IrekiaLF_es: a new open benchmark and baseline systems for Spanish Automatic Text Simplification

Itziar Gonzalez-Dios
HiTZ Center - Ixa, University of the Basque Country (UPV/EHU)
itziar.gonzalezd@ehu.eus

Iker Gutiérrez-Fandiño
University of Deusto
ikergutierrez@opendeusto.es

Oscar M. Cumbicus-Pineda
Univ. Nac. de Loja and Ixa (UPV/EHU)
oscar.cumbicus@unl.edu.ec

Aitor Soroa
HiTZ Center - Ixa, University of the Basque Country (UPV/EHU)
a.soroa@ehu.eus

Abstract

Automatic Text simplification (ATS) seeks to reduce the complexity of a text for a general or a target audience. In the last years, deep learning methods have become the most used systems in ATS research, but these systems need large and good-quality datasets to be evaluated. Moreover, these data are available on a large scale only for English and in some cases with restrictive licenses. In this paper, we present IrekiaLF_es, an open-license benchmark for Spanish text simplification. It consists of a document-level corpus and a sentence-level test set that has been manually aligned. We also conduct a neurolinguistically-based evaluation of the corpus in order to reveal its suitability for text simplification. This evaluation follows the Lexicon-Unification-Linearity (LeULi) model of neurolinguistic complexity assessment. Finally, we present a set of experiments and baselines of ATS systems in a zero-shot scenario.

1 Introduction

According to the UN, illiteracy affects 16 per cent of the world population (759 million adults)[1] and the number of children experiencing reading difficulties has increased from 460 million to 584 million after the COVID-19 pandemic.[2] Moreover, in the OECD countries, between 4.9 % and 27.7 % of adults only has proficiency at the lowest levels in literacy (OECD, 2013) and regarding the young people 10 % of new graduates have low literacy skills (OECD, 2015).

Due to these facts, plain language[3] and easy-to-read initiatives[4] give some guidelines to make texts more accessible. Basically, their recommendations can be summarised in writing for the audience, organising the information, using short and positive sentences, using active instead of passive voice, choosing words carefully, being concise, employing an appropriate design for smooth reading and, in the case of online communication, following the web standards.

In this line, text simplification seeks to reduce the complexity of texts (at the lexical, syntactic and discourse levels) for a general public or a target audience. As adapting the texts manually is a hard-working task, researchers in Natural Language Processing (NLP) have tried to automatise it since the mid 90s. The pioneers were Chandrasekar et al. (1996) and their main motivation was related to the problems that long and complex sentences caused in advanced NLP applications. Nowadays, however, most of the research on Automatic Text Simplification (ATS) focuses on human target audience (Štajner, 2021).

As other NLP tasks, ATS has evolved from rule-based systems, statistic systems, hybrid systems mixing hand-crafted rules and machine learning, to the present deep learning paradigm. The interested reader is referred to the following state-of-the-art reports for more information about the evolution of ATS (Gonzalez-Dios et al., 2013; Shardlow, 2014; Siddharthan, 2014; Saggion, 2017; Alva-Manchego et al., 2020b).

Current deep learning techniques, however, need extensive data and these data are mainly available for English. Moreover, some corpora do not have open licenses. There are NLP techniques to alleviate this problem such as transfer learning and cross-lingual learning, but, even in these cases, high-quality evaluation benchmarks are needed.

In this paper, we present IrekiaLF_es, an open corpus and benchmark for Spanish ATS systems. IrekiaLF_es compiles texts published by the Basque Government in both original and easy-to-read format. The corpus is divided in a document-

[1] https://www.un.org/en/chronicle/article/education-all-rising-challenge
[2] https://news.un.org/en/story/2021/03/1088392
[3] https://www.plainlanguage.gov/
[4] https://www.inclusion-europe.eu/easy-to-read/

Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022), pages 86 - 97

level version containing 288 documents, and a sentence-level version, where 35 of them have been manually aligned. The corpus is available with an open license.[5] Furthermore, in order to reveal its quality and suitability for ATS, we have evaluated neurolinguistic complexity of the corpus following the Lexicon-Unification-Linearity (LeULi) model (Gutiérrez-Fandiño, 2022). This model of neurolinguistic complexity assessment is entirely inspired by Hagoort (2005, 2013, 2014, 2017, 2019, 2020)'s Memory, Unification, Control (MUC) model of language neurobiology. Finally, we have evaluated three different systems that will serve as baselines for future research in this corpus.

This paper is structured as follows: right after this introduction (Section 2), we present the related work on simplified corpora; in Section 3 we describe the methodology to build the IrekiaLF_es corpus; in Section 4 we summarise the LeULi model and provide the rationale for its use; in Section 5 we carry out a LeULi-based complexity evaluation of the corpus; in Section 6 we show the experiments with the three baseline systems; and we conclude with the take-home messages and outline the future work in Section 7.

2 Related work

Corpora or datasets for ATS are built with two main objectives: on the one hand, to study the process and operations carried out when simplifying texts e.g. by annotating the operations (Caseli et al., 2009; Bott and Saggion, 2014; Brunato et al., 2015; Gonzalez-Dios et al., 2018), and, on the other hand, to use them as resources to build and evaluate machine learning systems (see next paragraph for references). When creating and compiling corpora of simplified texts, the strategies (intuitive or structural) and the target audiences can be different (Gonzalez-Dios et al., 2018). Hence, there is no unique answer or simplified correct sentence for a given complex or original sentence. A recent overview on the creation of ATS corpora can be found in Brunato et al. (2022).

The main research in text simplification, both from an educational perspective and from an NLP perspective, has focused on English and, therefore, the majority of corpora as well as the largest ones are in such language (Petersen and Ostendorf, 2007; Pellow and Eskenazi, 2014; Vajjala and Lučić, 2018). The most used datasets for NLP are i) those derived from Wikipedia and Simple Wikipedia (and therefore with open licences): WikiSmall, originally created by Zhu et al. (2010) and adapted by Zhang and Lapata (2017), WikiLarge, compiled by Zhang and Lapata (2017) and usually used for training, TurkCorpus (Xu et al., 2016) and Asset (Alva-Manchego et al., 2020a) used for evaluation and ii) Newsela (Xu et al., 2015), which has proprietary licence but can be obtained for research. In order to study the document level simplification, D-Wikipedia dataset has been proposed (Sun et al., 2021).

However, there are also corpora for other languages such as Brazilian Portuguese (Caseli et al., 2009; Hartmann et al., 2018, 2020), Danish (Klerke and Søgaard, 2012), German (Klaper et al., 2013; Battisti et al., 2020; Säuberli et al., 2020), Italian (Brunato et al., 2015; Tonelli et al., 2016; Brunato et al., 2016), Basque (Gonzalez-Dios et al., 2018) and French (Gala et al., 2020). Recently, a multilingual corpus of news has been compiled that includes Finnish, French, Italian, Swedish, English and German (Hauser et al., 2022).

Regarding Spanish, the first ATS corpus was developed in Saggion et al. (2011)'s project, which aimed to build an ATS system guided by the so-called easy-to-read principles. The corpus, named Simplext, was created manually by trained experts and the target audience were students with Down Syndrome. An analysis of the operations needed to simplify the original text revealed that the most frequent operations in Simplext were change (transformation), delete, insert and split (Bott and Saggion, 2011, 2014). This corpus is available upon request from the authors. The Newsela corpus (Xu et al., 2015), which is also available upon request, contains a portion in Spanish, but there is no information about the particulars of the Spanish subset. There are three resources for Spanish focusing on lexical simplification: the LexSiS corpus (Bott et al., 2012) (obtained upon request), and the EASIER corpus (Alarcón García, 2022) (available at GitHub, but without explicit license), and ALEXSIS (Ferrés and Saggion, 2022), which will be available after what the authors call *embargo period* (but so far there is no explicit license). Finally, a bilingual (EN/ES) dataset about Covid-19 texts (Simple TICO-19) has been released (Shardlow and Alva-Manchego, 2022) and a corpus for Spanish medical text simplification, the CLARA-MeD comparable corpus, is made up of 24 298 pairs of pro-

[5] https://github.com/itziargd/IrekiaLF

	orig	e2r
Word number	185,070	135,659
Token number	231,332	177,402
Sentence number	5,389	2,408

Table 1: Basic statistics of the document level corpus.

fessional and simplified texts (Campillos-Llanos et al., 2022).

3 Building IrekiaLF_es

Irekia is the open-government communication channel of the Basque Government. This web site contains, among others, news about the Government, written in a non-administrative language, both in Spanish and Basque. Some of the news are adapted to the easy-to-read format, thereby making the site very valuable as a source to compile complex-simple parallel texts, which, moreover, can be bilingual. The portal has CC-BY license, so that its content can be used to derive research datasets.[6] Based on this resource, the aim of this work is to create a good-quality corpus of Spanish, IrekiaLF_es, with original and adapted/simplified texts. We release the corpus under an open license, thus expanding the options for ATS researchers to train and test systems for Spanish.

IrekiaLF_es is built by crawling all the Spanish news that have an easy-to-read counterpart until 17/11/2021 (unfortunately the last adapted version was published in 28/12/2021). The first document's date is 1/04/2017. After removing the duplicates, we have compiled a document-level corpus comprising 288 parallel documents. The dataset is publicly available[7] under CC BY-SA 4.0 license.

Table 1 shows the number of words, tokens and sentences of the complex and simple parts of the corpus.[8] As it can be seen, the *orig* texts are much longer than the *e2r*. This is in line with what is found for example in English corpora (Amancio and Specia, 2014), where simplified texts have also fewer words.[9]

As in the original web site, some complex and unfamiliar words in the documents are linked to their definitions (see Figure 1). We keep these definitions at the document level dataset so that they can serve for both complex word identification and generation of explanations (elaboration). There are 1624 definitions in the corpus that explain complex legal denominations, named entities, and complex words.

In addition to the document-level corpus, we have created a subcorpus, a part of the document-level corpus, that is manually aligned at a sentence level,[10] and which comprises 705 aligned sentences from 35 documents.[11] We have followed this methodology to align the sentences:

1. Preliminary alignment: As a preliminary step, two persons (a computational linguist expert in ATS and a linguistics student) aligned the sentences in five documents and discussed the doubts and unclear cases. As a result, the following guidelines were defined:
   - Align sentences according to information preservation.
   - Do not manipulate easy-to-read texts to improve the simplifications.
   - Regarding sentence boundaries: periods (.) and ellipses (...) indicate the end of a declarative sentence; exclamation marks (!) indicate the end of an exclamative or imperative sentence; question marks (?) indicate the end of an interrogative sentence. Colons (:) are also sentence boundaries, but only when they are used to introduce new paragraphs, not when they link two clauses in a subordinate relationship.
2. Agreement alignment: once the guidelines were fixed, the annotators aligned ten additional documents, so that inter-annotator agreement could be calculated. The resulting percentage of agreement (or observed agreement) was 80.5 % and Cohen's kappa was 0.7. We considered this as a substantial agreement and, therefore, as a good basis to align the rest of the corpus.

[6] We are not aware of other governmental initiatives that could serve as data source under the same conditions. Research should be done at local and regional levels to consider the possibility of adding other data sources to augment the dataset.

[7] https://github.com/itziargd/IrekiaLF

[8] In this paper, we will call *orig* to the original, complex text and *e2r* to the simplified, easy-to-read counterpart.

[9] We do not have the data of the other Spanish corpora to make this comparison.

[10] We have also assessed the possibility of using ATS alignment tools for an automatic alignment, but none of them was good enough to maintain the quality of the alignments.

[11] The definitions of complex words are not considered when aligning the sentences.

Figure 1: Example of words with hyperlinks to definitions.

3. General alignment: one of the annotators has aligned the rest of the documents in the sentence-aligned subcorpus (20 documents).

In Table 2 we present an example of two sentences aligned to their *e2r* counterparts. The translations of the examples are provided in Table 7 (Appendix A).

In Table 3 we present the number and percentage of the alignment scales, that is, how many sentences have been created/removed out of the original one. The percentage of merge operations is remarkable, close to the Italian *Teacher* subcorpus (Brunato et al., 2015), but high in comparison to the Basque CBST (Gonzalez-Dios et al., 2018) and the Italian *Terence* subcorpus (Brunato et al., 2015).[12] Most alignments are at scale 1:1, that is, when no splitting is performed, followed by 1:0, where the sentence has no equivalent *e2r* version. In the vast majority of splitting cases, the original sentence has been split into two (1:2), or into three (1:3) *e2r* sentences. Splitting into more than three sentences is residual.

4 LeULi model of complexity assessment

Evaluation of simplified corpora and ATS systems is more often than not largely based on formal, purely linguistic complexity metrics. However, determining the so-called readability can only be achieved in terms of neurolinguistic complexity: text comprehension takes place in the brain, and as such it requires a cognitive assessment.

On the basis of Hagoort (2005, 2013, 2014, 2017, 2019, 2020)'s Memory, Unification, Control (MUC) model of language neurobiology, Gutiérrez-Fandiño (2022)[13] proposes a model of three categories of neurolinguistic complexity assessment: Lexicon-Unification-Linearity (LeULi). Complexity assessment of IrekiaLF_es has therefore been conducted according to the LeULi model, which is synthesised in Table 4.

Regarding Lexical complexity, it has been shown that the less frequent a word is, the more effortful it is retrieve from long-term memory (LTM), resulting in higher levels of neural activation (Fiebach et al., 2002; Nakic et al., 2006). Infrequent words are not just more difficult to access, they are also more likely to be unknown to the reader, in which case they do not even exist in the mental lexicon and are hence impossible to retrieve from LTM.

The Unification of information from different language modules of the brain is a costly operation in language processing. Accordingly, to lighten complexity of the Unification category, linguistic phenomena involving several language modules should be strongly avoided in ATS: for instance, coreference (principally measured by pronoun incidence) demands the integration of information from the syntactic and semantic language modules.[14] Similarly, the large presence of elements that help avoid the presence of coreference is also an indicator of the lack of Unification complexity: the ratio of proper nouns for all nouns and content word overlap, for example, are metrics showing how often referents are repeated, instead of being pronominalised, coreferenced.

Lastly, linearity affects sentence comprehension as it determines both the temporal separation of chunks provisionally stored in working memory (WM) and the number of chunks stored in such temporary buffer during the processing of a sentence. It is primarily measured by sentence length, but also by non-selected constituents (adjunction and coordination), since selected constituents such as complements are effortlessly retrieved from LTM as part of the syntactic template of any lexical item.

5 Complexity assessment of IrekiaLF_es

In this section we present the results of the complexity evaluation of the whole IrekiaLF_es (document-level corpus). As explained in Section 4, we wanted

[12] We compare the alignments to these works because they are the only available to our knowledge. We show the statistics of the other datasets in the Appendix, in Table 8.

[13] This project will be soon publicly available at `https://dkh.deusto.es/en/community/learning/tfg`. Yet, anyone interested can obtain it in advance upon request to the author via email (ikergutierrez@opendeusto.es).

[14] The present analysis focuses on the syntactic-semantic Unification, since phenomena involving the remaining phonological module seem to have no significant presence in text processing.

orig	e2r
La Consejera de Empleo y Políticas Sociales, Beatriz Artolazabal, y el Consejero de Hacienda y Economía, Pedro Azpiazu, se han reunido con los responsables de EHLABE-Euskal Herriko Lan Babestuaren Elkartea-Asociación vasca de entidades no lucrativas que fomentan la inclusión sociolaboral de las personas con discapacidad, en la sede de Lantegi Batuak, en Loiu, Bizkaia.	Beatriz Artolazabal es la Consejera de Empleo y Políticas Sociales del Gobierno Vasco. Pedro Azpiazu es el Consejero de Hacienda y Economía del Gobierno Vasco. Euskal Herriko Lan Babestuaren Elkartea (EHLABE) es una asociación que impulsa la inclusión en la sociedad y en el trabajo de las personas con discapacidad. Beatriz Artolazabal y Pedro Azpiazu se han reunido con los responsables de EHLABE en la sede de Lantegi Batuak, en Bizkaia.
En el encuentro se ha analizado el trabajo de estas empresas y se han propuesto nuevas fórmulas de colaboración.	En la reunión han estudiado el trabajo de estas empresas y se han propuesto nuevas maneras de colaborar.

Table 2: Examples of aligned sentences (English translations in Table 7 in the Appendix).

Alignment scale	Sentences	Percentage
Merge (2:1)	53	7.6%
1:0	154	22.0%
1:1	310	44.0%
1:2	123	17.5%
1:3	50	7.1%
1:4	14	2.0%
1:5	1	0.1%

Table 3: Statistics of the alignment scales (the sentences created/removed out of the original sentences).

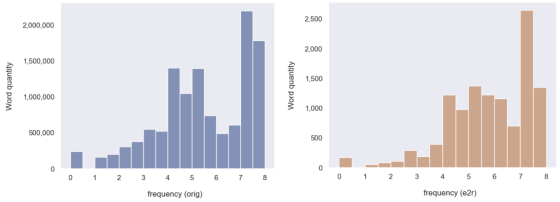

Figure 2: Histograms of word frequencies, grouped in eight bins according to the frequency logarithm.

to base our complexity evaluation on recent evidence on sentence and text processing complexity, and thus we have decided to follow the metrics provided by the LeULi model (Gutiérrez-Fandiño, 2022). For the automatic measurement of complexity metrics, we have employed MultiAzterTest (Bengoetxea and Gonzalez-Dios, 2021), an open-source multilingual text analysis tool which examines more than 130 features at various linguistic levels. To calculate word frequencies we have used Python's *wordfreq* package (Speer et al., 2018). Specifically, we have grouped words in eight bins according to the logarithm of their frequencies.[15] Next, we present histograms and violin plots (Hintze and Nelson, 1998) comparing the scores of *orig* and *e2r* texts according to the neurolinguistic complexity metrics of the LeULi model.

Regarding the Lexicon category, there are substantially more infrequent words (0-4 levels) in *orig* than in *e2r* texts (Figure 2), where words around frequency level 6 are more regularly distributed, as a result of the conscious, purposeful use of frequent words.

When it comes to the Unification category, in Figure 3 it can be observed that there is a positive but too slight difference in the incidence of pronouns: *e2r* texts should have a markedly lower score than *orig* in this metric. The higher the pronoun incidence, the higher the Unification complexity.

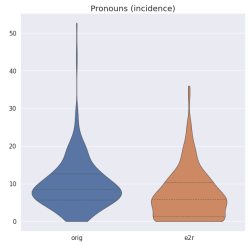

Figure 3: Incidence of pronouns.

As previously explained, the ratio of proper nouns for all nouns and content word overlap are metrics that show the extent to which coreference

[15] Which corresponds to the `zipf_frequency` of the *wordfreq* package.

Category	Constrainer	Constrainee	Metrics
Lexicon	LTM	Infrequent words: hard to access or not stored	**Word frequency**
Unification	WM	The integration of information from different modules	**Multi-module phenomena** (mainly coreference in syntactic-semantic Unification)
Linearity	WM	Time and volume of temporary storage	**Sentence length** mainly, but also adjunction and coordination

Table 4: Constrainers and constrainees of the LeULi categories of neurolinguistic complexity and their metrics.

is being avoided. Ratio/Mean scores in these metrics should therefore be notably higher in *e2r* than in *orig* texts (contrary to pronoun incidence), and standard deviation should be the lowest possible (as in any *e2r* metric). Thus, coreference would be avoided in a consistent manner and the easy-to-read principle “use the same word for the same term" would be observed.

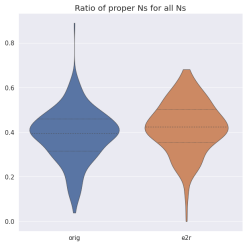

Figure 4: Plots of the ratio of proper nouns for all nouns (mean).

In this corpus, the ratio of proper nouns for all nouns is just slightly and hence insufficiently higher in *e2r* texts (Figure 4). Content word overlap (c.f. Figure 5) should not have a higher mean in *e2r* than in *orig* texts. Besides, it should not have a higher standard deviation in *e2r* than in *orig* texts.

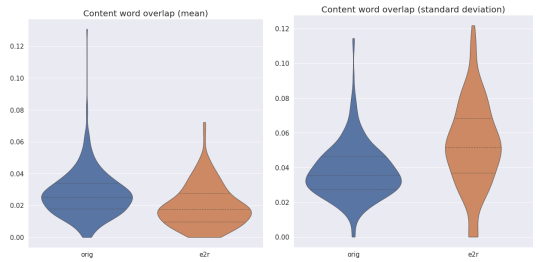

Figure 5: Plot of the content word overlap: mean (left) and standard deviation (right).

With respect to the Linearity category, Figure 6 displays the plots of sentence length and sentence depth. In both cases, there are lower mean scores in *e2r* texts and the difference between *orig* and *e2r* texts is similarly large for both metrics. These two are highly correlated metrics (Gutiérrez-Fandiño, 2022) and their plots are accordingly similar, but only sentence length actually contributes to processing complexity. This is because hierarchical structures (sentence depth, subordinate clauses) are effortlessly processed whereas linear phenomena that contribute to sentence length (number of words per sentence, coordination) are costly: WM consumption occurs in the horizontal extension of the syntactic tree, not in the vertical one.

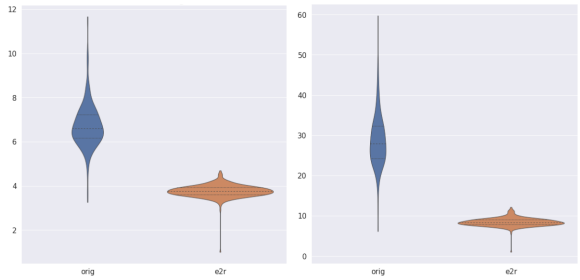

Figure 6: Plots of sentence depth (mean) (left) and words per sentence (mean) (right).

In Figure 7, it is shown that there is a large difference between *orig* and *e2r* texts in propositions per sentence than in subordinate clauses. Such difference accords with neurolinguistic simplicity since coordinated clauses should always be split into different sentences, to lighten the processing load, whereas subordinate clauses are not a problem themselves, as long as they do not include coreference.

In Figure 8, we see that there is a bigger difference in NP decendents (adjuncts+complements) than in NP modifiers (adjuncts) between *orig* and *e2r*. In terms of neurolinguistic simplicity, however, such difference should be bigger in adjuncts,

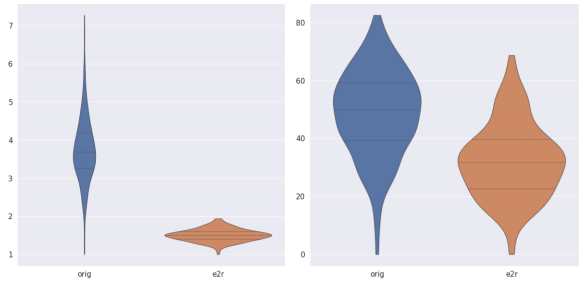

Figure 7: Plots of propositions per sentence (mean) (left) and subordinate clauses (incidence) (right).

which are non-selected constituents incurring an extra processing cost.

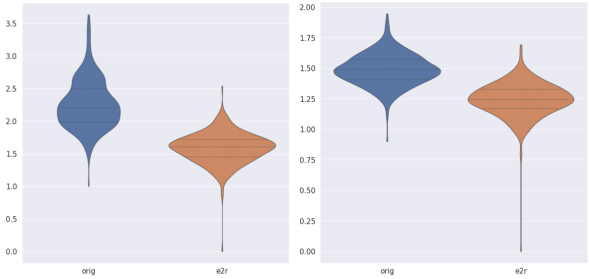

Figure 8: Plots of decendents per NP (mean) (left) and modifiers per NP (mean) (right).

In sum, taking into account the LeULi framework (Table 5), results on the word frequency metric are positive and hence the IrekiaLF_es corpus harmonises with neurolinguistic simplification as regards the Lexicon category of complexity. Regarding the Unification category, results are considerably worse: scores on 2 out of 3 metrics are halfway —not negative but highly improvable— and scores on the remaining metric are negative. Finally, results of the LeULi-based evaluation show positive scores in 2 out of 3 metrics of the Linearity category and negative scores on the remaining metric. Consequently, the text simplifications of the IrekiaLF_es corpus are good at the lexical and sentence level, but not at the discourse level owing to the substantial presence of multi-module phenomena (Unification category).

6 Experiments

In this section we report the results of three neural-based ATS systems when evaluated in the sentence-level IrekiaLF_es dataset. Given its small size, we followed a zero-shot scenario where the neural ATS systems are trained using Simplext (Saggion et al., 2015), and tested on IrekiaLF_es. Simplext comprises 200 manually simplified news texts from

Category	Metric	Assessment
Lexicon	Word frequency	POSITIVE
Unification	Ratio of proper nouns per all nouns	HALFWAY
	Pronouns (incidence)	HALFWAY
	Content word overlap (mean)	NEGATIVE
Linearity	Words per sentence	POSITIVE
	Coordination	POSITIVE
	NP adjunction	NEGATIVE

Table 5: Results of the LeULi-based evaluation of IrekiaLF_es.

different domains in Spanish. We have decided to train the baseline systems on Simplext because the text genre and the Spanish variety (peninsular Spanish) are similar. We followed Martin et al. (2020) and split the corpus in 574 sentences for training, 143 for development, and 693 for testing. We used the training set for finetuning, and the development set for model selection (the test split of Simplext was not used). Regarding IrekiaLF_es, we discarded the 1:0 alignments as well as the merge operations (29.3 % of the sentence level corpus), which results in a test set with 498 sentence pairs. The ATS systems are the following:

Edit+Synt is an edit-based seq2seq system that adds syntactic information at the word level (Cumbicus-Pineda et al., 2021). In the preprocessing stage, the training dataset was lowercased and the sentences were tokenised and parsed with SpaCy,[16] using the large model. The model was trained for 50 epochs, with a batch size of 64, a learning rate of 10^{-3}, a hidden dimension of 200, a decay factor of 10^{-6}.

mBART is a multilingual encoder-decoder model based on the transformers architecture, which is pretrained on 25 languages using the cc5 corpus (Liu et al., 2020). We used mBART-large, and fine-tuned it on Simplext (train set) for 50 epochs following default hyperparameters.[17]

mT5 is an encoder-decoder system similar to mBART but pre-trained using a different learning function and corpora (Xue et al., 2021). We used

[16] https://spacy.io/

[17] Learning rate of 5^{-5}, number of beams for beam search of 4, number of steps between val check of 500, number of steps between logs of 50, batch size of 2

System	BLUE	SARI
Edit+synt	5.44	37.95
mBART	4.38	38.90
mT5	7.12	42.19

Table 6: Results of the baseline systems.

the large version of mT5, and pretrained it on Simplext train using the same hyperparameters used in mBART.

We used default values for the hyperparameters, and did not perform any hyperparameter tuning. Regarding model selection, we selected the checkpoints that obtained the best SARI (Xu et al., 2016) score in the Simplext development test. To evaluate the models, we followed usual practice[18] and computed the BLUE (Papineni et al., 2002) and SARI metrics, using EASSE (Alva-Manchego et al., 2019).

In Table 6 we present the results obtained by the ATS systems. In general, all systems obtain SARI values that are similar to other ATS datasets, and in particular, to Simplext (Cumbicus-Pineda et al., 2021), with mT5 yielding the best results. While comparing figures across datasets cannot be used to draw meaningful conclusions, the relatively high SARI values might suggest the suitability of IrekiaLF_es for evaluating Spanish ATS systems. BLUE scores are however low in all systems, which we attribute to the followed zero-shot approach. Because Simplext simplifications are very short and highly compressed, the ATS systems produce sentences that are much shorter than the reference simplifications of IrekiaLS_es.

7 Conclusion and future work

In this paper, we have presented IrekiaLF_es, a new open corpus for Automatic Text Simplification in Spanish. The corpus compiles a document-level version with 288 parallel original and easy-to-read texts and a sentence-level version, where 35 of the documents have been manually aligned to create a test set of 705 sentences. The aim of this test set is to serve as a benchmark to evaluate ATS systems at the sentence level.

We have evaluated the neurolinguistic complexity of the corpus by following the Lexicon-Unification-Linearity (LeULi) model. The evaluation yields positive results regarding the Lexicon category of complexity, mostly negative regarding the Unification category and mostly positive regarding the Linearity category. Therefore, we can conclude that this corpus is suitable for ATS training and evaluation regarding lexical simplification and sentence simplification, but it may hinder end users' comprehension when it comes to discourse simplification due to the significant presence of multi-module phenomena (Unification category). This important drawback should be considered for future work. A specific quantitative benchmark for establishing the boundaries of the qualitative assessment (positive/halfway/negative) of the LeULi-based evaluation results should also be addressed in future work.

We have also evaluated three different systems that will serve as baselines for future research with this corpus. Results show good SARI values for all systems, but very low BLUE scores, which we attribute to the used zero-shot approach. Such results suggest the need of simplification datasets in Spanish where ATS systems can be trained or finetuned on.

In the future, we plan to align automatically the rest of the corpus by developing/adapting specific tools. From a linguistic point of view, we also foresee to study the operations carried out to simplify the texts. From an experimental point of view, we would like to carry out crosslingual experiments and test them in this corpus. Finally, we plan to create the Basque version of the corpus.

[18] We are aware that these metrics are flawed and may not be suitable for the quality evaluation of automatic simplifications but they are used as reference by the community (Sulem et al., 2018; Alva-Manchego et al., 2021).

Acknowledgements

This research has been partially funded by the Basque Government (Ixa excellence research center, IT1570-22) and the Spanish government (DeepKnowledge project, PID2021-127777OB-C21).

References

Rodrigo Alarcón García. 2022. Lexical simplification for the systematic support of cognitive accessibility guidelines. *https://doi.org/10.1145/3471391.3471400*.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020a. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics, pages 4668–4679, Online. Association for Computational Linguistics.

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. Easse: Easier automatic sentence simplification evaluation. In *EMNLP-IJCNLP 2019-Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (demo session)*, pages 49–54.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020b. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un) suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Marcelo Adriano Amancio and Lucia Specia. 2014. An analysis of crowdsourced text simplifications. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 123–130.

Alessia Battisti, Dominik Pfütze, Andreas Säuberli, Marek Kostrzewa, and Sarah Ebling. 2020. A corpus for automatic readability assessment and text simplification of german. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3302–3311.

Kepa Bengoetxea and Itziar Gonzalez-Dios. 2021. Multiaztertest: a multilingual analyzer on multiple levels of language for readability assessment. *arXiv preprint arXiv:2109.04870*.

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

Stefan Bott and Horacio Saggion. 2011. An unsupervised alignment algorithm for text simplification corpus construction. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 20–26.

Stefan Bott and Horacio Saggion. 2014. Text simplification resources for spanish. *Language Resources and Evaluation*, 48(1):93–120.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, and Giulia Venturi. 2016. Paccss-it: A parallel corpus of complex-simple sentences for automatic text simplification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 351–361.

Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2022. Linguistically-based comparison of different approaches to building corpora for text simplification: A case study on italian. *Frontiers in Psychology*, 13.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2015. Design and annotation of the first italian corpus for text simplification. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 31–41.

Leonardo Campillos-Llanos, Ana Rosa Terroba Reinares, Sofía Zakhir Puig, Ana Valverde Mateos, and Adrián Capllonch Carrión. 2022. Building a comparable corpus and a benchmark for spanish medical text simplification.

Helena M Caseli, Tiago F Pereira, Lucia Specia, Thiago AS Pardo, Caroline Gasperin, and Sandra M Aluisio. 2009. Building a brazilian portuguese parallel corpus of original and simplified texts. *Advances in Computational Linguistics*, page 59.

Raman Chandrasekar, Christine Doran, and Srinivas Bangalore. 1996. Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Oscar M Cumbicus-Pineda, Itziar Gonzalez-Dios, and Aitor Soroa. 2021. A syntax-aware edit-based system for text simplification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 324–334.

Daniel Ferrés and Horacio Saggion. 2022. Alexsis: A dataset for lexical simplification in spanish. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, page 3582–3594.

Christian J Fiebach, Angela D Friederici, Karsten Müller, and D Yves Von Cramon. 2002. fmri evidence for dual routes to the mental lexicon in visual word recognition. *Journal of cognitive neuroscience*, 14(1):11–23.

Núria Gala, Anaïs Tack, Ludivine Javourey-Drevet, Thomas François, and Johannes C Ziegler. 2020. Alector: A parallel corpus of simplified french texts with alignments of misreadings by poor and dyslexic readers. In *Language Resources and Evaluation for Language Technologies (LREC)*.

Itziar Gonzalez-Dios, María Jesús Aranzabe, and Arantza Díaz de Ilarraza. 2013. Testuen sinplifikazio automatikoa: arloaren egungo egoera. *Linguamática*, 5(2):43–63.

Itziar Gonzalez-Dios, María Jesús Aranzabe, and Arantza Díaz de Ilarraza. 2018. The corpus of basque simplified texts (cbst). *Language Resources and Evaluation*, 52(1):217–247.

Iker Gutiérrez-Fandiño. 2022. Toward complexity assessment in automatic text simplification: Evidence from neurobiological models of language. Bachelor's thesis. University of Deusto.

Peter Hagoort. 2005. On broca, brain, and binding: a new framework. *Trends in cognitive sciences*, 9(9):416–423.

Peter Hagoort. 2013. Muc (memory, unification, control) and beyond. *Frontiers in psychology*, 4:416.

Peter Hagoort. 2014. Nodes and networks in the neural architecture for language: Broca's region and beyond. *Current opinion in Neurobiology*, 28:136–141.

Peter Hagoort. 2017. The core and beyond in the language-ready brain. *Neuroscience & Biobehavioral Reviews*, 81:194–204.

Peter Hagoort. 2019. The neurobiology of language beyond single-word processing. *Science*, 366(6461):55–58.

Peter Hagoort. 2020. The core and beyond in the language-ready brain. Talk presented at Abralin ao Vivo – Linguists Online.

Nathan Hartmann, Gustavo Paetzold, and Sandra Aluísio. 2020. Simplex-pb 2.0: A reliable dataset for lexical simplification in brazilian portuguese. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 18–22.

Nathan S Hartmann, Gustavo H Paetzold, and Sandra M Aluísio. 2018. Simplex-pb: A lexical simplification database and benchmark for portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 272–283. Springer, Cham.

Renate Hauser, Jannis Vamvas, Sarah Ebling, and Martin Volk. 2022. A multilingual simplified language news corpus. In *2nd Workshop on Tools and Resources for REAding DIfficulties (READI)*, page 25.

Jerry L. Hintze and Ray D. Nelson. 1998. Violin plots: A box plot-density trace synergism. *The American Statistician*, 52(2):181–184.

David Klaper, Sarah Ebling, and Martin Volk. 2013. Building a german/simple german parallel corpus for automatic text simplification. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 11–19.

Sigrid Klerke and Anders Søgaard. 2012. Dsim, a danish parallel corpus for text simplification. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 4015–4018.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2020. Multilingual unsupervised sentence simplification. *CoRR*, abs/2005.00352.

Marina Nakic, Bruce W Smith, Sarah Busis, Meena Vythilingam, and R James R Blair. 2006. The impact of affect and frequency on lexical decision: the role of the amygdala and inferior frontal cortex. *Neuroimage*, 31(4):1752–1761.

OECD. 2013. *OECD Skills Outlook 2013*.

OECD. 2015. *OECD Skills Outlook 2015*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

David Pellow and Maxine Eskenazi. 2014. An open corpus of everyday documents for simplification tasks. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 84–93.

Sarah E Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *Workshop on speech and language technology in education*. Citeseer.

Horacio Saggion. 2017. Automatic text simplification. *Synthesis Lectures on Human Language Technologies*, 10(1):1–137.

Horacio Saggion, Elena Gómez-Martínez, Esteban Etayo, Alberto Anula, and Lorena Bourg. 2011. Text simplification in simplext: Making texts more accessible. *Procesamiento del lenguaje natural*, (47):341–342.

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. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):1–36.

Andreas Säuberli, Sarah Ebling, and Martin Volk. 2020. Benchmarking data-driven automatic text simplification for german. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 41–48.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.

Matthew Shardlow and Fernando Alva-Manchego. 2022. Simple tico-19: A dataset for joint translation and simplification of covid-19 texts. In *Proceedings of the 13th Language Resources and Evaluation Conference, Marseille, France, June. European Language Resources Association*.

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

Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. 2018. Luminosoinsight/wordfreq: v2.2.

Sanja Štajner. 2021. Automatic text simplification for social good: Progress and challenges. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-level text simplification: Dataset, criteria and baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7997–8013.

Sara Tonelli, Alessio Palmero Aprosio, and Francesca Saltori. 2016. Simpitiki: a simplification corpus for italian. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*.

Sowmya Vajjala and Ivana Lučić. 2018. Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 297–304.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361.

A Translation of the examples

In Table 7 we provide the translations into English of the examples of aligned sentences in Table 2.

B Alignment scales of the Italian and Basque corpora

In Table 8 we show the statistics in percentages of the Italian (IT) and Basque (EU) corpora as reported by Brunato et al. (2015) and Gonzalez-Dios et al. (2018) respectively.

orig	e2r
The Councilor for Employment and Social Policies, Beatriz Artolazabal, and the Councilor for Finance and Economy, Pedro Azpiazu, met with the heads of EHLABE-Euskal Herriko Lan Babestuaren Elkartea-Basque Association of non-profit organisations that promote the social and labor inclusion of people with disabilities, at the headquarters of Lantegi Batuak, in Loiu, Bizkaia.	Beatriz Artolazabal is the Councilor of Employment and Social Policies of the Basque Government. Pedro Azpiazu is the Councilor of Finance and Economy of the Basque Government. Euskal Herriko Lan Babestuaren Elkartea (EHLABE) is an association that promotes the inclusion of people with disabilities in society and at work. Beatriz Artolazabal and Pedro Azpiazu met with the heads of EHLABE at the headquarters of Lantegi Batuak, in Bizkaia.
During the meeting, the work of these companies was analyzed and new formulas for collaboration were proposed.	At the meeting they studied the work of these companies and proposed new ways to collaborate.

Table 7: English translations of the examples of aligned sentences in Table 2.

Alignment scale	Terence (IT)	Teacher (IT)	CBST- structural (EU)	CBST- intuitive (EU)
Merge (2:1)	2.88	13.74	0.88	0.44
1:0	0.67	1.15	-	-
1:1	92.1	68.32	76.21	73.25
1:2	3.75	11.45	18.50	19.74
1:3	0.19	0.76	3.52	4.39
Other	0.38	-	0.88	2.19

Table 8: Statistics (percentage) of the alignment scales of the Italian and Basque corpora.