Lexical Simplification in Foreign Language Learning: Creating Pedagogically Suitable Simplified Example Sentences

Jasper Degraeuwe
LT^3 / MULTIPLES
Ghent University
Belgium
Jasper.Degraeuwe@UGent.be

Horacio Saggion
LaSTUS / TALN
Universitat Pompeu Fabra
Spain
horacio.saggion@upf.edu

Abstract

This study presents a lexical simplification (LS) methodology for foreign language (FL) learning purposes, a barely explored area of automatic text simplification (TS). The method, targeted at Spanish as a foreign language (SFL), includes a customised complex word identification (CWI) classifier and generates substitutions based on masked language modelling. Performance is calculated on a custom dataset by means of a new, pedagogically-oriented evaluation. With 43% of the top simplifications being found suitable, the method shows potential for simplifying sentences to be used in FL learning activities. The evaluation also suggests that, though still crucial, meaning preservation is not always a prerequisite for successful LS. To arrive at grammatically correct and more idiomatic simplifications, future research could study the integration of association measures based on co-occurrence data.

1 Introduction

The rise of digital corpora has been steadily transforming the FL learning domain. As corpora are an easy-to-compile source of natural text which can be consulted in a highly efficient fashion (Granger et al., 2007; Pilán et al., 2016), they are arriving at a stage of being seamlessly embedded in several aspects of the everyday language learning practice (Chambers, 2019). This "normalisation" (Bax, 2003) of corpora is especially evidenced in the growing interest in data-driven learning (DDL; Johns, 1990). In its broadest sense, this area refers to both teachers and learners "using the tools and techniques of corpus linguistics for pedagogical purposes" (Gilquin and Granger, 2010, p. 359). While learner-led DDL activities tend to consist in analysing concordance lines (e.g. to discover collocations), teacher-focused DDL usually corresponds to accessing corpora directly in order to generate resources such as vocabulary lists and fill-the-gap exercises, which has led to the concept of "corpus-informed language teaching" (Jablonkai and Csomay, 2022).

However, working with corpora also entails challenges and limitations, for instance with regard to learner proficiency levels. To begin with, DDL has been found to be beneficial for intermediate and advanced learners (Boulton and Cobb, 2017), but with lower levels the credentials of using corpora still have to be established (Boulton and Vyatkina, 2021). Furthermore, also between intermediate and advanced learners considerable differences can exist, for example with respect to vocabulary knowledge: the lower one's language proficiency, the less extensive one's vocabulary will be (Laufer and Nation, 1995). In DDL, this can lead to the following scenario: while preparing a language for specific purposes (LSP) class on economics, an SFL teacher is using a corpus query tool to create a fill-the-gap exercise for the target item *arancel* and finds sentence (a) below in the query output. However, if this sentence were to be included in the final exercise, low- or intermediate-proficiency learners could find themselves unable to solve it, as their limited vocabulary knowledge might prevent them from understanding essential parts of the context, such as the word *esquivar*.

(a) La planta local también permitirá **esquivar** los aranceles. ('The local plant will also allow **evading** tariffs.')

To overcome this limitation, (part of) the corpus data can be simplified according to the needs of the target audience (Gilquin and Granger, 2010). As manual simplification constitutes a time-consuming task, automating the simplification procedure could provide a more viable solution, especially when large corpora are involved. This study aims to contribute to this barely explored area of automatic TS for FL learning purposes (Section 2.1). We specifically focus on the natural language processing (NLP) technique of lexical

Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022), pages 98 - 110

simplification (Section 2.2), which will be used to adapt DDL activities to the needs of Dutch-speaking B2-level SFL learners. Apart from presenting this novel LS method (Section 3), the study also introduces a new type of human-based evaluation, distinguished by its particular pedagogical focus (Section 4).

2 Related Research

2.1 Text Simplification

Automatic TS, usually subdivided into syntactic and lexical simplification, is the computer-driven operation of "transforming a text into another text which, ideally conveying the same message, will be easier to read and understand by a broader audience" (Saggion, 2017). TS methods have been applied in a wide range of areas, where they have been proven useful for developing reading aids for children and people with cognitive disabilities (Rello et al., 2013; Watanabe et al., 2009), and for improving NLP tasks such as information extraction and machine translation in the form of a preprocessing step (Evans, 2011; Štajner and Popović, 2016).

The field of FL learning, however, has seen little attention being devoted to automatic TS, despite having a long tradition in manual TS (Shardlow, 2014a; Siddharthan, 2014). As one of the few existing studies related to automatic TS, Paetzold and Specia (2016) focus on unsupervised word embedding-based LS for non-native English speakers. Their aim is to satisfy the needs of this target audience by constructing a custom evaluation dataset based on a user study. Uchida et al. (2018) also present a language learning-oriented dataset for English, containing sentences taken from university textbooks. All B2+ words in those sentences were marked as complex, and substitution candidates were identified after manually revising a thesaurus-based selection of possible replacements. Finally, Martin et al. (2020) propose a controllable sentence simplification system based on Sequence-to-Sequence models, in which attributes such as sentence length and lexical complexity can be conditioned by the user. Although they do not specifically target FL learners, the controllable nature of their system can enable adjusting the simplification procedure to this target audience.

Even though TS is sometimes tackled as a generic task with a one-size-fits-all simplified output, it is agreed that different user groups often require different simplification methodologies (Martin et al., 2020; Shardlow, 2014a; Uchida et al., 2018). Datasets annotated by native speakers, for instance, have shown to be unsuitable for evaluating a TS system for non-native speakers, since word complexity as perceived by mother-tongue speakers does not correspond to word complexity for non-natives (Paetzold and Specia, 2016). Moreover, to further define this "non-native word complexity" (and to identify the simplification needs of FL learners in general), linguistic and pedagogical insights could be taken into account, such as Krashen's (1985) theory that learners acquire language when the input they are exposed to is comprehensible, but just somewhat beyond their current knowledge. It is, however, also important to highlight that manipulating corpus data is an intervention which needs to be undertaken with caution, since it may jeopardise the authentic character of the DDL activities (Boulton, 2009; Siddharthan, 2014).

Finally, by choosing Spanish as the target language, this study aims to continue the line of TS research which focuses on languages other than English. Since LexSis (the first implemented LS system for Spanish; Bott et al., 2012), Spanish has been included in more and more studies (Alarcón et al., 2021; Sheang, 2019; Saggion et al., 2015) and shared tasks (Yimam et al., 2018). The methodology and models presented in this study will contribute to further developing the Spanish TS domain.

2.2 Lexical Simplification

In LS, the goal is to replace "words in a given sentence in order to make it simple, without applying any modifications to its syntactic structure" (Paetzold and Specia, 2017, p. 549). LS systems can have different types of architectures, ranging from rule-based pipelines in which a predefined set of complex words is linked to synonyms (Devlin and Tait, 1998), over systems which exploit parallel corpora and the corresponding edit information (Biran et al. 2011), to word embedding approaches, which are designed to be less resource-dependent (Glavaš and Štajner, 2015). The LS process typically consists of four steps, presented below.

2.2.1 Complex Word Identification

A first important step within the CWI process is the definition of "complexity", as this concept may refer to absolute/objective or relative/agent-related complexity (North et al., 2022). While the former type refers to the linguistic properties of a word (e.g. word length, number of diphthongs and number of senses), the latter reports how individuals perceive a word based on their individual experiences or psycholinguistic factors (e.g. cognitive load and level of familiarity with a particular typography). In the field of CWI, however, a more general definition is adopted which combines elements from both complexity types. Therefore, when using the terms "complex" and "complexity" in this paper, we refer to the difficulty an individual may have in understanding a particular target word as a result of the target word's linguistic properties as well as factors belonging to the individual (North et al., 2022).

As for types of CWI methods, four categories can be discerned: threshold-based, lexicon-based, implicit and machine-learning assisted CWI. In threshold-based strategies, words are usually categorised as simple or complex based on word frequency. However, despite being intuitive and easy to implement, they lead to many simple words being unnecessarily labelled as complex (Shardlow, 2014b). Next, lexicon-based approaches look up words in human-curated lexicons, a strategy which yields good results but suffers from low coverage. Third, implicit CWI integrates this step into later stages of the simplification process, for example by only replacing words for which the top substitution candidate has a higher frequency (Glavaš and Štajner, 2015). In machine learning strategies, finally, classifiers such as support vector machines (Shardlow, 2013) or convolutional neural networks (Sheang, 2019) are trained based on training data with word embeddings, morphological data (word frequency, word length, number of syllables, etc.) and (psycho)linguistic information (age-of-acquisition value, part-of-speech [POS] tag, dependency relation, etc.) as features. As can be concluded from the 2018 CWI shared task (Yimam et al., 2018), of all strategies machine-learning assisted approaches obtain the best results.

Finally, it should be highlighted that as an alternative for CWI, the task of lexical complexity prediction (LCP) is also attracting more attention, as appears from the corresponding SemEval 2021 task (Shardlow et al., 2021). In LCP, a word's complexity is evaluated by assigning a value from a continuous scale, instead of providing a binary complex versus non-complex judgement as in CWI.

2.2.2 Substitution Generation

In the substitution generation step, candidate substitutions for the complex words are proposed. The generation can take two forms: linguistic database querying or automatic generation (Paetzold and Specia, 2017). In the former scenario, synonyms and/or other related words are looked up in human-curated databases such as WordNet (Fellbaum, 1998). Although the approach generally leads to suitable substitution candidates, both its coverage and potential to be extended to other languages are limited, since building such databases constitutes an expensive and time-consuming process (Shardlow, 2014b).

As for automatic generation, parallel resources such as English Wikipedia and Simple English Wikipedia can be exploited to automatically generate simplification pairs. Recently, the introduction of first static word embedding models such as word2vec (Mikolov et al., 2013) and later contextualised word embedding models such as BERT (Devlin et al., 2019) opened a whole new range of opportunities for the automatic generation of substitution candidates. Especially the masked language modelling feature of BERT and other models with transformer-based architectures has proven to bear great potential (Qiang et al., 2021; Zhou et al., 2019), as it is able to predict a masked word in a sentence such as (b) below while attending to both its left and right context. Introducing this sequence into the base, cased version of RoBERTa-BNE (Gutiérrez-Fandiño et al., 2021) results in *reducir* (0.09 probability; 'to reduce'), *cobrar* (0.05; 'to collect') and *bajar* (0.05; 'to decrease') as the top predictions. Finally, it should be noted that a hybrid approach, which combines embeddings with database information, can further improve performance (Paetzold and Specia, 2017).

(b) La planta local también permitirá <mask> los aranceles.

2.2.3 Substitution Selection

To determine which candidate substitutions fit the sentence context, a selection process needs to be carried out. The most common approaches to

substitute selection are sense labelling (Baeza-Yates et al., 2015), POS tag filtering (Aluísio and Gasperin, 2010) and semantic similarity filtering (Biran et al., 2011). In the resource-dependent sense labelling approach, substitution selection is modelled as a word sense disambiguation task, in which classification methods are used to check which candidates have the same sense label as the original complex word in a given database. Next, POS tag filtering consists in excluding all substitution candidates which do not have the same POS as the word to be simplified. For semantic similarity filtering, finally, the similarity between the substitution candidate and the word to be simplified is measured, after which all candidates which do not pass a certain threshold are removed.

2.2.4 Substitution Ranking

The fourth and final LS step encompasses ranking the selected candidates, for which three main strategies can be adopted: frequency-based, simplicity-based or machine learning-assisted (Paetzold and Specia, 2017). The first approach draws on the notion that the more frequent the word, the more familiar it will be to readers. Ranking from highest to lowest frequency is a very intuitive and straightforward operation, but the calculation of the frequency values can take many forms (token-based vs. lemma-based, raw frequencies vs. "transformed" logarithmic frequencies, extracting frequencies from different corpora, etc.). Simplicity measures and machine learning-assisted approaches expand on this frequency-based strategy by incorporating word frequency together with other features such as word length into, respectively, handcrafted metrics (Biran et al., 2011) or machine learning methods (Horn et al., 2014). The output of these metrics or machine learning models are designed to capture the complexity of words, after which candidates are ranked from lowest to highest complexity. Finally, substitution ranking can also be obtained by combining several ranking strategies and calculating one single average ranking score in the end (Glavaš and Štajner, 2015). In this case, aspects of the substitution selection stage (e.g. cosine similarity scores) can also be used as an additional ranker, instead of serving as a threshold-based selection parameter (Qiang et al., 2021).

3 Methodology

3.1 Setting

As mentioned in the introduction, a DDL-flavoured Spanish LSP course for Dutch-speaking B2-level learners is taken as the target setting, with business vocabulary as the specific purpose. The DDL character of the course is twofold: on the one hand, it includes a series of DDL activities in which learners analyse concordance lines of a selection of target vocabulary items they have to learn. On the other, the teacher of the course uses a corpus to create fill-the-gap exercises for another series of target vocabulary items. We specifically adopt a teacher-focused perspective on DDL, meaning that the goal of this study is to tailor the corpus data (i.e. the concordance lines and the sentences used for the fill-the-gap exercises) to the (lexical) needs of the B2-level target audience as perceived by the teacher. However, it is important to highlight that, in an ideal scenario, this operation is complemented by data on how SFL learners themselves perceive their lexical needs.

3.2 Datasets

Given the specific FL learning setting, we cannot make use of general benchmarking datasets such as ALEXSIS (Ferrés and Saggion, 2022). Instead, we generate datasets from a 11M tokenised, POS-tagged and lemmatised corpus containing newspaper articles on economics available within the pedagogically-oriented Spanish Corpus Annotation Project (SCAP; scap.ugent.be; Goethals, 2018). To arrive at the selection of target vocabulary items to be learned in the DDL activities, we first extract all candidate key vocabulary items from the corpus by means of a keyness calculation methodology (Gabrielatos, 2018). We use the Log Ratio metric (Hardie, 2014) to compare the frequency of each lemma in the economic corpus with its frequency in a 94M reference corpus and calculate the effect size of the difference in frequencies. Next, only the candidate items with a statistically significant effect size according to the Bayesian Information Criterion (values ≥2; Wilson, 2013) are maintained. Finally, the resulting list is ranked from highest to lowest keyness and all items are assigned a difficulty level by the dictionary-based difficulty level classifier of SCAP, after which the top 25 nouns of a C1 level (i.e. the proficiency level to be acquired) are selected as the final set of target vocabulary items.

Lemma (Log Ratio)	Original	Selection	≥1 CW	1 CW (noun/verb)	Changed
Arancel (7.72)	837	65	46	16	16
Desaceleración (7.68)	272	39	23	3	3
Competitividad (7.34)	699	83	58	16	14
Depreciación (7.31)	258	36	26	12	12
Competidor (6.61)	1272	113	67	27	25
Revalorización (6.34)	313	58	24	12	12
Liberalización (5.71)	347	25	15	3	3
Puja (5.59)	311	51	27	11	11
Remuneración (559)	645	93	57	23	21
Robótica (5.47)	150	24	12	3	3
Carburante (5.22)	173	25	14	10	10
Anunciante (5.14)	169	27	20	6	6
Canje (5.11)	232	31	24	13	13
Cancelación (4.73)	364	37	21	11	11
Emprendedor (4.62)	389	49	25	9	8
Encarecimiento (4.57)	128	16	13	4	4
Fomento (4.52)	167	19	15	5	5
Dígito (4.5)	353	44	18	12	12
Solvencia (4.23)	673	67	48	16	14
Factoría (4.17)	331	24	16	8	8
Plusvalía (4.1)	412	45	25	12	11
Homologación (4.05)	140	15	12	3	3
Normativa (3.73)	1680	148	112	31	29
Captación (3.71)	274	45	29	16	15
Provisión (3.59)	881	117	83	23	22
	11 470	1296	830	305	291

Table 1: Dataset statistics.

For each of the 25 selected target items, all sentences in which the lemma of the item occurs are then extracted from the 11M economic corpus.

3.3 Example Selection

Prior to performing LS on them, the datasets can already be brought one step closer to the needs of FL learners by filtering out unsuitable sentences, an intervention which has often been neglected in previous research (Pilán et al., 2016). This filtering consists in applying a series of criteria sentences need to comply with in order to be comprehensible in isolation (Kilgarriff, 2009). To perform this automatic sentence selection for FL learning purposes, we develop an example sentence selection methodology based on the HitEx framework for Swedish (Pilán et al., 2016). A complete overview of the criteria and the definition of the corresponding parameters is to be found in Appendix A. Table 1 presents the dataset sizes before ("Original") and after ("Selection") applying the example selection methodology.

3.4 Lexical Simplification

3.4.1 Complex Word Identification

To tailor the CWI strategy to the teacher-focused DDL setting, we build a classifier which is able to predict, for all words in a given sentence, different complexity labels based on the proficiency levels described in the Common European Framework of Reference (CEFR). To this end, we first build a lexicon based on the PortaVoces (Buyse et al., 2005) and Thematische Woordenschat (Navarro and Navarro Ramil, 2010) SFL vocabulary learning resources for Dutch-speaking learners, whose contents we combine into a single lexicon of 2823 A-level lemmas, 1557 B1 lemmas, 1998 B2 lemmas and 3584 C lemmas. Drawing on insights from FL learning research and taking into account criteria ranging from frequency to learner-specific features such as familiarity, these vocabulary learning resources are often taken as reference points in many SFL curricula for Dutch-speaking learners, and can thus serve as an indicator for the lexical needs of SFL students at a given stage of their language learning careers. In other words, the output of the classifier can help teachers identify

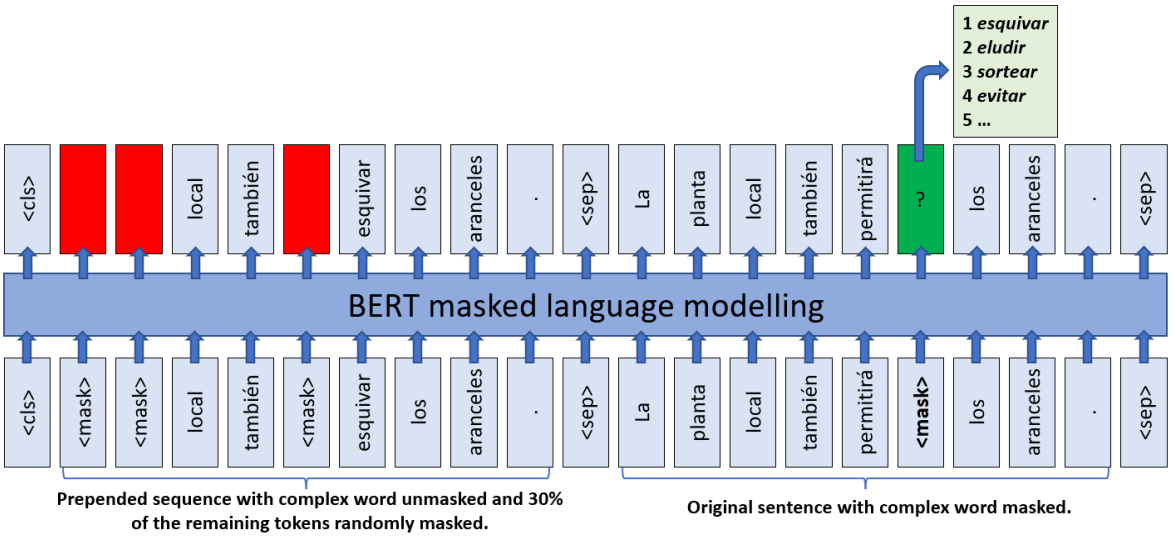

Figure 1: Illustration of masked language modelling using next sentence prediction (the special "<cls>" and "<sep>" tokens are used to initiate the input sequence and separate the items of the sentence pair, respectively). For the masked complex word *esquivar* in the original sentence, the top candidates from the probability distribution of the model's vocabulary (50 262 entries) are collected. For the randomly masked tokens in the prepended sentence no predictions are generated.

potentially complex words for a B2 target audience.

To train the classifier, we fine-tune the base, cased version of RoBERTa-BNE for token classification, thus adopting a machine learning-assisted approach. Apart from the pretrained model weights, training the token classifier also requires labelled sequences as input. To obtain this labelled data, all sentences from the SCAP corpora in which every content word has a matching entry in the previously elaborated lexicon are gathered (all non-content words receive the A label) and split into a training (1 511 387 sentences), validation and test set (both 188 924 sentences). We train the token classifier for 3 epochs with a learning rate of $2e^{-5}$, AdamW as the optimiser and a weight decay of 0.01, and obtain a 0.9983 macro F1 score on the test set.

The classifier [1], which thus offers unlimited coverage, is then applied to the 25 datasets. Every C-labelled token is identified as complex, unless it has a Dutch cognate[2] or appears amongst the 25 target items. As cognates have shown to be easily processed and learned by foreign language learners (De Groot and Keijzer, 2000), they usually possess low lexical complexity and thus should not be identified as such. In total, 64% of the selected sentences contain one or more potentially complex content words for a B2 target audience (see "≥1 CW" in Table 1), which highlights the need for LS in a FL learning setting. Next, the column "1 CW (noun/verb)" presents the number of sentences in which exactly one complex noun or verb was found: these sentences constitute the final dataset to be used in this study, as a one-per-sentence setup is most suitable to measure the exact impact of the simplification procedure.

3.4.2 Substitution Generation

To generate substitution candidates, we build upon the line of research of Qiang et al. (2021). Their automatic generation method, which offers potentially unlimited coverage, exploits the masked language modelling and next sentence prediction features of BERT models to get the probability distribution of the model's vocabulary $p(\cdot|S\backslash\{w\})$ corresponding to the masked word w in sentence S. The input introduced into the BERT model is a sequence pair: the original sequence with the complex word being masked, preceded by the exact same sequence, but now with the complex word unmasked and a given percentage of the remaining tokens randomly masked (see Figure 1).

In this study we use RoBERTa-BNE, apply 0.3 as the ratio for randomly masking tokens in the prepended sequence, and bring the top 25

[1] huggingface.co/JasperD-UGent/roberta-base-bne-complexity-classifier-v1

[2] A pair of words in different languages which are related and look similar, or which have the same origin. Spanish – Dutch example: 'proyecto' – 'project' (EN 'project').

Number	Criterion
1	Probability value obtained from masked language modelling
2	Language model score (Qiang et al., 2021)
3	Lemma frequency in SCAP corpora
4	Token frequency in SUBTLEX-ESP (Cuetos et al., 2011)
5	Cosine similarity with complex word using word2vec and fastText (fasttext.cc) pretrained static word embeddings
6	Cosine similarity with complex word using RoBERTa-BNE contextualised word embeddings

Table 2: Substitution ranking criteria.

candidates from the probability distribution to the substitution selection phase. As a novel aspect, we also add contextual information to the sequence in the form of the previous and following sentence of the corpus text from which the target sentence was taken. This adjustment is particularly useful in cases where the complex word is situated at the sentence-final position. In fact, when the previous and following sentence are not added, for 17 of the 25 target sentences with the complex word at the sentence-final position no suitable substitution candidate is found, because almost all of the 25 suggestions appeared to be punctuation marks. With the extra contextual information being added, this number decreases to 7 out of 25.

3.4.3 Substitution Selection

The first component of our substitution selection strategy is a POS filter, which excludes every candidate whose POS tag does not correspond to the POS tag of the complex word. Next, given the morphological richness of the Spanish language, an additional filter is applied: using spaCy's (spacy.io) v3.3.1 morphologiser ("es_core_news_lg" model), the morphological features of the complex word are determined, after which all substitution candidates without matching features are discarded. The feature set consists of gender (masculine, feminine) and number (singular, plural) for nouns, and mood (indicative, subjunctive, imperative), person (1, 2, 3), number (singular, plural) and verb form (finite, infinitive, past participle, gerund) for verbs. To tailor the selection strategy to our FL learning context, a third component replaces the complex word by the substitution candidates, introduces each of these modified sentences into the CWI classifier (see Section 3.4.1) and eliminates every candidate for which the classifier predicted C as the complexity label. Finally, all morphological variants of the complex word are also excluded, as well as words whose lemma appears in the target sentence.

3.4.4 Substitution Ranking

In the last phase, all remaining substitution candidates are ranked based on the criteria described in Table 2. The six individual rankings are averaged to obtain one single final ranking.

4 Results

4.1 Evaluation of Suitability

The LS method changed 291 of the 305 complex words included in the final datasets (see "Changed" in Table 1), corresponding to a 95.41% score on the "changed" metric (Horn et al., 2014). Apart from the 7 sentence-final cases mentioned earlier, the main reason for which no candidates are found is that the morphological filter appears to be too strict for, amongst others, sentence structures which allow both singular and plural replacements for a complex noun. If in such case none of the generated candidates shares the number of the complex noun, no candidates pass the selection phase.

To evaluate the 291 simplified sentences, a novel evaluation method with SFL teachers as evaluators is applied, which is in line with the teacher-focused DDL perspective we adopted. After presenting them the background information explained in Section 3.1, we ask them to indicate, for each of the 3 top-ranked substitution candidates of a given sentence, if replacing the complex word by the candidate results in a better, similar or worse example sentence (see Table 3). For each sentence, responses from 3 different teachers are collected (2619 annotations in total). Importantly, in the instructions we explicitly mention that changes in

Sentence	Substitution	Better	Similar	Worse
La planta local también permitirá **esquivar** los aranceles.	evitar			
	escapar			
	olvidar			

Table 3: Illustration of the annotation task, with *aranceles* as the vocabulary item to be learned. Teachers are asked to indicate if the substitutions for the complex word *esquivar* result in a more ("Better"), equally ("Similar") or less ("Worse") suitable example sentence to be used in the setting described in Section 3.1.

Metric	All	R1	R2	R3
IAA	.26	.28	.24	.26
% 3/3 agreement	35.4	36.08	32.65	37.46
% better (≥2/3)	33.68	37.46	35.74	27.84
% similar (≥2/3)	11.57	12.37	12.37	9.97
% worse (≥2/3)	44.9	40.21	42.96	51.55
% suitable (binary)	38.95	42.96	40.89	32.99

Table 4: Performance results. "IAA" reports the inter-annotator agreement as measured by Fleiss' Kappa, "≥2/3" refers to agreement between at least 2 participants, and "binary" refers to the results of classifying the annotations into suitable (at least 2 "better" annotations or 1 "better" and 2 "similar" annotations) and non-suitable (all other cases) simplifications. Percentages for "binary" correspond to the precision metric of Horn et al. (2014).

meaning should not be taken into account during evaluation, as long as the end result is a pedagogically suitable example sentence. This enabled us to analyse if FL learning as the target setting affects the importance of the meaning preservation criterion (see Section 4.2).

Table 4 presents the main descriptive statistics taken from the experiment, with the "All" column reporting the results for all annotations combined and the three "R" columns showing the results broken down according to ranking position. Overall, the results show moderate agreement between the teachers (IAA of 0.26 and 35.4% of the sentences annotated equally by all 3 annotators), without any considerable differences between the ranking groups. Although these statistics suggest that evaluating the added value of LS for FL learning purposes is not a straightforward task, we consider the number of times in which at least two of the three teachers coincide (90.15% across all labels in "All") as an indication that agreement is sufficiently high to draw valuable conclusions.

First of all, the statistics reveal mixed performance results: when converting the annotations into a binary classification, 42.96% of the "R1" simplifications come out as suitable, a score which highlights both the potential of the method and its room for improvement. Next, the ranking component seems to perform well, as the first-ranked substitutions are considerably more annotated as better and considerably less as worse compared to "R3". Third, the CWI classifier can still be improved: despite being found pedagogically suitable, 11.57% of the sentences are not evaluated as more simple compared to the original text, which indicates that the classifier labelled an equally complex word as more simple.

4.2 Evaluation of Meaning Preservation

For this supplementary analysis, we annotate all substitutions according to meaning preservation (see Table 5). The results suggest that meaning preservation is not a sine qua non for successful LS in a FL learning context, as replacing the complex word by an unrelated or even opposite concept does not prevent 45 sentences from being found suitable. However, meaning preservation does remain a key criterion, as is evidenced by the majority of the suitable sentences having the same meaning as the original word (189 sentences), or at least being related to it (106 sentences). Finally, it should be noted that substitutions which share the meaning of the complex word do not necessarily result in better example sentences. Many of those cases can be linked to the "idiomaticity" of the simplified sentence, which comes to the fore as an additional important criterion. This is evidenced in the results for sentence (c) in Table 5: despite being semantically equal to *aliviar*, none of the 3 teachers

Label	Total	Suitable (binary)	Example
Preserved	311	189	(c) Los aranceles pueden **aliviar** la presión que sufren los fabricantes. ('Tariffs can **alleviate** pressure on manufacturers.') → *reducir* ('to reduce')
Related	258	106	(d) Estoy muy contento con los 100.000 millones de dólares en aranceles que llenan nuestras **arcas**. ('I am very happy with the $100 billion in tariffs that fill our **treasury**.') → *cuentas* ('bank accounts')
Unrelated	257	42	(e) Las importaciones de **baldosas** chinas se cargarán con aranceles del 30% al 69%. ('Chinese **tile** imports will be charged tariffs from 30% to 69%.') → *alfombras* ('carpets')
Opposite	47	3	(f) Sus principales competidores presentan **retrocesos** anuales muy fuertes. ('Its main competitors show very strong annual **declines**.') → *incremento* ('increase')

Table 5: Overview of meaning preservation annotations.

annotated the substitution candidate *relajar* ('to relax') as "better", because *relajar la presión* is a rather uncommon collocation. A similar case is that of multiword expressions, in which often only one formulation sounds idiomatic (e.g. *agrupación de acciones* → ?*unión de acciones*; 'consolidation of shares → ?union of shares').

4.3 Evaluation of Grammaticality

Finally, a manual grammaticality check of the output reveals that 50 simplifications result in incorrect sentences, with preposition issues being the main cause. The *escapar* prediction in Table 3 is such an example, as this verb needs to be followed by the preposition *de*. Non-surprisingly, virtually all of these substitutions were annotated as "worse", which suggests that, in a future version of the method, excluding non-grammatical simplifications alone can lead to considerable increases in performance.

5 Discussion and Conclusion

In this study, we presented a Spanish LS method tailored to FL learning as the target setting. By simplifying all potentially complex words except the vocabulary item to be studied, the method adapts DDL activities to a given proficiency level while also taking into account the language acquisition theory of providing comprehensible input which is just somewhat beyond the current knowledge of the target audience (Krashen, 1985). As we specifically focused on SFL learners with Dutch as their mother tongue, the findings of this study primarily contribute to LS for this particular language combination. However, if equivalent resources (graded vocabulary learning resources, language models, etc.) are available, the methodological design of the LS pipeline can be applied to any language.

To analyse performance, a new type of human-based evaluation was carried out, which revealed the potential of the system (43% of the top-ranked predictions being found suitable) and suggested that meaning preservation is an important though not always necessary condition for obtaining both successfully simplified and pedagogically suitable example sentences. However, the results also showed that the custom-made CWI classifier leaves room for improvement, that many simplifications lack idiomaticity and that the substitution selection component is not yet able to exclude all non-grammatical replacements.

To overcome these limitations in the future, we first of all aim to further develop the CWI classifier and evaluate it in a separate experiment. Next, to arrive at grammatically correct and more idiomatic substitution candidates, we also plan to implement "typicality"-related measures (e.g. association measures based on co-occurrence data; Gries, 2013) into the substitution selection and ranking components. Finally, we will study the addition of weights to the ranking calculation, in order to balance the relative importance of the criteria.

As a final observation, it should be highlighted that the teacher-focused DDL perspective adopted in this study also comes with its limitations. The expert knowledge of teachers and the contents of scientifically grounded vocabulary learning resources (as the ones used in this study to train the CWI classifier) can be valuable indicators of lexical complexity, but they often exhibit a lack of systematicity and do not capture how FL learners perceive lexical complexity themselves (Tack et al., 2021). Therefore, in future research we will also collect "non-native data" and integrate them into the LS methodology. The pedagogically oriented evaluation, for instance, could be performed by SFL learners in the form of a best-worst scaling experiment in which learners have to indicate the best and worst item in a set of four versions of the same sentence (the original sentence plus three sentences with the complex word being replaced by the three top-ranked substitution candidates).

Acknowledgements

The first author acknowledges the support from the IVESS project (file number 11D3921N), a PhD fellowship funded by the Research Foundation – Flanders (FWO). The second author acknowledges support from the project Context-aware Multilingual Text Simplification (ConMuTeS) PID2019-109066GB-I00/AEI/10.13039/501100011033 individual grant awarded by the Ministerio de Ciencia, Innovación y Universidades (MCIU) and by the Agencia Estatal de Investigación (AEI) of Spain. Finally, both authors also wish to express their sincere gratitude to the reviewers for their valuable comments.

References

Alarcón Rodrigo, Moreno Lourdes, & Martínez Paloma (2021). Exploration of Spanish Word

Embeddings for Lexical Simplification. *Proceedings of the First Workshop on Current Trends in Text Simplification (CTTS 2021)*, 29–41.

Aluísio Sandra, & Gasperin Caroline (2010). Fostering Digital Inclusion and Accessibility: The PorSimples project for Simplification of Portuguese Texts. *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, 46–53.

Baeza-Yates Ricardo, Rello Luz, & Dembowski Julia (2015). CASSA: A Context-Aware Synonym Simplification Algorithm. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1380–1385.

Bax Stephen (2003). CALL—past, present and future. *System*, *31*(1), 13–28.

Biran Or, Brody Samuel, & Elhadad Noémie (2011). Putting it Simply: A Context-Aware Approach to Lexical Simplification. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 496–501.

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

Boulton Alex (2009). Data-driven learning: reasonable fears and rational reassurance. *Indian Journal of Applied Linguistics*, 2009, *35*(1), 1–28.

Boulton Alex, & Cobb Tom (2017). Corpus Use in Language Learning: A Meta-Analysis. *Language Learning*, *67*(2), 348–393.

Boulton Alex, & Vyatkina Nina (2021). Thirty years of data-driven learning: Taking stock and charting new directions over time. *Language Learning & Technology*, *25*(3), 66–89.

Buyse Kris, Delbecque Nicole, & Speelman Dirk (2005). *Portavoces: Thematische woordenschat Spaans*. Wolters Plantyn.

Chambers Angela (2019). Towards the corpus revolution? Bridging the research–practice gap. *Language Teaching*, *52*(4), 460–475.

Cuetos Fernando, Glez-Nosti Maria, Barbon Analia, & Brysbaert Marc (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *PSICOLOGICA*, *32*(2), 133–143.

De Groot Annette M.B., & Keijzer Rineke (2000). What Is Hard to Learn Is Easy to Forget: The Roles of Word Concreteness, Cognate Status, and Word Frequency in Foreign-Language Vocabulary Learning and Forgetting. *Language Learning*, *50*(1), 1–56.

Devlin Jacob, Chang Ming-Wei, Lee Kenton, & Toutanova Kristina (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *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)*, 4171–4186.

Devlin Siobhan, & Tait John (1998). The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic Databases*, *1*, 161–173.

Evans Richard (2011). Comparing methods for the syntactic simplification of sentences in information extraction. *Literary and Linguistic Computing*, *26*(4), 371–388.

Fellbaum Christiane (Ed.). (1998). *WordNet: An Electronic Lexical Database*. The MIT Press.

Ferrés Daniel, & Saggion Horacio (2022). ALEXSIS: A Dataset for Lexical Simplification in Spanish. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, 3582–3594.

Gabrielatos Costas (2018). Keyness analysis: Nature, metrics and techniques. In Charlotte Taylor & Anne Marchi (Eds.), *Corpus Approaches To Discourse* (pp. 225–258). Routledge.

Gilquin Gaëtanelle, & Granger Sylviane (2010). How can data-driven learning be used in language teaching? In *The Routledge Handbook of Corpus Linguistics*. Routledge.

Glavaš Goran, & Štajner Sanja (2015). Simplifying Lexical Simplification: Do We Need Simplified Corpora? *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)*, 63–68.

Goethals Patrick (2018). Customizing vocabulary learning for advanced learners of Spanish. In Read, Timothy and Sedano Cuevas, Beatriz and Montaner-Villalba, Salvador (Eds.), *Technological innovation for specialized linguistic domains: Languages for digital lives and cultures, proceedings of TISLID'18* (pp. 229–240). Éditions Universitaires Européennes.

Granger Sylviane, Kraif Olivier, Ponton Claude, Antoniadis Georges, & Zampa Virginie (2007). Integrating learner corpora and natural language processing: A crucial step towards reconciling technological sophistication and pedagogical effectiveness. *ReCALL*, *19*(3), 252–268.

Gries Stefan (2013). 50-something years of work on collocations: What is or should be next ….

International Journal of Corpus Linguistics, *18*(1), 137–166.

Gutiérrez-Fandiño Asier, Armengol-Estapé Jordi, Pàmies Marc, Llop-Palao Joan, Silveira-Ocampo Joaquín, Carrino Casimiro Pio, Gonzalez-Agirre Aitor, Armentano-Oller Carme, Rodriguez-Penagos Carlos, & Villegas Marta (2021). *MarIA: Spanish Language Models*. Computer Science Repository, arXiv:2107.07253, Version 5.

Hardie Andrew (2014). Log ratio: An informal introduction. *ESRC Centre for Corpus Approaches to Social Science (CASS)*, 1–2.

Horn Colby, Manduca Cathryn, & Kauchak David (2014). Learning a Lexical Simplifier Using Wikipedia. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 458–463.

Jablonkai Reka, & Csomay Eniko (Eds.). (2022). *The Routledge handbook of corpora and English language teaching and learning*. Routledge.

Johns Tim (1990). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *CALL Austria*, *10*, 14–34.

Kilgarriff Adam (2009). Corpora in the classroom without scaring the students, *Proceedings from the 18th International Symposium on English Teaching*.

Krashen Stephen (1985). *The input hypothesis: Issues and implications*, Vol. 1. London: Longman.

Laufer Batia, & Nation Paul (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics*, *16*(3), 307–322.

Martin Louis, de la Clergerie Éric, Sagot Benoît, & Bordes Antoine (2020). Controllable Sentence Simplification. *Proceedings of the 12th Language Resources and Evaluation Conference*, 4689–4698.

Mikolov Tomas, Chen Kai, Corrado Greg, & Dean Jeffrey (2013). *Efficient Estimation of Word Representations in Vector Space*. Computer Science Repository, arXiv:1301.3781, Version 1.

Navarro José María, & Navarro Ramil Axel (2010). *Thematische woordenschat Spaans*. Intertaal.

North Kai, Zampieri Marcos, & Shardlow Matthew (2022). Lexical Complexity Prediction: An Overview. *ACM Computing Surveys*, 3557885.

Paetzold Gustavo, & Specia Lucia (2016). Unsupervised Lexical Simplification for Non-Native Speakers. *Proceedings of the AAAI Conference on Artificial Intelligence*, *30*(1).

Paetzold Gustavo, & Specia Lucia (2017). A Survey on Lexical Simplification. *Journal of Artificial Intelligence Research*, *60*, 549–593.

Pilán Ildikó, Volodina Elena, & Borin Lars (2016). Candidate sentence selection for language learning exercises: From a comprehensive framework to an empirical evaluation. *Revue Traitement Automatique Des Langues*, *57*(3), 67–91.

Qiang Jipeng, Li Yun, Zhu Yi, Yuan Yunhao, Shi Yang, & Wu Xindong (2021). LSBert: Lexical Simplification Based on BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3064–3076.

Rello Luz, Baeza-Yates Ricardo, Bott Stefan, & Saggion Horacio (2013). Simplify or help?: Text simplification strategies for people with dyslexia. *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility - W4A '13*, 1.

Saggion Horacio (2017). *Automatic Text Simplification*. Springer International Publishing.

Saggion Horacio, Štajner Sanja, Bott Stefan, Mille Simon, Rello Luz, & Drndarevic Biljana (2015). Making It Simplext: Implementation and Evaluation of a Text Simplification System for Spanish. *ACM Transactions on Accessible Computing*, *6*(4), 1–36.

Shardlow Matthew (2013). A Comparison of Techniques to Automatically Identify Complex Words. *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, 103–109.

Shardlow Matthew (2014a). A Survey of Automated Text Simplification. *International Journal of Advanced Computer Science and Applications*, *4*(1).

Shardlow Matthew (2014b). Out in the Open: Finding and Categorising Errors in the Lexical Simplification Pipeline. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 1583–1590.

Shardlow Matthew, Evans Richard, Paetzold Gustavo, & Zampieri Marcos (2021). SemEval-2021 Task 1: Lexical Complexity Prediction. *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, 1–16.

Sheang Kim Cheng (2019). Multilingual Complex Word Identification: Convolutional Neural Networks with Morphological and Linguistic Features. *Proceedings of the Student Research Workshop Associated with RANLP 2019*, 83–89.

Siddharthan Advaith (2014). A survey of research on text simplification. *ITL - International Journal of Applied Linguistics*, *165*(2), 259–298.

Štajner Sanja, & Popovic Maja (2016). Can Text Simplification Help Machine Translation? *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, 230–242.

Tack Anaïs, Desmet Piet, Fairon Cédrick, & François Thomas (2021). *Mark My Words! On the Automated Prediction of Lexical Difficulty for Foreign Language Readers*.

Uchida Satoru, Takada, Shohei & Arase Yuki (2018). CEFR-based Lexical Simplification Dataset. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Watanabe Willian Massami, Junior Arnaldo Candido, Uzêda Vinícius Rodriguez, Fortes Renata Pontin de Mattos, Pardo Thiago Alexandre Salgueiro, & Aluísio Sandra (2009). Facilita: Reading assistance for low-literacy readers. *Proceedings of the 27th ACM International Conference on Design of Communication - SIGDOC '09*, 29.

Wilson Andrew (2013). Embracing Bayes factors for key item analysis in corpus linguistics. In Markus Bieswanger & Amei Koll-Stobbe (Eds.), *New Approaches to the Study of Linguistic Variability* (pp. 3–11). Peter Lang.

Yimam Seid Muhie, Biemann Chris, Malmasi Shervin, Paetzold Gustavo, Specia Lucia, Štajner Sanja, Tack Anaïs, & Zampieri Marcos (2018). A Report on the Complex Word Identification Shared Task 2018. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 66–78.

Zhou Wangchunshu, Ge Tao, Xu Ke, Wei Furu, & Zhou Ming (2019). BERT-based Lexical Substitution. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3368–3373.

Appendix A. Example Selection Criteria

Table 6 in this appendix includes the criteria and values applied in the example sentence selection methodology (Section 3.3). Custom criteria which have been added to take into account the particularities of Spanish as the target language are indicated as "(CUSTOM)". The tools used to process the corpus are the SCAP tokeniser, POS tagger (list of POS tags available at scap.ugent.be/static/SCAP_POS-tags_details.pdf) and lemmatiser, as well as spaCy's v3.3.1 dependency parser ("es_core_news_lg" model).

Criterion	Values applied
Search term	
Number of matches	= 1
Position of search term	Anywhere in the sentence
Well-formedness	
Dependency root	= 1
Ellipsis	Not allowed: sentence has to contain a finite verb, a subject and a verbal root
Incompleteness	Sentence has to start with a capital letter and end with a punctuation mark
Non-lemmatised tokens	≤ 5% of the tokens (non-lemmatised tokens are identified as tokens without a matching entry in the SCAP lemma list)
Non-alphabetical tokens	≤ 5% of the tokens (non-alphabetical tokens are identified as tokens which have been assigned the "SYM" POS tag)
Subject type (CUSTOM)	Sentence has to contain an explicit subject, not an implicit subject integrated into the verb form
Context independence	
Structural connective in isolation	Not allowed: sentence cannot contain connectives in sentence-initial position unless it consists of more than one clause
Pronominal anaphora	Not allowed: sentence cannot contain tokens which have been assigned the "DM" tag or which have *eso*, *esto*, *aquello* or *tal* as their lemma
Adverbial anaphora	Not allowed: sentence cannot contain time or location adverbs which behave anaphorically, such as *entonces* ('then')
L2 complexity	
L2 complexity in CEFR level	This criterion is excluded, as complex words are supposed to be identified in the CWI step and replaced by simpler alternatives
Additional structural criteria	
Negative formulations	Not allowed: sentence cannot contain tokens which have been assigned the "CCNEG" or "NEG" tag
Interrogative sentence	Not allowed: sentence cannot contain question marks
Direct speech	Allowed
Answer to closed questions	Not allowed: sentence cannot start with adverbs or interjections such as *sí* ('yes') or *no* ('no') preceded and followed by delimiters such as commas
Modal verbs	Allowed
Sentence length	≤ 40 tokens
Additional lexical criteria	
Difficult vocabulary	This criterion is excluded, as complex words are supposed to be identified in the CWI step and replaced by simpler alternatives
Word frequency	No limitations
Sensitive vocabulary	Not allowed: sentence cannot contain tokens which appear in a self-compiled list of swear words
Typicality	No limitations
Proper names	Not allowed: sentence cannot contain tokens which have been assigned the "XP" tag
Abbreviations	Not allowed: sentence cannot contain tokens which have been assigned the "ACRNM" or "UMMX" tag

Table 6: Example selection criteria.