Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3582–3594
Marseille, 20-25 June 2022

ALEXSIS: A Dataset for Lexical Simplification in Spanish

Daniel Ferrés, Horacio Saggion
Large Scale Text Understanding Systems Lab / TALN
Department of Information & Communication Technologies
Universitat Pompeu Fabra, Barcelona (Spain)
{daniel.ferres,horacio.saggion}@upf.edu

Abstract

Lexical Simplification is the process of reducing the lexical complexity of a text by replacing difficult words with easier to read (or understand) expressions while preserving the original information and meaning. In this paper we introduce ALEXSIS, a new dataset for this task, and we use ALEXSIS to benchmark Lexical Simplification systems in Spanish. The paper describes the evaluation of three kind of approaches to Lexical Simplification: a thesaurus-based approach, a single transformers-based approach, and a combination of transformers. We also report state of the art results on a previous Lexical Simplification dataset for Spanish.

Keywords: Lexical Simplification, Text Simplification, Evaluation Dataset

1. Introduction

Text Simplification (Saggion, 2017) is a Natural Language Processing task to transform a text with the aim of reducing its lexical and syntactic complexity while preserving its original meaning. This task can be potentially useful for different audiences, specially children, second language (L2) learners (Petersen and Ostendorf, 2007), low literacy readers (Aluísio and Gasperin, 2010) and people with cognitive disabilities (Saggion et al., 2015) among others. Automatic Text Simplification has usually been 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 information and meaning of the original text. Lexical Simplification systems (Shardlow, 2014; Paetzold and Specia, 2017a) usually have components for: 1) identification of complex terms (Complex Word Identification - CWI), 2) generation of substitution words (Substitute Generation - SG), 3) selection of the substitutes that can fit in the context (Substitute Selection - SS), 4) ranking substitutes by their simplicity (Substitute Ranking - SR), and 5) morphological generation and context adaptation.

The availability of several lexical simplification datasets for the English language has intensified research in this area making system comparison possible. In this paper we present a dataset for benchmarking lexical simplification in Spanish hoping this will help other researchers who are developing technology in this area. Additionally, we present the first set of experiments on this dataset therefore establishing a number of benchmarks for the task. The dataset, systems, and systems' outputs will be made available to the research community for reproducible research[1].

The new dataset contains a set of sentences in which a complex work has been identified and a set of substitutes has been proposed by several annotators. Although the first phase of a LS system is CWI, the systems evaluated in this paper do not perform the CWI task and they start in the SG phase, using the provided complex words to propose and rank simpler substitutes. In this work we do not address the morphological and context adaptation task.

The contributions of this paper are:

- ALEXSIS, a new dataset for benchmarking Lexical Simplification in Spanish, its compilation methodology, and its comparison with other available datasets.
- Experiments with several neural and unsupervised systems for the different phases of LS, thus establishing benchmarks for LS in Spanish.

The rest of the paper is organized as follows: in Section 2 we describe related work on Lexical Simplification including the description of available datasets for several languages. Section 3 presents the ALEXSIS dataset, its compilation procedure and comparison with other datasets. Sections 4 and 5 describe, respectively, the Substitution Generation and Substitution Selection approaches. Section 6 describes the Substitution Ranking approaches. Section 7 describes the evaluation metrics and presents the experimental results. Section 8 discusses the results of the experiments while Section 9 concludes the paper and presents future work.

2. Related Work

Initial work on Lexical Simplification was developed for the English language. (Devlin and Tait, 1998) made use of Wordnet to identify synonyms for target words and word frequencies from the Kucera-Francis psycholinguistic database for synonyms ranking. This initial approach was then followed by corpus-based approaches that used Latent Words Language Models (De Belder and Moens, 2010) or Wikipedia (Biran et al., 2011; Yatskar et al., 2010). (Horn et al., 2014) used a dataset of 137K aligned sentence pairs between English Wikipedia and Simple English Wikipedia to learn simplification rules. (Glavaš and Štajner, 2015) proposed an unsupervised approach for LS based on distributional lexical semantics for languages for which lexical resources are scarce. (Paetzold and Specia, 2017b) presented a LS approach that combines learned

[1] `https://github.com/LaSTUS-TALN-UPF/ALEXSIS_lexsim`

substitutions from the Newsela corpus using neural networks with a retrofitted context-aware word embeddings model. They use a neural ranking model which learns to rank from annotated data.
More recently, (Qiang et al., 2020b) presented LSBert, a LS framework that uses a pretrained representation of BERT (Devlin et al., 2019) for English to propose substitution candidates with high grammatical and semantic similarity to a complex word in a sentence. LSBert uses the masked language model (MLM) of BERT to predict a set of candidate substitution words and their substitution likelihood. They feed BERT with the original sentence concatenated with a copy of the sentence in which the complex word has been masked.
LSBert combines five different strategies for Lexical Simplicity Ranking: BERT prediction order, a BERT-based language model, the PPDB database, word frequency and word semantic similarity with fasttext.
Regarding Lexical Simplification in Spanish, there are seven systems reported in the literature:

- LexSiS (Bott et al., 2012) was the first LS system for Spanish. It uses a word vector model derived from a 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.

- (Štajner, 2014) presented a system that uses phrase-based Statistical-Machine-Translation (PBSMT) for LS in Spanish with language models derived from the Spanish Europarl corpus.

- CASSA (Baeza-Yates et al., 2015) is a LS system 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.

- (Ferrés et al., 2017a), based on LexSiS, proposed a hybrid LS system which employs a combination of knowledge-based resources and corpus-based approaches: the Freeling NLP tool in combination with candidates extracted from a thesaurus, a corpus-based Words Sense Disambiguation approach based on words vector model from Wikipedia, and synonyms ranking based on word frequencies lists in combination with morphological generation and context adaptation.

- (Štajner et al., 2019) carried out a set of experiments with PBSMT Lexical Simplification using nine different training datasets and three language models. Their systems are trained to simplify phrases longer than one word, outperforming the CASSA approach (Baeza-Yates et al., 2015).

- (Alarcon et al., 2021b) developed several approaches for CWI, SG, and SS. In SG they experimented with combinations of lexical resources (PPDB, Babelnet and a Thesaurus database) that achieved the best performance evaluated with the EASIER SG/SS dataset.

- In follow up experiments, (Alarcón et al., 2021a) tested the following approaches for SG and SS: 1) Word2vec, with a model pre-trained on the Spanish Billion Words Corpus (SBWC), 2) Sense2Vec, with a model trained with SBWC, 3) fasttext: with a model pre-trained on Wikipedia with character n-grams of length 5, and 4) BETO[2], a pre-trainted BERT-base model for Spanish (Cañete et al., 2020).

2.1. Available Datasets

Several datasets exist for English including: SemEval2012, LSEVAL, LexMTurk, BENCHLS, NNSEVAL and CEFR-LS. SemEval2012 is a LS corpus used at the shared task on English Lexical Simplification at SemEval-2012, (Specia et al., 2012). The dataset has a total of 2,010 instances and 201 target words (i.e., 10 contexts per complex word). LexMTURK (Horn et al., 2014) is a LS dataset of 500 sentences from Wikipedia with marked complex words and lexically simpler replacements suggested by 50 English speaking annotators (mturk). The BENCHLS dataset (Paetzold and Specia, 2016a) is a union of the LSeval (De Belder and Moens, 2012) and LexMTurk datasets in which spelling and inflection errors were automatically corrected (Paetzold and Specia, 2016a). The NNSEVAL dataset (Paetzold and Specia, 2016b) is a filtered version of the BenchLS adapted to evaluate LS for non-native English speakers (Paetzold and Specia, 2016b). The CEFR-LS dataset (Uchida et al., 2018) for English contains substitutions at different levels of simplicity according to the Common European Framework of Reference for Languages (CEFR).

The HanLS corpus is a dataset for LS for Chinese (Qiang et al., 2021) which was annotated by 6 native speakers. For Japanese, two datasets exist: SNOW E4 (Kajiwara and Yamamoto, 2015) and BCCWJ (Kodaira et al., 2016).
For French, it exists the FrenLys dataset (Rolin et al., 2021), in which its set of synonyms for each complex word was: 1) compiled using automatic generation methods (that were assessed by 3 expert linguists), and 2) ranked by 20 non-expert native speakers. The SIMPLEX-PB 3.0 corpus (Hartmann and Aluísio, 2020) is a version of an existing dataset to evaluate LS for Brazilian Portuguese.

The EASIER[3] dataset (Alarcón, 2021) is a Spanish Dataset which was annotated by a linguist expert in easy-to-read language, and was used for the CWI and SG/SS tasks (Alarcon et al., 2021b). Its quality was verified by two additional experts and a LS user. While the full dataset (EASIER SS/SG) contains about 5,130 instances (Alarcón et al., 2021a) with at least one proposed substitute per complex word, there is a smaller portion of the dataset which

[2] https://huggingface.co/dccuchile/bert-base-spanish-wwm-uncased

[3] https://data.mendeley.com/datasets/ywhmbnzvmx/2

contains 575 instances in which the complex word has three proposed substitutes. A set comprised by the first 500 instances (EASIER-500) of the smallest portion of the EASIER dataset was used by (Alarcon et al., 2021b) and (Alarcón et al., 2021a) to evaluate SG and SS approaches. The instances in this subset contain sentences, a target complex word per sentence and three context-aware substitutions suggested by the expert linguist.

3. ALEXSIS: A new dataset for Lexical Simplification in Spanish

The ALEXSIS Spanish Dataset for Lexical Simplification contains 381 instances[4]. Each instance is composed by a sentence, a target complex word, and 25 candidate substitutions. The dataset format is similar to that of LexMturk (Horn et al., 2014) but in ALEXSIS the sentences are not tokenized (see in Table 1 an instance of the dataset).

The sentences and complex words of this dataset were extracted from the CWI Shared Task 2018 dataset[5] for Spanish. The CWI dataset for Spanish contains 7,015 complex words in which 4,712 are single-words (see (Yimam et al., 2018) for more details about this dataset). A set of 588 pairs <sentence, complex word> was extracted from the Train, Dev and Test files of the Spanish CWI dataset.

The criteria to extract these pairs was that the complex word that appears in the sentence had to be marked as complex word by 5 or more native language annotators and was not a multi-word expression or a word having at least one uppercase letter. From the set of 588 pairs a manual judgment process, involving 2 computational linguists experts, was conducted to decide if the complex word is ”simplifiable”[6] in its context or not. The reviewers had 3 options to mark: simplifiable, not simplifiable or dubious. From this process 3 sets of judgements were obtained: 1) a set of 256 pairs in which both reviewers agreed that the complex word is simplifiable, 2) a set of 113 pairs in which both reviewers agreed that the complex word is not simplifiable, and 3) a set of 219 pairs in which there is disagreement between the reviewers or at least one of the reviewers indicated that had doubts about the simplification. This reviewing process was done using the aid of online dictionaries and thesaurus. Finally, after a joint revision of the 219 dubious pairs previously selected, 146 pairs from these pairs were agreed to be also simplifiable, thus having a total of 402 simplifiable pairs. Then, after deleting few repeated pairs, a set of 393 unique simplifiable pairs was obtained. Afterwards, we applied an additional filtering step that involved removing few cases of: 1) very similar pairs, 2) complex words from other languages that are not yet commonly used and not (yet) accepted as valid words in Spanish (e.g. [hoax]) , 3) complex words that have a sense in the sentence that is used in very specific locations. The final dataset has 384 pairs <sentence, complex word>.

In order to obtain a synonym or a suitable replacement for each complex word, we relied in the crowdsourcing platform *prolific.co*[7] to hire annotators at fair payment who match, to ensure quality, specific minimum language proficiency and education requirements. The requirements to be accepted as annotator were 1) to be fluent in Spanish, and 2) to have an undergraduate (BA/BSc/other) or graduate level (MA/MSc/MPhil/other) as the highest level of completed education. Annotators were asked to propose a single word that is a valid simpler synonym or replacement for the complex word in the context but is easier to understand. If a single-word is not possible then phrases or multi-words are allowed. Otherwise the same complex word should be written. The complete instructions given to the annotators for Spanish (and its translation in English) are shown in the appendices. Based on a pilot study, the estimated average time to complete the task was set to 96 minutes (about 45 seconds per instance). The crowdsourcing platform set the maximum time to perform the task in 196 minutes.

Once the annotation process was finished, the authors decided to delete 2 instances with the complex word repeated two times in its sentence and a sentence which has a typographical error leaving the final number of sentences of the dataset in 381. There are 356 different target words in the dataset: 333 words appear once, 21 words appear twice, and 2 words appear three times.

There are a total of 9,524 substitutions in the dataset and after joining the repeated substitutions in each instance we get a total of 3,918 different substitutions.

One of the authors reviewed the annotations of the dataset and detected that: there are 137 incorrect substitutions (1.43%) and 93 dubious substitutions[8] (0.976%), 230 substitutions are equal to the complex word (2,414%), and 9,064 substitutions are correct (95,17% of the total substitutions).

See in Table 2 a comparison of existing datasets for LS evaluation in several languages. This table shows the language, number of instances, average number of unique synonyms per instance and indicates if the dataset could be used for Lexical Simplicity Ranking.

ALEXSIS and LexMTurk have more unique average number of synonyms per instance (10.28 and 12.85 respectively) among all datasets. LexMTurk has slightly more average number of synonyms per instance while ALEXSIS has a better variability with respect to the mean with a StdDev value of 3.42 (ALEXSIS) and 6.6 (LexMTurk) synonyms per instance. ALEXSIS can only be compared to EASIER. The ALEXSIS dataset has 381 instances but has a higher ratio of unique annotated synonyms per instance (10.28) with respect to the EASIER (full) and the EASIER-500 dataset.

[4] The ALEXSIS dataset will be available after a short embargo period at: https://github.com/LaSTUS-TALN-UPF/ALEXSIS and https://doi.org/10.5281/zenodo.5837149

[5] https://sites.google.com/view/cwisharedtask2018/datasets

[6] That is if the experts found a simpler substitute for the complex word.

[7] https://www.prolific.co

[8] Dubious substitutions are those which, according to the reviewer, are probably incorrect.

EASIER	Sentence	*El colágeno es la proteína más abundante en el organismo, cuya presencia mejora* **notablemente** *la función física.*
	Annotations	mucho, destacadamente, apreciablemente
ALEXSIS	Sentence	*Sufrió una importante reducción en su capacidad para poder* **acogerse** *a las normas de la FIFA para los estadios de fútbol.*
	Annotations	adaptarse (6), refugiarse (2), apegarse (2), ampararse (2), aceptar (2), incorporarse (2), sumarse, recurrir, obedecer, cumplir con, asimilarse, aplicarse, amparar, admitirse, aceptarse

Table 1: Examples of the sentence, complex word (in bold font) and annotations fields of an instance from the EASIER SG/SS dataset and the ALEXSIS dataset.

Dataset	lang	#instances	AvgUniqSyns
SemEval-2012*	en	2,010	4.99
LexMTURK*	en	500	12.85
BENCHLS*	en	929	7.36
NNSEVAL*	en	239	7.49
CEFR-LS*	en	406	2.35
SIMPLEX-PB 3.0*	pt[9]	1,719	7,31
Frenlys*	fr	196	4.03
Chinese-LS*	ch	524	8.51
SNOW E4*	jp	2,330	4.50
BCCWJ*	jp	2,010	4.30
EASIER	es	5,128	1.53
EASIER-500	es	500	3
ALEXSIS*	es	381	10.28

Table 2: Comparison of different LS datasets. Datasets with asterisk (*) indicate that they could allow evaluate Lexical Simplicity Ranking.

4. Substitution Generation Approaches

This section describes the Substitution Generation approaches that we have evaluated with ALEXIS and EASIER-500:

Thesaurus-based approach. This approach is an adaptation of an existing LS system for Ibero-Romance languages based on thesaurus (Ferrés et al., 2017a). The adaptation of the LS system has 4 phases: Document Analysis, Word Sense Disambiguation (WSD), Synonyms Ranking, and Morphological Generation. The WSD algorithm used is based on the Vector Space Model approach for lexical semantics and uses a model extracted from Spanish Wikipedia. The Spanish thesaurus used for WSD was derived from Multilingual Central Repository (MCR) 3.0. The Synonyms Ranking phase ranks synonyms by their lexical simplicity using the ESWIKI-2014 (described in Subsection 6.1) word frequency list (i.e. more frequent is simpler). The Morphological Generation phase combines lexicon-based generation and predictions from Decision-Trees (Ferrés et al., 2017b).

LSBert-es: We have adapted the LSBert[10] (Qiang et al., 2020b) system for Spanish to retrieve and rank the top 80 candidates from a Spanish pre-trained BERT-based model. The language specific resources used to adapt the system to Spanish were: 1) BETO, 2) Snowball stemmer, 3) Fasttext CBOW model for Spanish[11], and 4) SUBTLEX-ESP word frequencies (with zipf values).

Single Transformers with Masked Language Model: We followed the BERT-LS (Qiang et al., 2020a) and LSBert (Qiang et al., 2020b) approach of using the *unmasking* properties of the Masked Language Model of existing pre-trained BERT-based and RoBERTa-based transformers for Spanish to get the top 80 candidates with their associated probabilities (<substitute,score>). We apply the approach consisting on concatenating the original sentence (*S*) with the same sentence with the complex word being masked (*S'*) to obtain the substitute and its likelihood[12]. In this approach, we also perform these procedures: 1) candidates that are equal to the target word or are prefixes (substring) of it are removed, 2) candidates with 2 or less characters are removed, 3) candidates that have the same letters but with accentuation and/or capitalization changes (e.g. publico/*público*, Actual/*actual*) are unified in one unique lowercased/accented candidate with the scores added, and 4) filtering out candidates that are equal to the complex word but without accents or with capitalized letters.

We experimented with 6 transformers models for Spanish. Two of them are derived from BERT (Devlin et al., 2019): BETO and mBERT[13] (Devlin et al., 2019). And four of them are derived from RoBERTa (Liu et al., 2019): SpanBERTa[14], BERTIN[15] (De la Rosa et al., 2022), RoBERTa-base-BNE[16] (Gutiérrez-Fandiño et al., 2021), and RoBERTa-large-BNE[17] (Gutiérrez-Fandiño et al., 2021).

BETO is of size similar to BERT-base and was trained with the Whole Word Masking technique using a Spanish corpus of about 3 billion tokens. mBERT is a pretrained BERT-base model with 102 languages using Wikipedias and masked language modeling (MLM) objective. The SpanBERTa model has the same size as RoBERTa-base and

[9] SIMPLEX-PB 3.0 is a dataset for Brazilian portuguese.

[10] `https://github.com/qiang2100/BERT-LS`

[11] `https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.es.300.vec.gz`

[12] Our adaptation of LSBert does not use probability masking on the original sentence (*S*).

[13] `https://huggingface.co/bert-base-multilingual-uncased`

[14] `https://huggingface.co/chriskhanhtran/spanberta`

[15] `https://huggingface.co/bertin-project/bertin-roberta-base-spanish`

[16] `https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne`

[17] `https://huggingface.co/PlanTL-GOB-ES/roberta-large-bne`

	mBERT	BETO	SpanBERTa	BERTIN	RbaseBNE	RbaseBNE
transformer type	BERT-base	BERT-base	RoBERTa-base	RoBERTa-base	RoBERTa-base	RoBERTa-large
train. corpus	Wikipedia	Several corpus	OSCAR (es)	mC4 (es)	BNE	BNE
size corpus	102 languages	3B (tokens)	18GB (text)	1TB (text)	570GB (text)	570GB (text)
#self-attention layers	12	12	12	12	12	24
#attention heads	12	12	12	12	12	16
#hidden layer size	768	768	768	768	768	1024
#parameters	110M	110M	125M	125M	125M	355M

Table 3: Comparison of BERT and RoBERTa models for Spanish.

was trained on a portion of 18 GB of OSCAR's Spanish corpus. BERTIN is a RoBERTa-based model for Spanish trained on the Spanish portion of mC4 corpus (Xue et al., 2021). RoBERTa-base BNE (RbaseBNE) and RoBERTa-large BNE (RlargeBNE) are Spanish RoBERTa-base and RoBERTa-large models trained with the BNE (National Library of Spain) corpus.

Combination of Results of Single Transformers: In this approach we use the combination of the sets of results obtained by two or more different single transformers to generate a new ranked list of results. There are two modes of use this approach: *Union* and *Intersection*. In the *Intersection* mode ($\cap$) the sets of results' tuples <substitute, score> are intersected and the scores of the substitutes that pertain to the intersection are added. In the *Union* mode ($\cup$) the sets of results' tuples <substitute, score> are joined and the scores are added.

5. Substitution Selection Approaches

We experimented with two SS approaches that can be used independently or in combination: Morphological filtering and POS-tag filtering:

Morphological filtering: Using the Freeling[18] (Padró and Stanilovsky, 2012) dataset of morphosyntactic information to: 1) filter out morphological variations of the complex word in the list of candidates and 2) combine all the morphological variations of candidates in the list of results in a single unique form. These procedures are performed using sets of lemmas associated to morphological forms (556,424 forms with 669,216 form-lemma-Part-of-Speech associations). As an example, the candidates *tocar*, *tocando* and *toca* (which are morphological variations of the verb *tocar* [touch/play]) share the same lemma and are joined in only one candidate form (the form with the highest score among them) and the individual scores are added to the selected candidate form.

POS-tag filtering: The Freeling NLP tool was used for tokenization, morphological analysis[19] , Part-of-Speech (POS) tagging and Named Entity Recognition of both the original sentence (that includes the complex word) and the original sentence with the complex word replaced by the substitution candidate. Then the lexical categories (e.g. adjective, adverb, noun, verb,...) of both the complex word and the substitution candidate are compared and the candidate is discarded if the lexical categories do not match.

6. Substitution Ranking Approaches

We employed two kind of approaches for Substitution Ranking: the ranking based on corpus-based lists of word frequencies (assuming the hypothesis of that a word more frequent that another word is simpler) and a BERT-based fine tuning approach for Lexical Complexity prediction.

6.1. Corpus-based Ranking

SUBTLEX-ESP[20]: a subtitles-based word form frequencies list for Spanish of 94,338 words (Cuetos et al., 2011).

ESWIKI-2014: a set of 2,645,049 word forms frequencies that were extracted from a Spanish Wikipedia dump of 2014 (Ferrés et al., 2017a).

OpenSubtitles-2016[21]: This list (version 2016) has 1,882,198 word forms with its associated frequency in a database of subtitles.

6.2. BERT fine-tuned for Lexical Complexity prediction

A BERT-base based pre-trained model was fine-tuned with a set of 9,607 instances extracted from the CWI2018 dataset (excluding the words included in the ALEXSIS dataset, which was extracted from the same dataset). This set of 9,607 instances has a subset of 3,143 words that are considered complex words (with a score range from 0.1 to 1.0 depending on the judgement of the annotator) and another subset of 6,464 words that are considered simple (with a score of 0.0). After shuffling all instances in random order we assigned a 90% of the dataset for training (8,646 instances), an 8% for model development (768 instances) and a 2% for testing (193 instances).

We followed an approach based partially on the work of (Bani Yaseen et al., 2021) which achieved the best results at the single-word subtask of the SemEval 2021 Lexical Complexity Prediction (LCP) shared task (Shardlow et al., 2021). The model learns to predict a real number that represents the complexity of the word (less complexity indicates more simplicity). We used only single tokens to learn without taking into account the contexts or the sentences in which the word appears. In our approach we used a pre-trained BETO model for fine-tuning with the *BertForSequenceClassification* class of the HuggingFace transformer's library for regression. The values of the parameters used to learn were: learning rate = 5e-6, epsilon = 1e-8,

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

[19] Multiword detection was disabled in this experiment.

[20] http://crr.ugent.be/archives/679

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

batch size= 8, epochs = 4. We achieved a Pearson value of 0.54 for Spanish with the test set of 193 instances.

7. Evaluation Procedure

This section presents the evaluation metrics of the different LS phases evaluated: Substitution Generation and Selection, Substitution Ranking and Full Pipeline. In each subsection we show the tables of results of the datasets tested. The following datasets, which were not used for training the systems in any way, were used to evaluate the approaches: ALEXSIS and EASIER-500.

The Substitution Selection and Substitution Generation phases are evaluated with the same metrics proposed by (Paetzold and Specia, 2016a): Potential, Precision, Recall and F1: 1) *Potential (Pot.)* is the ratio of instances for which at least one of the substitutions generated is present in the gold standard, 2) *Precision (Prec.)* is the ratio of generated candidates that are in the gold standard, 3) *Recall (Rec.)* is the ratio of gold-standard substitutions which are included in the generated substitutions, and 4) *F1* is the harmonic mean between Precision and Recall. Although SG in English is usually evaluated on the top-10 (k=10) candidates, we use in this paper the top-50 (k=50) candidates to evaluate the approaches with ALEXSIS and the EASIER-500 in order to have results that allow a comparison with the evaluation carried out by (Alarcón et al., 2021a; Alarcon et al., 2021b).
In order to evaluate the SS approaches, we followed the methodology presented by (Paetzold and Specia, 2016a) of using the candidates produced by all the generators altogether. The Substitution Selection approaches were evaluated only with the ALEXSIS dataset using the union (∪) of the results of all the approaches that use transformers (∪-all-TFS), but excluding results obtained with the LSBert-es system.

System	Pot.	Prec.	Rec.	F1
Baselines (Alarcón et al., 2021a)				
Word2vec	0.358	0.019	0.188	0.034
FastText	0.464	0.029	0.289	0.053
Sense2Vec	0.506	0.056	0.298	0.095
BETO	0.348	0.03	0.282	0.054
Our Approaches				
Thesaurus-based	0.198	**0.124**	0.089	**0.104**
LSBert-es (BETO)	0.764	0.027	0.464	0.051
BETO	0.697	0.025	0.422	0.048
SpanBERTa	0.848	0.035	0.601	0.067
mBERT	0.328	0.010	0.161	0.019
BERTIN	0.800	0.033	0.564	0.063
RbaseBNE	0.826	0.035	0.589	0.067
RlargeBNE	0.824	0.035	0.585	0.067
BETO ∪ SpanBERTa	0.782	0.031	0.533	0.059
BETO ∩ SpanBERTa	0.674	0.052	0.406	0.093
SpanBERTa ∪ RbaseBNE	**0.866**	0.036	**0.618**	0.069
SpanBERTa ∩ RbaseBNE	0.808	0.049	0.561	0.090

Table 4: SG evaluation on the EASIER-500 dataset with top-k=50.

System	Pot.	Prec.	Rec.	F1
Thesaurus+eswiki	0.146	**0.132**	0.021	0.037
LSBert-es (BETO)	0.860	0.047	0.245	0.079
mBERT	0.545	0.023	0.118	0.039
BETO	0.782	0.042	0.213	0.071
SpanBERTa	0.892	0.059	0.308	0.099
BERTIN	0.853	0.057	0.294	0.096
RbaseBNE	0.913	0.061	0.317	0.103
RlargeBNE	0.910	0.061	0.318	0.103
BETO ∪ SpanBERTa	0.876	0.053	0.277	0.089
BETO ∩ SpanBERTa	0.745	0.093	0.188	0.124
BERTIN ∪ RbaseBNE	**0.921**	0.062	0.325	0.105
BERTIN ∪ RbaseBNE	0.837	0.085	0.263	0.129
SpanBERTa ∪ RbaseBNE	0.918	0.064	**0.332**	0.107
SpanBERTa ∩ RbaseBNE	0.881	0.087	0.284	**0.133**

Table 5: SG evaluation on the ALEXSIS dataset on the top-k=50.

System	Potential	Prec.	Rec.	F1
top-k=3				
∪-all-TFS	**0.632**	**0.300**	**0.093**	**0.142**
+morpho.	0.629	0.298	0.092	0.141
+POStag	0.627	0.293	0.090	0.138
+morpho+POS	0.627	0.293	0.091	0.139
top-k=10				
∪-all-TFS	0.792	0.170	0.176	0.173
+morpho.	0.790	**0.171**	**0.177**	**0.174**
+POStag	0.795	0.167	0.173	0.170
+morpho+POS	**0.797**	0.169	0.174	0.171
top-k=50				
∪-all-TFS	**0.921**	**0.059**	**0.309**	**0.100**
+morpho.	0.913	0.058	0.304	0.098
+POStag	0.910	0.058	0.300	0.097
+morpho+POS	0.902	0.057	0.294	0.096

Table 6: SS results on the ALEXSIS U-all-TFS dataset.

The Substitution Ranking evaluation metric used is TRank-at-n (n=1,2,3)[22] modified to take into account cases in which two or more candidates have the same predicted weight and rank. *Trank-at-n (TRnk-n)* is the ratio of instances in which a candidate of gold-rank r≤n was ranked first.

System	TRnk-1	TRnk-2	TRnk-3
BERT fine-tuned	0.1811	0.3779	**0.6657**
SUBTLEX-ESP	0.1889	**0.3963**	0.6648
ESWIKI	**0.1916**	0.3597	0.6505
OPENSUBTITLES	0.1837	0.3805	0.664

Table 7: Evaluation of the Substitution Ranking Approaches on the ALEXSIS dataset.

The evaluation metrics for the Full Pipeline are: 1) *Precision*: the ratio of instances with the top ranked candidate is either the target word itself or is in the gold standard list of annotated candidates, 2) *Accuracy*: the ratio of instances with the top ranked candidate in the gold standard list of annotated candidates, and 3) *Changed*: the ratio of instances

[22] https://github.com/ghpaetzold/LEXenstein

in which the top ranked candidate is not the target word itself.

System	Prec.	Acc.	Changed
Thesaurus (eswiki)	**0.776**	0.096	0.320
LSBert-es (BETO)	0.236	0.228	0.992
Transformers			
RlargeBNE	0.33	0.312	0.982
SpanBERTa	0.281	0.281	1.0
BERTIN ∩ SpanBERTa	0.338	0.302	0.964
BERTIN ∪ SpanBERTa	0.290	0.290	1.0
Transformers+filterMorpho+filterPOS			
RlargeBNE	0.35	0.326	0.976
SpanBERTa	0.352	0.347	0.996
BERTIN ∩ SpanBERTa	0.402	0.347	0.946
BERTIN ∪ SpanBERTa	0.354	**0.350**	0.996

Table 8: Evaluation of the full pipeline on the EASIER-500 dataset. Approaches with single transformers or combined results use the top-k=1 candidate to evaluate.

System	Prec.	Acc.	Change
Thesaurus (eswiki)	**0.889**	0.089	0.199
LSBert-es (BETO)	0.278	0.278	1.0
Transformers			
RbaseBNE	0.438	0.438	1.0
SpanBERTa	0.409	0.409	1.0
SpanBERTa ∩ RbaseBNE	0.456	0.456	1.0
SpanBERTa ∪ RbaseBNE	0.454	0.454	1.0
Transformers + filterMorpho+filterPOS			
RbaseBNE	0.454	0.451	0.997
SpanBERTa	0.448	0.448	1.0
SpanBERTa ∩ RbaseBNE	0.475	0.461	0.986
SpanBERTa ∪ RbaseBNE	0.469	**0.469**	1.0

Table 9: Evaluation of some full pipeline combinations (SG, SG+SR) on the ALEXSIS dataset.

8. Results and Discussion

This section presents the results of the different systems evaluated in ALEXSIS and EASIER-500 and discusses our main findings in: SG, SS, SR and the Full pipeline.

8.1. Substitution Generation

We evaluated the Substitute Generation phase with both EASIER-500 and ALEXSIS (see the results in Tables 4 and 5 respectively). The results of evaluating our approaches and the adapted LSBert-es system with respect to the baselines (Alarcón et al., 2021a) with the EASIER-500 dataset show that: 1) all the tranformers-based approaches (with the exception of mBERT) outperform the baselines in Potential and Recall, so achieving state-of-the-art results, 2) the approaches that use the LSBert MLM strategy with BETO (LSBert-es and BETO) largely outperform the BETO baseline, 3) the best results for each metric are achieved by these approaches: Potential and Recall (the SpanBERTa ∪ RbaseBNE approach with 0.866 and 0.618 respectively), Precision and F1 (the Thesaurus-based approach with 0.124 and 0.104 respectively). On the other hand, the results of evaluating our approaches and the LSBert-es system with the ALEXSIS dataset are the following: 1) all the transformers-based approaches achieve better Potential, Recall and F1 with respect to the Thesaurus-based approach, 2) the Thesaurus-based approach achieves the best Precision (0.132), 3) the approach BERTIN ∪ RbaseBNE achieves the best results in Potential (0.921), 4) the approach SpanBERTa ∪ RbaseBNE achieves the best Recall (0.332), and 5) the approach SpanBERTa ∩ RbaseBNE achieves the best F1 (0.133).

In Table 10 at Section 13 (Appendix) we show a Qualitative Evaluation of the results in SG using the ALEXIS dataset with four of the top-performing approaches in this phase. That table contains 6 examples of: a) a sentence with the marked complex word, b) the annotated substitutions (gold-standard), and c) the top-k=10 candidates of the SG phase for four approaches: LSBert-es, RbaseBNE, BERTIN ∪ RbaseBNE, SpanBERTa ∪ RoBERTaBne. The comparison shows the problems that involve the retrieval of words that could fit semantically in the context but are not synonyms or correct replacements (e.g. in Sentence B with *folclórico* [folkloric] as complex word, *argentino* [argentinian] and other demonyms are retrieved as candidates). Sometimes the retrieved candidates can be antonyms of the complex words (e.g in Sentence A the target word *difunto* [deceased]) has the antonym *nuevo* [new] in the list of candidates).

8.2. Substitution Selection

The results of evaluating the SS approaches with the ALEXSIS dataset are shown in Table 6. The results show that these approaches achieve a slightly improvement in some metrics evaluated at top-k=10 for the morphological (morpho.) filter, for the POS-tag filter, and for the approach that combines both filters. The results indicate that these approaches can help to improve the filtering out of some incorrect candidates, but with low gains.

8.3. Substitution Ranking

The evaluation of the Substitution Ranking approaches was done with the ALEXSIS dataset. The gold-ranking of the unique annotated substitutes for each instance was determined by the number of annotators that had chosen each unique substitute as a measure of simplicity. The results of this evaluation (see Table 7) show that different approaches achieve the best results for different metrics: ESWIKI approach was the top one for TRnk-1, SUBTLEX-ESP is the top one for TRnk-2, and BERT fine-tuned for Lexical Complexity Prediction was the best for TRnk-3. But it is worth noticing that the differences in most of the results were not greater than 0.04 points.

8.4. Full Pipeline

Finally, regarding the full pipeline evaluation, we evaluated different combinations of the SG, SS and SR approaches. For the full pipeline evaluation with the EASIER-500 and ALEXIS datasets we evaluated the Thesaurus-based approach, the full LSBert-es adaptation (with SG+SR), and several approaches with single transformers and combinations of transformers[23] (with SG, SG+SS, SG+SR, and SG+SS+SR pipelines). As shown in Tables 8 and 9 the set of approaches named *Transformers* indicates the ones that used only the SG phase, selecting the top-k=1. The set of approaches named *Transformers+ filterMorpho + filterPOS* shows the results of those that used SG in combination with both SS aproaches applied sequentially after SG and selecting the top-k=1 candidate as the final selected replacement.

The evaluation of the full pipeline approaches with the EASIER-500 dataset (see results in Table 8) shows that the best results in Precision were achieved by the Thesaurus-based approach with a value of 0.776, but with a very low Accuracy (0.096) and Changed ratio (0.32) and the best results in Accuracy were achieved with the BERTIN ∪ SpanBERTa approach in combination with morphological and POS-tag filtering with a value of 0.350 and a change ratio of 0.9962.

The evaluation of the full pipeline approaches with the ALEXSIS dataset (see results in Table 9) shows that the best results in Precision were achieved by the Thesaurus-based approach (with a value of 0.889, but with a very low Accuracy (0.089) and Changed ratio (0.199) and the best results in Accuracy were achieved with the SpanBERTa ∪ RoBERTABbne approach for SG combined with the morphological filter and the POS-tag filter for Substitution Selection (with a value of 0.4698 and a Change ratio of 1.0).

The two best pairs of approaches with SG+SS pipelines based on single and combined transformers were: 1) the RlargeBNE, SpanBERTa, BERTIN ∩ SpanBERTa and BERTIN ∪ SpanBERTa approaches for the EASIER-500 dataset, and 2) the RbaseBNE, SpanBERTa, SpanBERTa ∩ RoBERTaBbne, SpanBERTa ∪ RoBERTaBbne, for the ALEXSIS dataset.

The Substitution Ranking algorithms were tested in pipelines that used combinations of SG+SR and SG+SS+SR approaches but generally, in almost all cases the results were not improving the baseline (SG or SG+SS) and were not reported in the tables of results.

9. Conclusions and Further Work

ALEXSIS is the first dataset for Lexical Simplification evaluation in Spanish that includes information potentially useful for Lexical Simplicity Ranking[24] and which has higher number of average unique synonyms per instance compared with EASIER/EASIER-500 (Alarcón et al., 2021a) (a previous existing LS dataset for Spanish) and most of the other datasets for other Languages. ALEXSIS and EASIER-500 datasets have been used to benchmark several neural-based approaches: 1) our adaptation of LSBert (Qiang et al., 2020b) for Spanish and, 2) other neural approaches based on pre-trained transformers and the MLM strategy of LSBert.

The manual inspection of the gold annotations and the proposed candidates by the systems (explained in Section 3) shows some minor but yet relevant issues about the annotation procedure: 1) some valid (and perhaps frequent) synonyms were not reported by the annotators (e.g. in one example we found that the generated substitution *destacada* [outstanding] could replace correctly *reputada* [reputed]), and 2) some annnotators reported a bad morphological form of the substitution which should not be accepted in the given context. (e.g. in another example, the proposed human substitution *prestigioso* [prestigious] instead of the correct inflected form in Spanish *prestigiosa* [prestigious]).

To solve these issues, further work can include the creation of a new version of the dataset without incorrect substitutions and dubious words and with new compiled synonyms and its evaluation and comparison with the original dataset.

The dataset could be extended for each instance (if necessary) in two ways 1) with other synonyms compiled manually using dictionaries and thesaurus during the dataset creation, and 2) with synonyms compiled manually by inspecting the output of Spanish pre-trained Transformers applying the Transformers plus complex word MLM proposed by (Qiang et al., 2020b) and other existing unsupervised systems.

On the other hand, both the manual inspection of the dataset and the low TRnk-1/TRnk-2/TRnk-3 results (compared with similar experiments with frequency counts lists in other languages (Paetzold and Specia, 2017a; Rolin et al., 2021)) could indicate that the Lexical Simplicity Ranking by aggregation (using the addition of repeated annotations to get gold-ranked substitutes by simplicity) may not be suitable for generic (age-independent / education-independent) ranking. For this reason, further investigation is needed and an additional step of ranking annotation or verification performed by linguistic experts might be required in order to have generic substitution Ranking.

The Substitution Generation phase could be improved with: 1) new experiments with combination of results of two and more transformers, 2) a combination of deep learning technologies and thesaurus-based approaches (Rolin et al., 2021).

Furthermore, the Substitution Selection approaches, should be improved to filter out wrong candidates. This phase could be improved in further work with the use of ngrams (Paetzold and Specia, 2016c).

Regarding the Substitution Ranking approaches, methods that combine neural ranking with several linguistic and corpus-based features such as (Maddela and Xu, 2018) could be applied to achieve better results in ranking.

10. Acknowledgements

We acknowledge support from the project Context-aware Multilingual Text Simplification (ConMuTeS) PID2019-109066GB-I00/AEI/10.13039/501100011033 awarded by Ministerio de Ciencia, Innovación y Universidades (MCIU) and by Agencia Estatal de Investigación (AEI) of Spain.

[23] Only the top-2 performing of each type (single or combinations) are reported in the results

[24] Assuming that the aggregation of repeated annotations can be used for Lexical Simplicity Ranking.

11. Bibliographical References

Alarcón, R., Moreno, L., and Martínez, P. (2021a). Exploration of Spanish Word Embeddings for Lexical Simplification. In *Proceedings of the First Workshop on Current Trends in Text Simplification CTTS 2021)*, volume 2944 of *CEUR Workshop Proceedings*. CEUR-WS.org, September.

Alarcon, R., Moreno, L., and Martínez, P. (2021b). Lexical Simplification System to Improve Web Accessibility. *IEEE Access*, 9:58755–58767, April.

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

Baeza-Yates, R. A., Rello, L., and Dembowski, J. (2015). CASSA: A Context-Aware Synonym Simplification Algorithm. In *Proceedings of NAACL HLT 2015*, pages 1380–1385.

Bani Yaseen, T., Ismail, Q., Al-Omari, S., Al-Sobh, E., and Abdullah, M. (2021). JUST-BLUE at SemEval-2021 Task 1: Predicting Lexical Complexity using BERT and RoBERTa Pre-trained Language Models. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 661–666, Online, August. Association for Computational Linguistics.

Biran, O., Brody, S., and Elhadad, N. (2011). Putting It Simply: A Context-aware Approach to Lexical Simplification. In *Proceedings of the ACL 2011*, pages 496–501. Association for Computational Linguistics.

Bott, S., Rello, L., Drndarevic, B., and Saggion, H. (2012). Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish. In *COLING*, pages 357–374. Indian Institute of Technology Bombay.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. (2020). Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR 2020*.

Cuetos, F., González-Nosti, M., Barbón, A., and Brysbaert, M. (2011). SUBTLEX-ESP: spanish word frequencies based on film subtitles. *Psicológica*, 32:133–143, 01.

De Belder, J. and Moens, M.-F. (2010). Text Simplification for Children. In *Proceedings of the SIGIR Workshop on Accessible Search Systems*, pages 19–26.

De Belder, J. and Moens, M.-F. (2012). A Dataset for the Evaluation of Lexical Simplification. In *Proceedings of the 13th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part II*, CICLing'12, page 426–437, Berlin, Heidelberg. Springer-Verlag.

De la Rosa, J., Ponferrada, E. G., Romero, M., Villegas, P., González de Prado Salas, P., and Grandury, M. (2022). BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling. *Procesamiento del Lenguaje Natural*, 68(0):13–23.

Devlin, S. and Tait, J. (1998). The Use of a Psycholinguistic Database in the Simplification of Text for Aphasic Readers. In *Linguistic Databases*, pages 161–173.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Ferrés, D., Saggion, H., and Gómez Guinovart, X. (2017a). An Adaptable Lexical Simplification Architecture for Major Ibero-Romance Languages. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 40–47, Copenhagen, Denmark, September. Association for Computational Linguistics.

Ferrés, D., Saggion, H., and Gómez Guinovart, X. (2017b). An Adaptable Lexical Simplification Architecture for Major Ibero-Romance Languages. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 40–47, Copenhagen, Denmark, September. Association for Computational Linguistics.

Glavaš, G. and Štajner, S. (2015). Simplifying Lexical Simplification: Do We Need Simplified Corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 63–68, Beijing, China, July. Association for Computational Linguistics.

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., Gonzalez-Agirre, A., Armentano-Oller, C., Penagos, C. R., and Villegas, M. (2021). Spanish Language Models. *CoRR*, abs/2107.07253.

Hartmann, N. S. and Aluísio, S. M. (2020). Adaptação Lexical Automática em Textos Informativos do Português Brasileiro para o Ensino Fundamental. *Linguamática*, 12(2):3–27, December.

Horn, C., Manduca, C., and Kauchak, D. (2014). Learning a Lexical Simplifier Using Wikipedia. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 458–463, Baltimore, Maryland, June. Association for Computational Linguistics.

Kajiwara, T. and Yamamoto, K. (2015). Evaluation Dataset and System for Japanese Lexical Simplification. In *Proceedings of the ACL-IJCNLP 2015 Student Research Workshop*, pages 35–40, Beijing, China, July. Association for Computational Linguistics.

Kodaira, T., Kajiwara, T., and Komachi, M. (2016). Controlled and Balanced Dataset for Japanese Lexical Simplification. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 1–7.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Maddela, M. and Xu, W. (2018). A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3749–3760, Brussels, Belgium, October-November. Association for Computational Linguistics.

Padró, L. and Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of LREC 2012*. ELRA.

Paetzold, G. and Specia, L. (2016a). Benchmarking Lexical Simplification Systems. In *Proceedings of LREC-2016*.

Paetzold, G. and Specia, L. (2016b). Unsupervised Lexical Simplification for Non-Native Speakers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), Mar.

Paetzold, G. and Specia, L. (2016c). Unsupervised Lexical Simplification for Non-Native Speakers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), Mar.

Paetzold, G. and Specia, L. (2017a). A Survey on Lexical Simplification. *Journal of Artificial Intelligence Research*, 60:549–593, 11.

Paetzold, G. and Specia, L. (2017b). Lexical Simplification with Neural Ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 34–40, Valencia, Spain, April. Association for Computational Linguistics.

Petersen, S. E. and Ostendorf, M. (2007). Text Simplification for Language Learners: a Corpus Analysis. In *In Proc. of Workshop on Speech and Language Technology for Education*.

Qiang, J., Li, Y., Yi, Z., Yuan, Y., and Wu, X. (2020a). Lexical Simplification with Pretrained Encoders. *Thirty-Fourth AAAI Conference on Artificial Intelligence*, page 8649–8656.

Qiang, J., Li, Y., Zhu, Y., Yuan, Y., and Wu, X. (2020b). LSBert: A Simple Framework for Lexical Simplification. *arXiv preprint arXiv:2006.14939*.

Qiang, J., Lu, X., Li, Y., Yuan, Y., and Wu, X. (2021). Chinese Lexical Simplification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1819–1828.

Rolin, E., Langlois, Q., Watrin, P., and François, T. (2021). FrenLyS: A Tool for the Automatic Simplification of French General Language Texts. In Galia Angelova, et al., editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Held Online, 1-3 September, 2021*, pages 1196–1205. INCOMA Ltd.

Saggion, H., Štajner, S., Bott, S., Mille, S., Rello, L., and Drndarevic, B. (2015). Making it Simplext: Implementation and Evaluation of a Text Simplification System for Spanish. *TACCESS*, 6(4):14.

Saggion, H. (2017). *Automatic Text Simplification*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Shardlow, M., Evans, R., Paetzold, G. H., and Zampieri, M. (2021). SemEval-2021 Task 1: Lexical Complexity Prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online, August. Association for Computational Linguistics.

Shardlow, M. (2014). A Survey of Automated Text Simplification. *International Journal of Advanced Computer Science and Applications*, 4, 01.

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

Štajner, S., Saggion, H., and Ponzetto, S. P. (2019). Improving Lexical Coverage of Text Simplification Systems for Spanish. *Expert Systems with Applications*, 118:80–91.

Štajner, S. (2014). Translating Sentences from Original to Simplified Spanish. *Procesamiento del lenguaje natural*, 53:61–68.

Uchida, S., Takada, S., and Arase, Y. (2018). CEFR-based Lexical Simplification Dataset. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June. Association for Computational Linguistics.

Yatskar, M., Pang, B., Danescu-Niculescu-Mizil, C., and Lee, L. (2010). For the Sake of Simplicity: Unsupervised Extraction of Lexical Simplifications from Wikipedia. In *Proceedings of HLT-NAACL 2010*.

Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G. H., Specia, L., Štajner, S., Tack, A., and Zampieri, M. (2018). A Report on the Complex Word Identification Shared Task 2018. *CoRR*, abs/1804.09132.

12. Language Resource References

Alarcón, R. (2021). Dataset of sentences annotated with complex words and their synonyms to support lexical simplification. https://doi.org/10.25592/uhhfdm.9722.

Paetzold and Specia. (2016a). BenchLS: A Reliable Dataset for Lexical Simplification, May. https://doi.org/10.5281/zenodo.2552393.

Paetzold and Specia. (2016b). NNSeval: Evaluating Lexical Simplification for Non-Natives, February. https://doi.org/10.5281/zenodo.2552381.

13. Appendix: Qualitative Evaluation

Sentence A	Además de partidos de fútbol americano, el estadio ha sido utilizado para una gran variedad de eventos, entre los que se destacan varios partidos de la selección nacional de fútbol de los Estados Unidos, y fue el hogar del ahora **difunto** club de la MLS, el Tampa Bay Mutiny.
annotated labels	extinto (7), muerto (5), fallecido (5), finado (3), desaparecido (2), inexistente (1), inactivo (1), acabado(1)
LSBert-es (BETO)	**extinto**, **desaparecido**, **fallecido**, ex, nuevo, antiguo, retirado, **muerto**, actual, viejo
RbaseBNE	poderoso, famoso, **desaparecido**, **fallecido**, nuevo, emblemático, histórico, mítico, último, gran
BERTIN ∪ RbaseBNE	poderoso, famoso, **desaparecido**, **fallecido**, nuevo, emblemático, histórico, mítico, último, gran
SpanBERTa ∪ RbaseBNE	poderoso, **fallecido**, **desaparecido**, famoso, nuevo, emblemático, histórico, mítico, último, gran

Sentence B	), Fue un pianista y compositor **folclórico**, y figura fundamental en lo que a interpretación pianística del folclore argentino se refiere.
annotated labels	tradicional (10), pintoresco (3), típico (3), de folclore (2), costumbrista (2), popular (1), local (1), de música folk (1), de folclor (1)
LSBert-es (BETO)	argentino, uruguayo, chileno, peruano, boliviano, mexicano, ecuatoriano, italiano, colombiano, **popular**
RbaseBNE	argentino, flamenco ,musical, **popular**, **tradicional**, clásico, mexicano, uruguayo, español, folk
BERTIN ∪ RbaseBNE	argentino, **popular**, flamenco, folclore, religioso, musical, **tradicional**, dramático, histórico, político
SpanBERTa ∪ RbaseBNE	argentino, flamenco, **popular**, musical, folklore, **tradicional**, folk, uruguayo, clásico, mexicano

Sentence C	El texto denominado ”Lamentos de Ipuwer” describe una situación caótica: reyes **desacreditados**, invasión asiática del Delta, desórdenes revolucionarios, destrucción de archivos y tumbas reales, ateísmo y divulgación de secretos religiosos.
annotated labels	desprestigiados (11), difamados (5), desacreditados (2), demeritados (1), denigrados (1), desmentidos (1), desprestigiado (1), malos (1), olvidados (1), sin prestigio (1)
LSBert-es (BETO)	falsos, secuestrados, perdidos, poderosos, muertos, abandonados, sospechosos, reconocidos, conocidos, corruptos, viejos
RbaseBNE	amenazados, prohibidos, encarcelados, corruptos, rebeldes, perseguidos, fascistas, expulsados, desconocidos, acusados
BERTIN ∪ RbaseBNE	perseguidos, derrotados, amenazados, encarcelados, prohibidos, asesinados, corruptos, rebeldes, marginados, destruidos
SpanBERTa ∪ RbaseBNE	amenazados, destruidos, prohibidos, perseguidos, corruptos, encarcelados, rebeldes, enemigos, descubiertos, derrotados

Sentence D	Floreció en la época clásica y tenía una **reputada** escuela de filosofía.
annotated labels	prestigiosa (6), famosa (4), afamada (2), respetada (2), renombrada (2), conocida (2), reconocida (1), muy reconocida (1), valorada (1), prestigioso (1), prestigiada (1), acreditada (1)
LSBert-es (BETO)	**reconocida**, **famosa**, importante, gran, excelente, **conocida**, propia, extensa, destacada, buena
RbaseBNE	**prestigiosa**, **reconocida**, importante, destacada, gran, brillante, notable, sólida, **conocida**, amplia
BERTIN ∪ RbaseBNE	**prestigiosa**, **reconocida**, importante, destacada, modesta, gran, respetable, estupenda, notable, magnífica
SpanBERTa ∪ RbaseBNE	**prestigiosa**, destacada,gran, **reconocida**, importante, buena, nueva, brillante, verdadera, larga

Sentence E	Pocos días más tarde, el 8 de septiembre, durante los Campeonatos de Europa de Roma, las mismas integrantes . ganaron el oro y **batieron** el récord mundial con 42,51
annotated labels	rompieron (12), vencieron (5), superaron (1), sobrepasaron (1), percutieron (1), mejoraron (1), lograron (1), consiguieron (1), conquistaron (1), alcanzaron (1)
LSBert-es (BETO)	**rompieron**, establecieron, **alcanzaron**, mantuvieron, destruyeron, rompen, **consiguieron**, **lograron**, pusieron, obtuvieron
RbaseBNE	**superaron**, **rompieron**, establecieron, **alcanzaron**, batió, **consiguieron**, **lograron**, pusieron, confirmaron, marcaron
BERTIN ∪ RbaseBNE	**rompieron**, **superaron**, **alcanzaron**, establecieron, batió, vencieron, **consiguieron**, **lograron**, pusieron, tocaron
SpanBERTa ∪ RbaseBNE	**superaron**, **rompieron**, **alcanzaron**, establecieron, batió, obtuvieron, **lograron**, **consiguieron**, pusieron, batir

Sentence F	Según algunos estudios se hace referencia a que la obra de construcción de la iglesia es anterior al S. XIV, pero debido a una gran reforma que sufrió entre los siglos XV y XVI que le da su imagen actual (excepto la fachada) hace que se **catalogue** como construida en el S. XVI.
annotated labels	clasifique (14), identifique (2), considere (2), categorice (2), registre (1), etiquete (1), entre (1), determine (1), anote (1)
LSBert-es (BETO)	**considere**, clasifica, conozca, declare, trate, vea, reconozca, **determine**, indique, consideren
RbaseBNE	declare, **considere**, mantenga, presente, quede, trate, cite, ocupe, encuentre, constituya
BERTIN ∪ RbaseBNE	**considere**, declare, **registre**, mantenga, **identifique**, presente, trate, reconozca, quede, cite
SpanBERTa ∪ RbaseBNE	**considere**, declare, mantenga, presente, quede, trate, cite, ocupe, encuentre, constituya

Table 10: Qualitative evaluation and comparison of top-k=10 results of four Substitution Generation approaches on the ALEXSIS dataset. Retrieved complex words or substrings of the complex words are not reported.

14. Appendix: Instructions for Annotators (in Spanish)

A continuación se presentan **128** oraciones en español, en cada oración hay una palabra marcada en negrita. Su tarea es escribir, en el espacio debajo de cada oración, una única palabra que tenga el mismo significado que la marcada, pero que sea más fácil de entender. Por ejemplo, en la oración ”Al mismo tiempo, se **atenuó** el ritmo de caída respecto del dólar” la palabra **atenuó** podría reemplazarse por la palabra más fácil de entender *disminuyó*. Escriba el reemplazo de manera que la sustitución sea válida en el contexto dado. En nuestro ejemplo, *disminuyó* es correcto mientras que disminuir no lo sería. En caso de que no fuera posible reemplazar por una única palabra, entonces usted podrá utilizar una substitución más compleja. Por ejemplo en la oración “Los vestidos eran **iraníes**” , la palabra **iraníes** podría reemplazarse por **de Irán**. Se admiten también reemplazos que comporten un cambio de género con respecto a la palabra marcada. Por ejemplo, en la oración ”Ganar es nuestra **meta**.”, el reemplazo *fin* es aceptable aunque hubiera que realizar cambios a la oración.

Nota1: si ocurriese la situación en que usted no encuentra una palabra más simple entonces debe escribir la misma palabra compleja en la zona de respuesta.

Nota2: para hacer la tarea se permite la ayuda y uso de todo tipo de recursos léxicos de consulta como diccionarios, diccionarios de sinónimos, etc, ya sean libros o por internet.

ADVERTENCIA: En esta tarea es importante que usted siga las instrucciones para recibir su pago. Al completar la tarea y clicar al botón en color morado ”Enviar” usted afirma haber leído y estar de acuerdo las condiciones de la ficha de información y consentimiento.

Ficha de Información y Consentimiento.
El estudio tiene como objetivo recopilar ejemplos de simplificación léxica para el español. Los datos recopilados se utilizarán únicamente con fines de investigación. Usted va a leer frases en las cuales aparecerá una palabra considerada compleja que usted debería de simplificar proponiendo otra palabra que tenga el mismo significado pero que sea más fácil de entender. Los datos recolectados serán utilizados en un proyecto de investigación y se facilitarán a investigadores que lo necesiten. Los resultados de esta investigación se podrán publicar en revistas científicas o conferencias y podrán ser utilizados en estudios posteriores. Para participar en este experimento usted debería:

- a) Ser hablante nativo de español,
- b) Tener al menos 18 años y ser competente para dar su consentimiento,
- c) Haber leído y comprendido esta Ficha de Información que explica el proyecto de investigación,
- d) Acepta que los datos recopilados se utilicen de forma anónima en el futuro,
- e) Aceptar participar en la investigación descrita anteriormente.

¡Grácias por participar!

15. Appendix: Instructions for Annotators (in English)

Below are **128** sentences in Spanish, in each sentence there is a word marked in bold. Your task is to write, in the space below each sentence, single word that has the same meaning as the one marked, but is easier to understand. For example, in the sentence ”At the same time, the rate of decline against the dollar was **attenuated**” the word **attenuated** could be replaced by the easier-to-understand word *decreased*. Write the replacement so that the replacement is valid in the given context. In our example, *decreased* is correct while decrease would not be correct. In that case that it is not possible to replace with a single word, then you can use a more complex substitution. For example in the sentence ”The dresses were **Iranian**”, the word **Iranian** could be replaced by “**from Iran**”. Replacements that involve a gender change with respect to the marked word are also allowed (Note that this is not applicable in English). For example, in the sentence ”Winning is our **goal**.”, The word *end* as a replacement is acceptable even if changes are required to the sentence (Note that this is not applicable in English).

Note1: if the situation occurs where you cannot find a simpler word then you must write the same complex word in the answer area.

Note2: to do the homework, the help and use of all kinds of lexical reference resources such as dictionaries, thesaurus, etc., whether books or online, is allowed.

WARNING: In this task it is important that you follow the instructions to receive your payment. By completing the task and clicking the purple button ”Send” you affirm that you have read and agree to the conditions of the information and consent form.

Information and Consent Form
The study aims to collect examples of lexical simplification for Spanish. The data collected will be used for research purposes only. You will read sentences in which a word considered complex will appear that you should simplify by proposing another word that has the same meaning but is easier to understand. The data collected will be used in a research project and will be provided to researchers who need it. The results of this research may be published in scientific journals or conferences and may be used in subsequent studies. To participate in this experiment you should:

- a) Be a native Spanish speaker,
- b) Be at least 18 years old and competent to give consent.
- c) Have read and understood this Information Form that explains the research project,
- d) You agree that the data collected will be used anonymously in the future,
- e) Agree to participate in the research described above.

Thanks for participating!