Lexical Complexity Prediction and Lexical Simplification for Catalan and Spanish: Resource Creation, Quality Assessment, and Ethical Considerations*

Horacio Saggion[1], Stefan Bott[1], Sandra Szasz[1], Nelson Pérez[2], Saúl Calderón[2], Martín Solís[2]
[1]Universitat Pompeu Fabra (Barcelona, Spain),
[2]Instituto Tecnológico de Costa Rica (Cartago)

Correspondence: horacio.saggion@upf.edu

Abstract

Automatic lexical simplification is a task to substitute lexical items that may be unfamiliar and difficult to understand with easier and more common words. This paper presents the description and analysis of two novel datasets for lexical simplification in Spanish and Catalan. This dataset represents the first of its kind in Catalan and a substantial addition to the sparse data on automatic lexical simplification which is available for Spanish. Specifically, it is the first dataset for Spanish which includes scalar ratings of the understanding difficulty of lexical items. In addition, we present a detailed analysis aiming at assessing the appropriateness and ethical dimensions of the data for the lexical simplification task.

1 Introduction

Various types of readers may have problems with the understanding of written text. These groups include, among others, language learners (Rets and Rogaten, 2021), children (Javourey-Drevet et al., 2022), people with cognitive disabilities (Licardo et al., 2021), and people with a generally low level of reading proficiency. On the other hand, some texts are written in a style that makes it hard to understand the content, for example, by being written in a difficult style or by the use of vocabulary that is unknown to the reader. Universal access to information in the form of understandable text is not only a desirable service to citizens, but it is a citizens' right that has started to be recognized by international institutions and national legislation in the last years.[1] Apart from recognized rights, there are also very serious general concerns about inclusion, the principled functioning of democracy and democratic institutions, as well as the right of citizens to be protected from political and economic abuse (Rennes, 2022; Johannessen et al., 2017). Democratic processes have serious shortcomings when certain groups are denied informed participation, just because essential information is not available in a form they can understand.

A common and effective, although costly strategy to remedy this is to adapt these texts by specialized human editors (Nomura et al., 2010). This approach is limited by the vast amount of texts which are available today. A much more economic alternative is to adapt texts automatically with computational algorithms. This Natural Language Processing task is known as *Automatic Text Simplification* (ATS) (Saggion, 2017). ATS may involve several transformations including sentence splitting, grammatical transformation or the exclusion of overly detailed content. *Automatic Lexical Simplification* (LS) (Shardlow, 2014a; Paetzold and Specia, 2017) is a well-defined sub-task of ATS, which only aims at finding i) words that are complex and should be simplified and ii) simpler substitutes for these complex words. These two sub-tasks are referred to as *Complex Word Identification* (CWI) (Zampieri et al., 2017) and *Substitute Generation* (SG). Finally, *Substitute Ranking* (SR) and *Substitute Selection* (SS) ensure that the best candidate(s) produced by SG are selected for the output. A similar task to CWI is *Lexical Complexity Prediction* (LCP) (Shardlow et al., 2021), which outputs an estimate for the lexical difficulty of each target unit, instead of only making a binary decision on whether a word should be substituted or not.

The availability of data that represent LCP and LS is a prerequisite for the development or fine-tuning of models to effectively handle these tasks. Data is needed to evaluate and benchmark them. As in the case of many other NLP tasks most work has been done for English. For Spanish the availability of suitable data is low and in the case of Catalan, it is, to the best of our knowledge, nonexistent. The work we present here aims to remedy this situation.

*This is a considerable modification to a preliminary version archived in arXiv.

[1]For example the plain writing act of 2010: https://www.govinfo.gov/app/details/PLAW-111publ274

Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024), pages 82–94

The main contributions of this paper are:

- We provide a detailed description of two datasets for *Lexical Simplification* and *Lexical Complexity Prediction* for Spanish and Catalan.
- We describe in full the data compilation process and provide a statistical description of the datasets.
- We assess the quality of the dataset for the lexical simplification task and consider ethical implications of the data.

This paper is organized as follows: Section 2 overviews of the state of the art in LS and describes existing comparable resources for Iberian Romance languages; Section 3 details the method for data collection and annotation; Section 4 describes the quality analysis of the data. In Section 5 we raise ethical concerns in LS while in Section 6 we close the paper with a discussion and future work.

2 Related Work

Foundational work on Lexical Simplification was developed for English by Devlin and Tait (1998) who used Wordnet to identify synonyms for target words and word frequencies from the Kucera-Francis psycho-linguistic database for synonyms ranking. This initial approach was followed by corpus-based approaches that used Language Models (De Belder and Moens, 2010) or Wikipedia (Biran et al., 2011; Yatskar et al., 2010; Horn et al., 2014). Deep learning approaches were explored by Glavaš and Štajner (2015) with an unsupervised approach for LS based on current distributional lexical semantics modelling, while Paetzold and Specia (2017) combine learned substitutions from a corpus using neural networks. Qiang et al. (2020) presented LS-BERT, a LS framework that uses a pre-trained 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.

Regarding LS in Spanish, few approaches are reported in the literature. They can be classified as: (i) knowledge-based approaches which rely on "curated" lists of synonyms and corpora to propose and rank synonyms by relying on frequency and other word characteristics (Bott et al., 2012a; Baeza-Yates et al., 2015; Ferrés et al., 2017a); (ii) translation-based approaches which cast simplification as translation (Stajner (2014) and Štajner et al. (2019) implicitly learn simplification rules) and (iii) current transformer-based approaches (Alarcón et al., 2021) which achieve a state of the art performance. In the context of the TSAR 2022 Lexical Simplification challenge (Saggion et al., 2022), several approaches have been proposed, mostly based on pre-trained language models. Controllable lexical simplification was introduced for English in Sheang et al. (2022) achieving state of the art in multilingual settings in Sheang and Saggion (2023). Contrary to current methods, Stajner et al. (2023) presents a light-weight text simplifier for Spanish claiming that it achieves good performance without the cost associated with current architectures.

In the earlier approaches to *Lexical Simplification*, CWI was treated as an implicit part of the simplification pipeline, even though it was often treated as a modular pipeline component (Carroll et al., 1998; Shardlow, 2014b; Bott et al., 2012b). Shardlow (2013) is the first work which frames CWI as an independent task "which may seem intuitively easy, but in reality is quite difficult and rarely performed". He presents a dedicated CWI classifier using Support Vector Machines. In 2016 and 2017 two shared tasks were held at SemEval and BEA (Paetzold and Specia, 2016; Yimam et al., 2018b) on CWI. The 2017 task also included an estimation of the probability of a target word being complex, which was a step towards *Lexical Complexity Prediction*, but it did not require a direct estimation of *Lexical Complexity*. ALexS (Ortiz-Zambrano and Montejo-Ráez, 2020) was a CWI competition for Spanish which unfortunately seldom attracted participants. In 2021, a SemEval shared task invited contributions for LCP (Shardlow et al., 2021), which now predicted grades of LC directly. This last task was based on previous work in Shardlow et al. (2020). The 2024 Multilingual Lexical Simplification Pipeline shared task (Shardlow et al., 2024) is a new challenge covering aspects of LCP and LS.

CWI and LCP has been tackled with the use of SVMs (Shardlow, 2013), decision trees (Quijada and Medero, 2016), random forests (Ronzano et al., 2016) and neural networks (Gillin, 2016). Recent approaches include the use of transformer models (Yaseen et al., 2021).

As for the coverage of Spanish and Catalan Ferrés et al. (2017b) presents a CNN classifier for CWI in Portuguese, Spanish, Catalan and Galician and Sheang (2019) builts a multilingual system based on a CNN and linguistic feature engineering for

multilingual CWI, which covers Spanish, English and German. So far, these systems tackled CWI, but not LCP with predictions on a complexity scale.

Concerning LS datasets, the aforementioned shared tasks produced valuable resources, mainly for English. There exist LS datasets for Portuguese (Hartmann et al., 2018) and Japanese (Kodaira et al., 2016). Uchida et al. (2018) present a dataset for the the educational domain.

For Iberian Romance Languages, to the best of our knowledge, there are only two datasets for LS in Spanish: EASIER and ALEXSIS. The EASIER dataset was used for CWI and SG/SS tasks (Alarcón et al., 2021); it contains about 5,130 instances (Alarcón et al., 2021) with at least one proposed substitute per complex word. A smaller portion of the dataset which contains 575 instances is more realistic for LS since it contains three proposed substitutes, although without ranking. The EASIER-500 dataset containing 500 instances[2] was used to evaluate SG and SS approaches (Alarcón et al., 2021; Alarcón et al., 2021). ALEXSIS (Ferrés and Saggion, 2022) contains 381 instances composed of a sentence, a target complex word, and 25 candidate substitutions. For every pair <sentence, complex word> a simpler substitute was annotated by a set of 25 annotators. The sentences and complex words of this dataset were extracted from the CWI Shared Task 2018 dataset[3] for Spanish (Yimam et al., 2018a) being its format similar to that of LexMturk (Horn et al., 2014) for English. Again, these datasets cover CWI, but not LCP. In the case of Catalan, there are, to the best of our knowledge, no available datasets at all.

3 Methodology of the Dataset Creation

Both datasets have been created within the data collection efforts for a lexical simplification shared task (Shardlow et al., 2024). The target selection and data collection process of the datasets for Spanish and Catalan was largely parallel, but there were some differences due to the availability of source texts and annotators. The initial goal was to select 600 target words per language in 200 contexts, with 3 targets per context. An additional 10 contexts (and 30 words) were required for pilot annotations. Due to the sparseness of resources we had to relax the goal for Catalan to 160 contexts. For each target a minimum of 10 annotations was required which were collected through on-line forms.

The annotation process collected two pieces of data for each target word: i) a rating on Lexical Complexity on a 5-point Likert scale (from "very easy" to "very hard") and ii) up to 3 lexical substitutes for the target that fit in the given context. Annotators were asked to simply repeat the target word if they could not find a suitable alternative.

In addition to the annotation itself, participants were asked to give some demographic data for the creation of simple statistics: age, years in education, average hours per week used for reading, whether the participant was a native speaker, the number of languages spoken and their native language. Education and weekly reading can be seen as proxies for stylistic and language proficiency and may be used in future studies. Personal data was stored anonymously and separate from annotation data and any data which would allow inferences on the identity of participants was deleted after the dataset compilation. Table 1 gives the resumed demographic information about the participants.

The structure of the datasets is similar to the one of ALEXSIS (described in Section 2), with two important differences: (1) ALEXSIS only contains words for which at least one lexical simplification could be found by the annotators, (2) target words in ALEXSIS do not contain lexical complexity values. Concerning the first point, our datasets also provide examples of non-substitutable words, which is also important for system developments.

The datasets presented here correspond to a combined scenario. This will help the development and assessment of systems that jointly or separately address the lexical simplification pipeline (Paetzold and Specia, 2017). The average ratings on Lexical Complexity are listed normalized to a scale from 0 to 1. Repeatedly proposed substitutions are listed as many times as they were proposed by different annotators. This implies a non-monotonic ranking of their preference. An example of a Catalan and a Spanish annotation is shown in Table 2.

3.1 Catalan Dataset

The Catalan dataset consists of 160 context sentences containing 475 target word tokens (454 distinct types). Sentences were selected from the Educational news section of the TeCla corpus[4] (Armengol-Estapé et al., 2021) of news texts.

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

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

[4] https://huggingface.co/datasets/projecte-aina/tecla

Catalan

Annotators	Av Age	Av Years in Education	Av Reading Hrs per Week	#Participants	#Native Speakers	Languages Spoken (L2)
Personal	58.21 (14.36)	17.93 (4.89)	10.21 (10.54)	14	8	2.21 (1.25)
Prolific	29.30 (8.54)	16.98 (3.24)	7.17 (6.06)	60	13	2.08 (0.81)
All	34.77 (15.02)	17.16 (3.59)	7.75 (7.14)	74	21	2.18 (0.90)

Spanish

Annotators	Av Age	Av Years in Education	Av Reading Hrs per Week	#Participants	#Native Speakers	Languages Spoken (L2)
Personal	34.50 (13.42)	21.78 (3.31)	14.00 (17.35)	10	7	4.1 (2.00)
University	17.98 (1.38)	12.16 (1.50)	2.73 (2.80)	60	60	1.93 (0.55)
All	22.11 (10.85)	13.69 (4.21)	5.67 (14.59)	70	67	2.31 (1.05)

Table 1: Demographic statistics on participants in the data collection. Standard Deviation is given in parentheses. *Personal* stands for personal contacts, *university* for university students and *prolific* for platform annotators.

Spanish Example	*Pero uno no puede dejar que el* **derrotismo** *lo detenga e impida que haga un presupuesto*
LC of target	0.7
Substitutes	desánimo (4), pesimismo (4), abatimiento (3), derrotismo (2), desesperanza (1), desaliento (1), catastrofismo (1), negativismo (1)
Catalan Example	*No poden tocar-se ni abraçar-se, no hi ha joc col·lectiu, s'ha* **sectoritzat** *el pati i la desinfecció per allà on passen és la nova rutina a l'escola.*
LC of target	0.6
Substitutes	dividit (5), segmentat (2), fragmentat (1), seccionar (1), sectorizat (1), divisió en sectors (1), sectoritzat (1), senyalitzat (1), compartimentat (1), dividit en parts (1), en grups (1), classificat (1), separat en zones (1)

Table 2: Examples from our datasets with complexity ratings and LS substitutes. The count of how many times the same word was proposed by different annotators is given in parentheses here, while in the datasets it is represented by the repetition of the words.

3.1.1 Data Preparation

A first pre-selection of candidate *contexts* was done with an automatic process that selected all sentences containing a minimum of 3 content words above a frequency threshold on lemma counts. This threshold was used as an approximate criterion of word difficulty. The frequency was measured with the Catalan Spacy[5] model. The selected contexts were then randomized in order and presented to two annotators (proficient L2 speakers) who had to decide for each word if it was a good simplification candidate because it i) was a complex word and ii) potentially any substitutes could be found for it.

3.1.2 Data Selection: Target Words and Context Sentences

Based on this pre-annotation, we selected target contexts that contained at least one target word unit on which both annotators agreed. For each context 3 targets were selected, giving first preference to units that were agreed on as being complex by the annotators, then those which were marked by only one of them. We did this in order to include words which are guaranteed to be complex and simplifiable. As a last resort, an infrequent word could be selected at random if less than 3 manually marked complex words were available in a sentence. This also allowed the inclusion of some words which might potentially not be simplifiable. This process gave us a total of 480 target words, embedded in 160 context sentences, with each context containing 3 targets. This data was divided into batches (3 batches of 10 targets for a pilot annotation and 9 batches of 50 targets for the rest). Each batch was annotated by a fixed set of annotators.

3.1.3 Annotation

Target words were annotated by proficient Catalan speakers (see Appendix A) We monitored the annotation process in Prolific to detect workers not following the annotation guidelines. For example, annotators who always returned target words as substitutes or provided synonyms in Spanish were contacted and allowed to re-annotate if they wanted.

[5] https://spacy.io/models/ca

Finally, we had to reject 11 annotators. Of the target words 5 had to be removed because they were not correctly presented to the annotators or did not potentially have a meaningful substitute (e.g. calendar dates).

3.2 Spanish Dataset

The Spanish dataset consists of 625 target words in 210 contexts from texts on educational books on finance (see also Appendix B).

3.2.1 Data preparation

Our lexical simplification dataset for Spanish derives from a corpus of over 5K sentences for sentence simplification currently under development. The sentences were simplified following a set of simplification guidelines borrowed from the Simplext project (Saggion et al., 2015). Each sentence was simplified by one of six annotators who were trained to follow the simplification guidelines. The corpus features interesting simplification phenomena such as the transformation of numerical information (*10%* → *diez por ciento*) – a well known simplification operation (Bautista and Saggion, 2014), the splitting of a long sentence into two shorter ones, and lexical substitutions (*derrotismo* → *pesimismo*).

3.2.2 Data Selection: Target Words and Context Sentences

Lexical simplification candidates were heuristically mined from the corpus in order to create our novel LS dataset for Spanish. We search specifically for sentence pairs in which a word was present in the original complex sentence but missing in the simplification. A Natural Language Processing pipeline for Spanish[6] was used to analyze original and simplified sentences and extract words and parts-of-speech tags. We restricted our analysis of lexical simplification to single content words with POS tags noun, verb, adjective or adverb, excluding Multi Word Expressions. The set of unique words in the original and simplification was compared to assess whether a *complex* → *simple* transformation could be identified. A transformation *complex* → *simple* was considered a priory valid substitution if the pair of words were semantically related and not a morphological derivation of one another. A semantic similarity threshold and a lexical similarity threshold were computed in order to implement this validation check using the test data from the ALEXSIS dataset to adjust parameters (see Section 2): all pairs of complex words and substitution words in ALEXSIS were compared using cosine similarity in a Spanish Word Embedding space[7] and the cosine values averaged to obtain a similarity threshold (i.e. similarities greater that the threshold used as an indication of word relatedness). A second value was computed to discard morphological similar (e.g. *obtenido* and *obtener*) pairs: the edit distance between candidates was computed and averaged over all ALEXSIS pairs. These two thresholds were used as a means to discard complex sentences containing a word without an equivalent simplification in the simple sentence, for example, in cases where the sentence underwent a delete operation or a different verb form was used in the simplification. With this, we obtained 1,533 complex sentences containing a potential target word, that is a word which was replaced by a related word in the simplification. This set provided the basis for the human annotation of the dataset.

The selected words in their sentence context were annotated by two annotators (one native Spanish speaker and one with Spanish as L2) on whether the word in question was a good simplification target (being complex and potentially "simplifiable"). In case of doubt dictinonaries were consulted. The process yield 601 valid contexts – contexts were at least one target word on which both annotators had agreed. The data was analyzed again to extract two additional content words from each sentence to provide words which could potentially be "non-simplifiable". From this set, we sampled 210 target contexts by taking into account the average sentence length, selecting sentences whose length deviated at most one standard deviation from the mean length. We ensured that each target word only appeared once in the dataset as a target.

3.2.3 Annotation

The resulting 630 target words were divided into a first batch of 30 contexts and target words to run a trial annotation and a batch of 200 contexts and target words to produce the final dataset. This task was undertaken by students who are native Spanish speakers and by social contacts of the authors. The trial annotation was done by personal contacts, while the main part of the dataset was annotated as part of a curricular activity.

[6] https://spacy.io/models/es

[7] Large Spanish Fasttext Word Embedding model https://zenodo.org/records/3255001

Spanish

LC Level	validity		equivalence		in-context fit		simplicity		
	V	NV	E	NE	F	NF	S	EQ	C
1 [0.00..0.20]	100%	0%	87%	13%	100%	0%	35%	50%	15%
2 (0.20..0.40]	100%	0%	87%	13%	81%	19%	42%	50%	8%
3 (0.40..0.60]	100%	0%	63%	37%	79%	21%	42%	58%	0%
4 (0.60..0.80]	100%	0%	77%	23%	74%	26%	65%	35%	0%
5 (0.80..1.00]	100%	0%	73%	27%	86%	14%	59%	41%	0%
ALL	100%	0%	77%	23%	84%	16%	48%	46%	6%

Catalan

LC Level	validity		equivalence		in-context fit		simplicity		
	V	NV	E	NE	F	NF	S	EQ	C
1 [0.00..0.20]	100%	0%	77%	23%	74%	26%	26%	61%	13%
2 (0.20..0.40]	97%	3%	93%	7%	70%	30%	44%	56%	0%
3 (0.40..0.60]	100%	0%	70%	30%	76%	24%	62%	38%	0%
4 (0.60..0.80]	93%	7%	71%	29%	75%	25%	45%	45%	10%
5 (0.80..1.00]	100%	0%	67%	33%	100%	0%	50%	50%	0%
ALL	97%	3%	78%	22%	58%	42%	44%	50%	6%

Table 3: Qualitative Assessment of the Analysed Substitutes in Spanish and Catalan by complexity level and overall. V: valid word, NV: not valid word, E: equivalent word, NE: non equivalent word, F: fit in context, NF: not fit in context, S: simpler, EQ: equaly simple/complex, C: more complex.

Target / Substitute	Sentence with target / Sentence with substitute / Sentence with correct
Tgt: **mercancías** (LC: 0.3)	✓ El mercado es el lugar donde se transan las **mercancías** y los servicios; es la expresión que define el lugar físico o figurado donde se encuentran vendedores y compradores. (*The market is the place where* **goods** *and services are traded; It is the expression that defines the physical or figurative place where sellers and buyers meet.*)
Sbs: **productos**	✗ El mercado es el lugar donde se transan las **productos** y los servicios; es la expresión que define el lugar físico o figurado donde se encuentran vendedores y compradores.
	✓ El mercado es el lugar donde se transan los **productos** y los servicios; es la expresión que define el lugar físico o figurado donde se encuentran vendedores y compradores.

Table 4: Substitution Amendment Examples in Spanish. In red we highlight the problems when the substitute is used as a direct replacement and in blue how it can be amended. Target = Tgt, Substitute = Subs.

Target / Substitute	Sentence with target / Sentence with substitute / Sentence with correct
Tgt: **manifest** (LC: 0.39)	✓ ... es va crear una comissió de seguiment que s'ha anat reunint d'aleshores ençà i a l'entorn de la qual es van posar de **manifest** algunes mancances ... (... *a follow-up commission was created which has been meeting ever since and around which some shortcomings became* **manifest** ...)
Sbs: **evidència**	✗ ... es va crear una comissió de seguiment que s'ha anat reunint d'aleshores ençà i a l'entorn de la qual es van posar de **evidència** algunes mancances ...
	✓ ... es va crear una comissió de seguiment que s'ha anat reunint d'aleshores ençà i a l'entorn de la qual es van posar en **evidència** algunes mancances ...

Table 5: Substitution Amendment Examples in Catalan. In red we highlight the problems when the substitute is used as a direct replacement and in blue how it can be amended. Target = Tgt, Substitute = Subs.

Each data point was annotated by 10 participants. Five data points had to be removed, 3 of them because no meaningful synonyms could be found (e.g. URLs) and two because there was an error in the annotation forms which prevented participants from giving meaningful answers. So the final dataset consists of 625 target words in 210 contexts.

3.3 Lexical Complexity Analysis

Lexical Complexity is perceived quite subjectively, although some factors, e.g. word frequency in day-to-day communication, are relatively objective factors, despite the fact that corpora may not always represent the day-to-day exposure of language to individuals faithfully. So, one important and interesting question is in how far different annotators agree in their complexity judgements. We expected to find a relatively strong, but not perfect agreement among raters. To assess inter-annotator agreement on complexity rating, it has to be considered that the values from the Likert-scale are ordinal and fall on an interval scale. The best way to treat this is by calculating agreement on the ranking of rated items. For this reason, we use Intraclass Correlation Coefficient (ICC) and Spearman's rho. ICC estimates were calculated using Pingouin (Vallat, 2018) statistical package version 0.5.4 based on a mean-rating, one-way random effects multiple

raters model (ICC1k) (Shrout and Fleiss, 1979). ICC values were calculated for each annotation batch (for which the set of raters was fixed) and then averaged. ICC1k was 0.78 for Spanish and 0.62 for Catalan. There is no generally accepted way to interpret ICC scores, but the value for Spanish can be described as *good* and the one for Catalan as *moderate* (Koo and Li, 2016). In the light of what we said above, this is an expected result.

4 Dataset Quality Analysis

In order to assess the quality of the datasets, we examined several contexts, target words and substitutes to check if those substitutes were simpler, meaning preserving, and fit for the context when used to replace the target word in the given context. While doing our analysis, we considered the top three (most frequent) suggested substitutes per target word hypothesising that they would satisfy the annotation requirements (see Section 3). We discover that, although a majority satisfy the desired properties, there is a considerable number of cases which do not comply with being appropriate in-context substitutes.

Our analysis consists on examining a sample of 270 data-points: 150 data-points for Spanish and 120 data-points for Catalan. The analysis is carried out by two native speakers of Spanish who additionally have C1 and B2 Catalan proficiency. For the assessment of the data-points speaker linguistic proficiency and knowledge of the language was considered while checking on dictionaries [8] to reinforce decisions. The method used for selecting the candidates was as follows: First the lexical complexity (LC) level of target words in the datasets was used to create five categories for analysis as shown for example in Table 3. From each category we selected 10 sentences and their targets (times their three top most human proposed substitutes). All categories in the Spanish dataset have at least ten sentences to select from by random sampling. As for Catalan, all categories, except category number 5, had enough sentences to sample from randomly. For category 5, we just selected the only sentence in that category.

The variables of interest for the analysis are as follows: (1) *validity* - whether or not the substitute is a valid word in the language (e.g. occurs in a dictionary or is a valid morphological derivation of a valid word); (2) *equivalence* - whether the substitute is equivalent to the target word; (3) *in-context fit*: whether the substitute can be used in the syntactic context as the target word; and (4) *simplicity* - whether the substitute is less complex, equally complex, or more complex word. Table 3 presents the overall quantitative results of the analysis as well as the results per lexical complexity category. By looking at the tables we can observe for the Spanish dataset that all proposed substitutes analysed are valid words of the language, however just 77% were considered as equivalent to the target word. Of those considered equivalent an overwhelming majority (84%) were considered to directly fit in the context while about half (48%) were assessed as simpler and 46% considered as equally complex (or simple). A trend can be perceived when looking at the analysis per lexical complexity categories, as complexity of the targets increases, the substitutes' equivalence and context fit decrease. A different trend can be observed with respect to simplicity, as the complexity of the target increases also does simplicity of the substitute. Contrary to the Spanish case, not all Catalan substitutes were valid words in the language (97% are valid words), however an overwhelming majority (78%) are equivalent to the target word but with only 58% being fit for direct replacement. As for simplicity, only 44% are considered simpler than the target. Looking at the complexity levels, the picture is not as clear for Catalan, and we speculate that differences with Spanish may be attributed to the target population who provided the crowd sourced substitutes (i.e. main language Spanish and knowledge of Catalan as second language, see Table 1).

We provide several examples of our analysis that qualitatively illustrate issues related to substitutes which are semantically unrelated, incorrect, or too specific to be used as replacements.

For example, in context "cifras **millionarias** de dinero" (**millionaire** *figures of money*) the substitute "acaudaladas" (wealthy) was considered not equivalent since it is an adjective which is used to qualify people and not to qualify abstract concepts such as "cifras" (figures of money).

Another example would be "...salario o sueldo que se percibe, cuando se tiene un empleo, **honorarios** que se cobran como prestaciones de servicios..." (... *salary that is received, when one has a job,* **honoraries** *that are charged as services...*), in this case the proposed substitute "pagos" (payment) was considered non equivalent to "honorar-

[8] For Spanish the Dictionary of the Spanish Royal Academy and for Catalan the Optimot and Diec2 dictionaries.

ios" (honoraries), the reason being an error in the gender of the word: although "pago" (payment) is a valid word, in Spanish it is the feminine "pagas" (wages) which could have been accepted as replacement. Finally, in the context "hay indicadores financieros que entregan información sobre el pulso **bursátil**, el número y los montos de las transacciones de acciones de sociedades" (*There are financial indicators that provide information about the pulse of the* **stock market***, the number and amounts of transactions in company shares.*) the proposed substitute "bancario" (banking) is a term referring to the banking domain, too specific to be considered equivalent to "bursàtil" (stock market) which is a broader term (which includes the banking domain).

As for Catalan, we illustrate three examples of incorrect substitutes due to problems of figurative language use or domain connected or semantically related – but not equivalent – words.

In the following context: "El Síndic també posa de manifest que una **sobreoferta** té efectes negatius sobre la segregació escolar..." (*The Ombudsman also points out that an* **oversupply** *has negative effects on school segregation...*) a substitute "sobresaturació" (oversaturation) would not provide a valid replacement for "sobreoferta" (oversupply) since this candidate substitution refers (figuratively) to people undergoing stress and it does not refer to an increase in (educational) course offer.

The context " l'Associació Celíacs de Catalunya ha denunciat la "situació d'indefensió" en la que es troben els 30.000 alumnes celíacs o sensibles al **gluten** que mengen als menjadors catalans" (*the Associació Celíacs de Catalunya has denounced the "helpless situation" in which the 30,000 celiac or* **gluten***-sensitive students who eat in Catalan canteens find themselves*) the candidate "ségol" (rye) can not be considered a valid replacement since "gluten" (gluten) is a proteine found in cereals like rye, but the terms are not equivalent.

Finally, in the context "Mitjançant la psicologia, el '**mindfulness**' i el ioga, els alumnes aprenen a resoldre conflictes i, alhora, valors com l'autoestima o el respecte." (*Through psychology,* **mindfulness** *and yoga, students learn to resolve conflicts and, at the same time, values such as self-esteem or respect.*) the word "meditació" (meditation) cannot be taken as an equivalent of "mindfulness" since these are two different but related concepts in psychology.

In Tables 4 and 5, we present examples of substitutes which are equivalent to the targets but nonetheless their use as direct replacement is not without consequences for the correctness of the resulting sentence. Indeed, a lexical simplification system should take into account context modification at the local and global level to guarantee grammaticality, coherence, and cohesion. We can observe that gender and governed prepositions have to be adapted to the substitution.

5 A Note on Ethical Considerations for Lexical Simplification Datasets

Although a very detailed analysis of the dataset could not be carried due to limited resources, we believe it is important to highlight aspects related to ethics which have not been addressed thus far in the field of lexical simplification. Since lexical simplification aims at substituting lexical items that may be unfamiliar and difficult to understand, the automated process may produce output which could raise concerns from the ethical viewpoint since the replacements may lead to unfair, unethical or false description of people or events. The following is a clear example of discriminatory, offensive language: Let's suppose we are given the sentence "She has a disabled brother." and the target word "disabled". English dictionaries list "retarded" as an offensive synonym of disabled, therefore in case a system does not take into account that metadata information, an offensive sentence could be produced as in "She has a retarded brother."[9] .The same goes without saying for the use of word-embedding models or LLMs which are trained on data which is not properly annotated for ethics.

The subset of data points we have analyzed already contains some traces of the problems described above, somehow concerning because it directly comes from human informants. Although only a few items entail ethical concerns a process of carefully revision and ethical disclosure as the one we have put forward here is necessary, specially in the case of a crowd annotated dataset, to understand the risk the provided data may entail. From a pure automated evaluation viewpoint, in the previous illustration if the offensive term is used to replace a non-offensive one, being considered

[9] In Spain pejorative terms were recently removed from the Spanish Constitution https://www.lamoncloa.gob.es/serviciosdeprensa/notasprensa/presidencia-justicia-relaciones-cortes/Paginas/2024/180124-congreso-aprobada-reforma-constitucion.aspx.

valid in the gold standard, the system producing such output would be rewarded (!).

Although in the Spanish data no serious problems were detected, two relevant cases are present in the Catalan data: The first case is a replacement suggested by a crowd workers which could be considered an euphemism and which, in this particular case, should be avoided: For the sentence "En una segona part, explica Campàs, els participants aprenen estratègies per abordar la violència masclista i que comportaments “poc visibles”, a la llarga es poden traduir en “assetjaments, violacions i **feminicidis**”." (*In a second part, explains Campàs, the participants learn strategies to address male violence and that “not very visible” behaviors, in the long run, can translate into “harassment, rape and* **femicide***”*) and the target word **feminicidi** (femicide), the substiture "assassinat" (murder) was proposed which does not carry the very meaning of the target word also diminishing its intended meaning.

The second case illustrate the proposal of two offensive terms: For the context "Alguns dels alumnes de 5è, amb qui també s’ha treballat una de les cançons del conte encara que no participen a la cantata, han explicat com mai abans havien sentit abans paraules com transsexual o **lesbiana**." (*Some of the 5th grade students, with whom one of the songs in the story was also worked on even though they do not participate in the cantata, explained how they had never heard words like transsexual or* **lesbian** *before.*) and the target word **lesbiana** (lesbian), one crowd annotator suggested the word "gallimarsot"[10] while another annotator proposed the term "marieta"[11] which can be considered pejorative terms to refer to a lesbian. But note that this example is also interesting in that the term "lesbiana" in this context is referring to the word itself, and therefore should not be replaced.

6 Conclusions and Future Work

As we have argued throughout the paper, there is a clear need to have more resources like the one presented here for Catalan and Spanish. Such datasets are a prerequisite for the development and evaluation of LS and LCP systems. We have described two novel datasets which allow the development and evaluation of Lexical Simplification Systems for Catalan and Spanish. We expect that these datasets are a valuable addition to the currently sparse data in this field. We have quantitatively and qualitatively assessed the dataset confirming the suitability of the dataset for lexical simplification research. Moreover we have also discussed ethical issues discovered through this analysis which should inform further dataset releases. The dataset has already been used in a shared task in lexical simplification (Shardlow et al., 2024) and our future work will consider a thorough analysis of system contributions, and in particular how to leverage system outputs to improve data creation and assessment. Given that target users of text simplification systems include vulnerable populations, we would like to launch *a call to arms* for better ethical control during data creation and annotation and evaluation of automatic systems so as to flag at early stages any sensitive issues which may affect the intended user of these systems.

Lay summary

For many people accessing information in written texts is too difficult, because the text is written in a style that is too hard for them. This can happen to elderly people, language learners and people with cognitive impairments, among others. Automatic Text Simplification can help to adapt texts for them. Lexical Simplification is one aspect of Text Simplification. It replaces difficult words with easier ones. For the creation of Automatic Text Simplification data sets are necessary which contain examples of good substitutions of words with simpler alternatives. We present two datasets of this type for Spanish and Catalan. For Spanish, there are only very few existing datasets so far and for Catalan there are none. Our contribution fills this gap and will make the development of Spanish and Catalan Text Simplification systems possible.

Acknowledgments

This document is part of a project that has received funding from the European Union’s Horizon Europe research and innovation program under the Grant Agreement No. 101132431 (iDEM Project). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the EU. Neither the EU nor the granting authority can be held responsible for them. We also thank the Departament de Recerca i Universitats de la Generalitat de Catalunya (ajuts SGR-Cat 2021) for its support.

[10] Zoomorphism to refer to a female who acts as a male. https://dlc.iec.cat/

[11] Despective for homosexual. Diccionario LGBT+ Catalán https://lgbt.fandom.com/es/wiki/Diccionario_LGBT

References

Rodrigo Alarcón, Lourdes Moreno, and Paloma Martínez. 2021. 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.

Rodrigo Alarcón, Lourdes Moreno, and Paloma Martínez. 2021. Lexical Simplification System to Improve Web Accessibility. *IEEE Access*, 9:58755–58767.

Jordi Armengol-Estapé, Casimiro Pio Carrino, Carlos Rodriguez-Penagos, Ona de Gibert Bonet, Carme Armentano-Oller, Aitor Gonzalez-Agirre, Maite Melero, and Marta Villegas. 2021. Are multilingual models the best choice for moderately underresourced languages? A comprehensive assessment for Catalan. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4933–4946, Online. Association for Computational Linguistics.

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

Susana Bautista and Horacio Saggion. 2014. Making Numerical Information more Accessible: Implementation of a Numerical Expressions Simplification Component for Spanish. *ITL- International Journal of Applied Linguistics*, (Special Issue on Readability and Text Simplification):299–323.

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

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

Stefan Bott, Luz Rello, Biljana Drndarević, and Horacio Saggion. 2012b. Can spanish be simpler? lexsis: Lexical simplification for spanish. In *Proceedings of COLING 2012*, pages 357–374.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 workshop on integrating artificial intelligence and assistive technology*, pages 7–10. Association for the Advancement of Artificial Intelligence.

Comité Económico y Social Europeo. 2011. Educación financiera y consumo responsable de productos financieros. *Recuperado el*, 27.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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. Association for Computational Linguistics.

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

Daniel Ferrés, Horacio Saggion, and Xavier Gómez Guinovart. 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. Association for Computational Linguistics.

Daniel Ferrés, Horacio Saggion, and Xavier Gómez Guinovart. 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.

Daniel Ferrés and Horacio Saggion. 2022. ALEXSIS: A Dataset for Lexical Simplification in Spanish. In *Language Resources and Evaluation Conference (LREC-2022)*.

Nat Gillin. 2016. Sensible at semeval-2016 task 11: Neural nonsense mangled in ensemble mess. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 963–968.

Goran Glavaš and Sanja Štajner. 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. Association for Computational Linguistics.

Nathan S Hartmann, Gustavo H Paetzold, and Sandra M Aluísio. 2018. Simplex-pb: A lexical simplification database and benchmark for portuguese. In *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*, pages 272–283. Springer.

Colby Horn, Cathryn Manduca, and David Kauchak. 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. Association for Computational Linguistics.

Ludivine Javourey-Drevet, Stéphane Dufau, Thomas François, Núria Gala, Jacques Ginestié, and Johannes C Ziegler. 2022. Simplification of literary and scientific texts to improve reading fluency and comprehension in beginning readers of french. *Applied Psycholinguistics*, 43(2):485–512.

Marius Rohde Johannessen, Lasse Berntzen, and Ansgar Ødegård. 2017. A review of the norwegian plain language policy. In *Electronic Government: 16th IFIP WG 8.5 International Conference, EGOV 2017, St. Petersburg, Russia, September 4-7, 2017, Proceedings 16*, pages 187–198. Springer.

Tomonori Kodaira, Tomoyuki Kajiwara, and Mamoru Komachi. 2016. Controlled and balanced dataset for japanese lexical simplification. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 1–7.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

Marta Licardo, Nina Volčanjk, and Dragica Haramija. 2021. Differences in communication skills among elementary students with mild intellectual disabilities after using easy-to-read texts. *The new educational review*, 64:236–246.

Misako Nomura, Gyda Skat Nielsen, and Bror Tronbacke. 2010. *Guidelines for easy-to-read materials*. International Federation of Library Associations and Institutions (IFLA).

Jenny A Ortiz-Zambrano and Arturo Montejo-Ráez. 2020. Overview of ALexS 2020: First workshop on lexical analysis at SEPLN. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*, volume 2664, pages 1–6.

Gustavo Paetzold and Lucia Specia. 2016. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

Gustavo Paetzold and Lucia Specia. 2017. 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. Association for Computational Linguistics.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2020. LSBert: A Simple Framework for Lexical Simplification. *arXiv preprint arXiv:2006.14939*.

Maury Quijada and Julie Medero. 2016. Hmc at semeval-2016 task 11: Identifying complex words using depth-limited decision trees. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1034–1037.

Evelina Rennes. 2022. *Automatic Adaptation of Swedish Text for Increased Inclusion*. Ph.D. thesis, Linköping University Electronic Press.

Irina Rets and Jekaterina Rogaten. 2021. To simplify or not? facilitating english l2 users' comprehension and processing of open educational resources in english using text simplification. *Journal of Computer Assisted Learning*, 37(3):705–717.

Francesco Ronzano, Luis Espinosa Anke, Horacio Saggion, et al. 2016. Taln at semeval-2016 task 11: Modelling complex words by contextual, lexical and semantic features. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1011–1016.

Horacio Saggion. 2017. *Automatic Text Simplification*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

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

Horacio Saggion, Sanja Štajner, Daniel Ferrés, Kim Cheng Sheang, Matthew Shardlow, Kai North, and Marcos Zampieri. 2022. Findings of the TSAR-2022 shared task on multilingual lexical simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 271–283, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Matthew Shardlow. 2013. A comparison of techniques to automatically identify complex words. In *51st annual meeting of the association for computational linguistics proceedings of the student research workshop*, pages 103–109.

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

Matthew Shardlow. 2014b. Out in the open: Finding and categorising errors in the lexical simplification pipeline. In *LREC*, pages 1583–1590.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Sanja Štajner, Marcos Zampieri, and Horacio Saggion. 2024. The BEA 2024 shared task on the multilingual lexical simplification pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. Complex: A new corpus for lexical complexity prediction from likert scale data. *arXiv preprint arXiv:2003.07008*.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. Semeval-2021 task 1: Lexical complexity prediction. *arXiv preprint arXiv:2106.00473*.

Kim Cheng Sheang. 2019. Multilingual complex word identification: Convolutional neural networks with morphological and linguistic features. In *Proceedings of the Student Research Workshop (RANLPStud 2019); 2019 Sep 2-4; Varna, Bulgaria.[Varna]: ACL; 2019. p. 83-9.* ACL (Association for Computational Linguistics).

Kim Cheng Sheang, Daniel Ferrés, and Horacio Saggion. 2022. Controllable lexical simplification for English. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 199–206, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Kim Cheng Sheang and Horacio Saggion. 2023. Multilingual controllable transformer-based lexical simplification. *Proces. del Leng. Natural*, 71:109–123.

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

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

Sanja Stajner, Daniel Ibanez, and Horacio Saggion. 2023. Less: A computationally-light lexical simplifier for spanish. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, RANLP 2023, Varna, Bulgaria, 4-6 September 2023*, pages 1132–1142. INCOMA Ltd., Shoumen, Bulgaria.

Satoru Uchida, Shohei Takada, and Yuki Arase. 2018. Cefr-based lexical simplification dataset. In *Proceedings of International Conference on Language Resources and Evaluation*, volume 11, pages 3254–3258. European Language Resources Association.

Raphael Vallat. 2018. Pingouin: statistics in python. *J. Open Source Softw.*, 3(31):1026.

Bruno Bastos Vieira de Melo, Mónica Silveira-Maia, and Sandra Barbosa Ribeiro. 2023. Full financial education programmes for people with disabilities: a scoping review. *Revista Brasileira de Educação Especial*, 29:e0222.

Tuqa Bani Yaseen, Qusai Ismail, Sarah Al-Omari, Eslam Al-Sobh, and Malak Abdullah. 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.

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

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo H. Paetzold, Lucia Specia, Sanja Stajner, Anaïs Tack, and Marcos Zampieri. 2018a. A Report on the Complex Word Identification Shared Task 2018. *CoRR*, abs/1804.09132.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo H Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018b. A report on the complex word identification shared task 2018. *arXiv preprint arXiv:1804.09132*.

Marcos Zampieri, Shervin Malmasi, Gustavo Paetzold, and Lucia Specia. 2017. Complex word identification: Challenges in data annotation and system performance. *arXiv preprint arXiv:1710.04989*.

Sanja Štajner, Horacio Saggion, and Simone Paolo Ponzetto. 2019. Improving Lexical Coverage of Text Simplification Systems for Spanish. *Expert Systems with Applications*, 118:80–91.

Appendix A: Selection criteria for annotators

For Catalan, annotators were in part recruited from persons of the social environment of the authors and in part from workers recruited over the Prolific[12] crowdsourcing platform.[13] All trial data was annotated by social contacts, as well as a part of the main annotations. In the case of Catalan it is difficult to select a pool of participants that consists only of native speakers because Catalonia is a largely bilingual territory. However, since Catalan has been used as the main vehicular language in the school system for several decades, most people who had their education in Catalonia have a high level of Catalan proficiency. Also a large part of the population grew up bilingually.

For Spanish, the trial annotation was done by personal contacts, while the main part of the dataset was annotated as part of a curricular activity within a course on written communication. This course was designed to foster the development of skills necessary for writing scientific and academic texts that are comprehensible to a broad audience. It required the texts to adhere to standards of clarity, precision, coherence, and readability, aligning with the principles of effective scientific communication. The primary intent behind this task was to enhance the student's ability to identify and modify the use

[12] https://www.prolific.com/

[13] Annotators received a fair pay.

of complex terminology, opting for more accessible alternatives without compromising the accuracy or depth of the content. This approach facilitates widespread dissemination and understanding.

The annotators recruited from personal contacts were mostly speakers of European Spanish , while the rest were speakers of the Costa Rican variety of Spanish.

Since the availability of annotators was limited, the main criterion for the recruitment of annotators from the personal contacts of the authors was their availability, both for Spanish and for Catalan. We made sure that all of them were proficient speakers of the language, either native or L2 speakers which use the language on a daily basis. Even without having any stricter selection criteria, in practice their annotations were much more reliable than annotations from crowdsourcing workers. For Catalan we had to discard 11 crowdsourcing annotators.

Appendix B: Selection criteria for texts

Both of datasets have been created within the context of the MLSP24 (Multilingual Lexical Simplification Pipeline) shared task (Shardlow et al., 2024), in which comparable datasets for 10 languages were created. In the guidelines for the data selection it was strongly suggested to use texts from the educational domain.

For Catalan, we could not find a sufficiently large corpus of educational text. So, entences were selected from the Educational news section of the TeCla corpus (Armengol-Estapé et al., 2021) of news texts.

For Spanish, we selected educational texts on finance due to their social relevance and the pressing need to make this knowledge accessible to vulnerable populations. Financial literacy, recognized as an essential tool for economic empowerment and inclusion, especially among individuals with disabilities, remains underexplored in text simplification (Vieira de Melo et al., 2023). Learning about personal finance is critical in fostering autonomy and improving decision-making. The specialized nature of these texts, characterized by domain-specific terminology and conceptual density, requires careful consideration in simplification approaches to maintain accessibility and accuracy (Comité Económico y Social Europeo, 2011). Our research addresses these challenges by focusing on this area, aligning with broader efforts to promote financial competence and social inclusion for underserved communities. The Spanish texts originate from publications in South America.