the threshold used for

[15] Graduates and university undergraduate students. None of them developed the simplifier.

[16] This step was performed to avoid interference of multiple simplifications.

the Wikipedia forms frequency list we used a list of subjective estimation of Age of Acquisition (AoA) words in Spanish (Alonso et al., 2015). The average AoA score for each word was based on 50 individual responses on a scale from 1 to 11 (indicating the age that this word was acquired). A set of 2,307 words estimated to be acquired at an age below 6 years (so supposing that these words have to be very simple) has been used for this test. Using the best threshold obtained in tuning procedure to estimate complex words has resulted in that 829 of these words (35.87%) were correctly not detected as complex words but 1,455 (62.99%) were incorrectly detected as complex words and 25 were not found (0.37%). This means that at least in Spanish (of the 4 languages used the one which has more documents in the Wikipedia) some words that are really simple such as "sopa" (soup), "to fish" (pescar), and "veinticinco" (twenty-five) among others have been detected as complex words.

In order to solve these problems, besides of increasing and balancing the corpora, the modularity of the resource allow these kind of solutions: 1) both the threshold and the frequency list files can be edited manually and change the frequency of those words, 2) generating manually or semi-automatically frequency lists of complex words or simple words that can be generic or adapted to specific target groups, and 3) combine both corpus-based frequency lists and manually generated. Previous competitive approaches to complex word identification are many times based on word frequency thresholding as we implement here (see (Wrobel, 2016) who obtained the best F-score in the recent Complex Word Identification task (Paetzold and Specia, 2016))

The results obtained through subjective manual assessment by native evaluators show that both the MFS baseline and the full simplifier obtain more than 50% of positive results (scores 4 and 5 in the five-point rating scale) in simplicity and adequacy a for *ca*,*es*, and *pt* and more than 40% for *gl*. These results mean that both the system and the MFS baseline can be useful for lexical simplification but the large percentage of negative results in adequacy (scores 1 and 2 in the five-point rating scales) indicates that more research is needed to avoid errors of meaning preservation. The reported errors in adequacy and the fact that the simplifier generally does not perform better than the MFS baseline point that the WSD algorithm and/or its resources need to be improved.

In general Lexical Simplification systems do not deal with Morphological Generation, for example CASSA (Baeza-Yates et al., 2015) has not morphological realization component, LexSiS morphological realization (Bott et al., 2012) is limited to a dictionary and set of handcrafted rules. Simplification systems based on machine translation (Specia, 2010; Stajner, 2014) generate words based on parallel/comparable original and simplified datasets being therefor limited in coverage (e.g. words not observed in the dataset will not be properly generated). Our approach instead is robust in terms of coverage and easily adapted to new languages with similar characteristics (e.g. Italian, French). It is worth notice that both approaches we present here: the baseline and our simplifier both take advantage of the morphological realization component. Moreover, the only module that is not used by the baseline is the Word Sense Disambiguator.

6 Conclusion

Automatic Lexical Simplification is a task that requires very complex and advanced resources in both Natural Language Processing and Natural Language Generation fields. In this paper we have presented a modular automatic Lexical Simplifier system that can deal with the four major Ibero-Romance Languages: Spanish, Portuguese, Catalan, and Galician. The experiments presented in this paper show that the corpus-based approaches tried, despite of being useful for generic prediction, are not yet sufficient to deal with the complexities of the task and manual effort from linguistic experts to create specific resources for the task is needed.

Future research includes: a) experiments with other available datasets, b) use more advanced vector representations (e.g. embeddings), c) update the thesaurus data of MCR 3.0 from release 2012 to release 2016 and apply some manual or automatic revision to prune or mark loosely related synonyms, d) experiments with the CHILDES corpus for complex word detection, and e) porting the system to other similar major Romance languages such as French, Italian and Romanian.

Acknowledgements

This work is (partly) supported by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502) and by the TUNER project (TIN2015-65308-C5-5-R and TIN2015-65308-C5-1-R, MINECO/FEDER, UE).

References

María Angeles Alonso, Angel Fernandez, and Emiliano Díez. 2015. Subjective Age-of-Acquisition norms for 7,039 Spanish Words. *Behavior Research Methods* 47(1):268–274.

Sandra M. Aluísio, Lucia Specia, Thiago A. S. Pardo, Erick G. Maziero, Helena M. Caseli, and Renata P. M. Fortes. 2008. A Corpus Analysis of Simple Account Texts and the Proposal of Simplification Strategies: First Steps Towards Text Simplification Systems. In *Proceedings of the 26th Annual ACM International Conference on Design of Communication*. ACM, SIGDOC '08, pages 15–22.

S.M. Aluísio and C. Gasperin. 2010. Fostering Digital Inclusion and Accessibility: The PorSimples Project for Simplification of Portuguese Texts. In *Proceedings of NAACL HLT 2010 YIWCALA*.

Ricardo A. Baeza-Yates, Luz Rello, and Julia Dembowski. 2015. CASSA: A Context-Aware Synonym Simplification Algorithm. In *NAACL HLT 2015*. pages 1380–1385.

Or Biran, Samuel Brody, and Noémie Elhadad. 2011. Putting It Simply: A Context-aware Approach to Lexical Simplification. In *Proceedings of the ACL 2011*. pages 496–501.

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

Jan De Belder and Marie-Francine Moens. 2010. Text Simplification for Children. In *Proceedings of the SIGIR Workshop on Accessible Search Systems*. pages 19–26.

Siobhan Devlin and John Tait. 1998. The Use of a Psycholinguistic Database in the Simplification of Text for Aphasic Readers. In *Linguistic Databases*. pages 161–173.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting Word Vectors to Semantic Lexicons. In *Proceedings of NAACL 2015*.

Daniel Ferrés, Ahmed AbuRa'ed, and Horacio Saggion. 2017. Spanish Morphological Generation with Wide-Coverage Lexicons and Decision Trees. *Procesamiento del Lenguaje Natural* 58:109–116.

Daniel Ferrés, Montserrat Marimon, Horacio Saggion, and Ahmed AbuRa'ed. 2016. YATS: Yet Another Text Simplifier. In *NLDB*. Springer, volume 9612 of *Lecture Notes in Computer Science*, pages 335–342.

Goran Glavaš and Sanja Štajner. 2015. Simplifying Lexical Simplification: Do We Need Simplified Corpora? In *Proceedings of ACL 2015*. pages 63–68.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of LREC 2012*. ELRA.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 Task 11: Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation*. pages 560–569.

Gustavo Henrique Paetzold. 2016. *Lexical Simplification for Non-Native English Speakers*. Ph.D. thesis, The University of Sheffield.

Sarah E. Petersen and Mari Ostendorf. 2007. Text Simplification for Language Learners: a Corpus Analysis. In *In Proc. of Workshop on Speech and Language Technology for Education*.

Horacio Saggion. 2017. *Automatic Text Simplification*. 32. Morgan & Claypool Publishers, 1 edition.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015. Making it Simplext: Implementation and Evaluation of a Text Simplification System for Spanish. *TACCESS* 6(4):14.

L. Specia, S. K. Jauhar, and R. Mihalcea. 2012. SemEval-2012 Task 1: English Lexical Simplification. In *Proceedings of *SEM 2012*.

Lucia Specia. 2010. Translating from Complex to Simplified Sentences. In *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language*. pages 30–39.

Sanja Stajner. 2014. Translating Sentences from Original to Simplified Spanish. *Procesamiento del lenguaje natural* 53:61–68.

P. D. Turney and P. Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *J. Artif. Int. Res.* 37(1):141–188.

Krzysztof Wrobel. 2016. PLUJAGH at SemEval-2016 Task 11: Simple System for Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation*. pages 953–957.

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 *Proceedings of HLT-NAACL 2010*.