Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3093–3102
Marseille, 20-25 June 2022

Simple TICO-19: A Dataset for Joint Translation and Simplification of COVID-19 Texts

Matthew Shardlow*, **Fernando Alva-Manchego**†
*Manchester Metropolitan University, †Cardiff University
m.shardlow@mmu.ac.uk, alvamanchegof@cardiff.ac.uk

Abstract

Specialist high-quality information is typically first available in English, and it is written in a language that may be difficult to understand by most readers. While Machine Translation technologies contribute to mitigate the first issue, the translated content will most likely still contain complex language. In order to investigate and address both problems simultaneously, we introduce Simple TICO-19, a new language resource containing manual simplifications of the English and Spanish portions of the TICO-19 corpus for Machine Translation of COVID-19 literature. We provide an in-depth description of the annotation process, which entailed designing an annotation manual and employing four annotators (two native English speakers and two native Spanish speakers) who simplified over 6,000 sentences from the English and Spanish portions of the TICO-19 corpus. We report several statistics on the new dataset, focusing on analysing the improvements in readability from the original texts to their simplified versions. In addition, we propose baseline methodologies for automatically generating the simplifications, translations and joint translation and simplifications contained in our dataset.

Keywords: Text Simplification, Machine Translation, Dataset, COVID-19

1. Introduction

The TICO-19 (Translation Initiative for COVID-19) dataset (Anastasopoulos et al., 2020) was developed in the wake of the COVID-19 pandemic as a benchmark for Machine Translation (MT) systems. It contains 3,000 sentences about COVID-19, translated from English into 36 languages to allow researchers, industry stakeholders and policy makers to make informed decisions about the abilities of MT technology for processing this type of specialised content.

The readability level of the texts in the TICO-19 dataset is generally high, with the English portion receiving a Flesch–Kincaid Grade Level (Kincaid et al., 1975) score of 12.14 for the development set and 12.06 for the test set.[1] These values indicate that the texts could be understood by those finishing a high school education (grade 12). The Flesch-Szigriszt Index (Szigriszt Pazos, 2001) for Spanish readability indicates that the development and test sets in the Spanish portion of the corpus are at a similar level of difficulty to the English sets. This is counter-intuitive as the TICO-19 initiative is designed to promote the wide distribution of scientific literature across speakers of different languages. However, if the translated version of the medical literature uses technical language that is inaccessible to a lay reader, then relevant information might not reach a wider audience in the general public.

Consider the example in Table 1. While the Spanish translation is faithful to the original English sentence, it can be difficult to understand, and reflects the medical language of the original text. However, the simplified translation could be more suitable for lay readers. For instance, the word "*subyacentes*" (underlying) in the original sentence is dropped, and the list of symptoms is replaced by the general term "*síntomas de COVID-19*" (symptoms of COVID-19). This translation is easier to read and records the most relevant original information. If the list of specific symptoms is important, this could be included as the next sentence.

Original (English)	*Persons with underlying health conditions who have symptoms of COVID-19, including fever, cough, or shortness of breath, should immediately contact their health care provider.*
Translation (Spanish)	*Las personas con afecciones médicas subyacentes que tienen síntomas de COVID-19, como fiebre, tos o disnea, deben comunicarse de inmediato con su proveedor de atención médica.*
Simplified Translation (Spanish)	*Las personas con afecciones médicas que tienen síntomas de COVID-19, deben comunicarse de inmediato con su proveedor de atención médica.*

Table 1: Example from the Simple TICO-19 dataset (ID: `PubMed_9:842`).

To better investigate how to produce translations of specialised content using "simple" language (also called joint translation and simplification), we introduce Simple TICO-19, a dataset with manual simplifications of the original English and translated Spanish subsections of the TICO-19 corpus.[2] This new resource will allow the study of the types of simplifications needed for medical English and Spanish (specifically related to the COVID-19 pandemic), as well as research on the challenges associated with translating from a "complex"

[1] Scores computed using TextStat `https://github.com/shivam5992/textstat`.

[2] Data and experimental code are available at `https://github.com/MMU-TDMLab/SimpleTICO19`

source language to a simplified target language.
After reviewing some related work (Sec. 2), we explain how we collected the manual simplifications in Simple TICO-19 (Sec. 3) and analyse their characteristics (Sec. 4). In addition, we study the performance of a strong Neural Machine Translation model in our dataset (Sec. 5), which serves as a baseline for future work. Furthermore, we propose three methodologies that may be explored when training systems for joint translation and simplification (Sec. 6). Finally, we outline the main conclusions of our work and provide ideas for future work (Sec. 7).

2. Related Work

Machine Translation. Neural Machine Translation (NMT) has become the dominating paradigm for developing translation models and achieving state-of-the-art results in most language pairs. Several frameworks and toolkits exist for implementing NMT models, such as MarianNMT (Junczys-Dowmunt et al., 2018), FAIRSEQ (Ott et al., 2019), and OpenNMT (Klein et al., 2017). MarianNMT, in particular, was originally developed for the WNMT2018 shared task (Birch et al., 2018), and uses an averaging attention network (Zhang et al., 2018) with teacher-student training (Kim and Rush, 2016). Models for MarianNMT are available in many language pairs, trained on the high-quality parallel data from the OPUS project (Tiedemann and Thottingal, 2020), which enables reliable translation between several language pairs. We test the performance of MarianNMT models in our new dataset.

Text Simplification. Automated Text Simplification (ATS) consists of rewriting texts into easier-to-read versions to improve their readability. For example, clinical letters in medical language can be rewritten into lay language to be better understood by patients (Shardlow and Nawaz, 2019). ATS systems are typically-based on end-to-end NMT approaches (Nisioi et al., 2017; Martin et al., 2020b; Martin et al., 2020a), hybrid architectures (Narayan and Gardent, 2014; Maddela et al., 2021), or edit-based frameworks (Dong et al., 2019; Omelianchuk et al., 2021), that are evaluated on purpose-built datasets for the task (Xu et al., 2016; Alva-Manchego et al., 2020).

Joint Translation and Simplification. At the interface of translation and simplification, some work has sought to develop NMT models that control the complexity of the generated translations. For instance, the output of a NMT system can be forced to be easier to understand than the original inputs (Agrawal and Carpuat, 2019), or can be controlled for varying levels of readability (Marchisio et al., 2019).

TICO-19. The TICO-19 dataset (Anastasopoulos et al., 2020) provides translation resources for health literature related to the coronavirus pandemic. It spans 36 languages and gives parallel translations for language pairs therein. Whilst it is possible to apply general domain translation resources for medical texts, they are generally low in quality as the domain specific language has not been learnt during training. Recent efforts have sought to adapt general language models for low-resource settings such as in TICO-19 (Vu et al., 2021; Ko et al., 2021). The inclusion of domain specific terminologies may help here, and new evaluation metrics have been recently proposed to determine their efficacy (Anastasopoulos et al., 2021).

Data Source	Domain	Sentences
CMU	medical, conversational	141
PubMed	medical, scientific	939
Wikinews	news	88
Wikivoyage	travel	243
Wikipedia	general	1,538
Wikisource	announcements	122
	Total	3,071

Table 2: Statistics of the TICO-19 benchmark.

3. Data Collection

We selected TICO-19 as our base corpus of original-translation sentence pairs in the domain of interest. This dataset contains health information related to the COVID-19 pandemic from a number of sources, ranging from formal language in academic publications to less formal settings, such as relevant speech corpora and news articles (see Table 2 for general statistics). TICO-19 provides parallel texts translated from English into 36 languages. In this work, we focused on the English-Spanish language pair.
For our purposes, we required a simplified version of the English and Spanish subsets of TICO-19. To achieve this goal, we engaged four translators: two native speakers of English and two native speakers of Spanish. All our translators were students at Manchester Metropolitan University, with knowledge and experience on editing and translating documents. We paid our translators at a rate of £11.35 per hour, which allowed us to annotate the entire corpus. Our annotators worked at an average rate of 26.4 sentences per hour.

3.1. Training of Annotators

Translators underwent a training session to get familiar with the data and the simplification task. We conducted several rounds of annotation and manual verification to ensure that the annotation guidelines were adequate and that the produced simplifications met the desired quality. The training process was as follows:

1. Annotation guidelines were provided, containing instructions on how to simplify sentences, explained examples, and practice sentences.
2. An online meeting was organised between the annotators and the project leaders to:
    - Collect comments/suggestions from the annotators regarding the guidelines.

Operations	Original (En)	Simplification (En)
Lexical Paraphrasing	*You may* ***be prohibited from changing*** *seats on the flight.*	*You may* ***not be allowed to change*** *seats on the flight.*
Lexical Paraphrasing + Splitting	***Thus*** *far, WIV1 represents the most closely related ancestor of SARS-CoV in bats****,*** ***sharing*** *95%* ***nucleotide sequence homology****.*	***So*** *far, WIV1 represents the most closely related ancestor of SARS-CoV in bats****.*** ***It shares*** *95%* ***of genetic similarity****.*
Compression	*All samples ~~collected~~ will be tested for ~~the presence of~~ influenza and COVID-19.*	*All samples will be tested for influenza and COVID-19.*

Table 3: Examples of the simplification operations that annotators were instructed to performed.

- Compare simplifications on practice sentences, and validate that the task was being understood.
- Answer any other type of questions that the annotators may have.

3. The guidelines were adjusted according to what was discussed in the meeting, and a new version was sent to the annotators. This new version included new practice sentences.

4. Steps 2 and 3 were repeated twice to ensure some level of understanding of the guidelines and expected outcome.

3.2. Annotation Guidelines

We gave the annotators a set of guidelines that defined a common style of expected simplifications. We centred these guidelines around three typical simplification operations: lexical paraphrasing, compression and splitting. **Lexical paraphrasing** allowed annotators to exchange a word or sequence for a simpler alternative if they felt it simplified the sentence. Annotators were encouraged to select meaning preserving swaps to the highest degree possible. **Compression** allowed annotators to drop words or clauses where the information being conveyed was either redundant or superfluous to the main text. **Splitting** is the operation of dividing a long and complex sentence into multiple parts. Typically, this is done where sentences contain multiple unrelated clauses, or explanations that can be better conveyed by distinct sentences. The three simplification operations could be performed in isolation or combined. Table 3 presents some examples. We invited our annotators to comment and provide feedback on the guidelines, and we updated these throughout the project as matters for clarification arose.[3]

[3]Annotation guidelines are available in our repository.

3.3. Annotation Process

As shown in Table 2, the TICO-19 dataset contains sentences from texts in various domains. Each domain can pose different challenges for manual simplification, mainly due to the vocabulary being used. As such, we conducted several annotation rounds, each focusing on different domains, starting from the most general data sources (Wiki*) and ending with the most technical ones (CMU and PubMed). Each round worked in the following way:

1. Each annotator for one language received half of the sentences from one specific data domain. After a certain number of days, the annotator sent the simplified sentences back to the project leaders.

2. The project leaders selected a sample of sentences from the submissions of each annotator for quality control. The sampled sentences of an annotator were sent to the other one of the same language, who verified that the simplifications were produced adequately (i.e. following the annotation guidelines) using a binary judgement. Any feedback was submitted to the project leaders and to the other annotator.

3. The project leaders provided feedback on the annotations in light of the quality control step, and asked annotators to update annotations based on the feedback where necessary.

Manual simplifications were performed in a monolingual setup, with annotators only having access to the original sentences in their native language. This means that annotators for Spanish did not have access to the original English versions, and directly simplified the Spanish translations available in the TICO-19 dataset. As as consequence, while original-Es sentences are direct translations of original-En sentences, that relationship is not necessarily preserved between simplified-Es and simplified-En sentence pairs.

Furthermore, our translators found that some of the sentences did not require simplification, and we allowed them to leave the sentences as they were in these cases. However, we required them to mark the cases that were deliberately left the same, allowing us to identify them when working with the corpus.

Following the round of annotation, the project leaders compiled the annotations into a single dataset, preserving the original splits for development and test from the TICO-19 corpus. We manually inspected the annotations and performed quality control operations such as correcting a few spelling or grammatical errors. For the English simplifications, one annotator was unable to complete the assigned tasks in the given time-frame, so the native English speaking author provided new simplifications for these, following the guidelines that had been used during the original annotation process.

Language	Complexity	Split	Instances	Words	W/S	Sy/W	FRE↑	S-P↑
English	Original	Dev	971	21,057	21.686	6.512	37.44	–
		Test	2,202	51,969	23.601	6.416	45.59	–
		All	3,173	73,026	23.015	6.444	45.69	–
	Simplified	Dev	971	19,321	19.898	6.365	53.21	–
		Test	2,202	49,970	22.693	6.286	52.49	–
		All	3,173	69,291	21.838	6.308	52.70	–
Spanish	Original	Dev	971	25,423	26.182	6.323	–	72.61
		Test	2,202	62,224	28.258	6.272	–	76.21
		All	3,173	87,647	27.623	6.287	–	75.17
	Simplified	Dev	971	22,836	23.518	6.324	–	77.83
		Test	2,202	55,694	25.292	6.250	–	80.19
		All	3,173	78,530	24.749	6.271	–	79.49

Table 4: Statistics on the completed corpus. For each language, complexity level and dataset split, we present: number of instances, total number of words, average number of words per sentence (W/S), average number of syllables per word (Sy/W), and the estimated readability level using Flesch Reading Ease (FRE) for English, and Szigriszt-Pazos (S-P) for Spanish.

As a result of this annotation process, we have a new corpus for Machine Translation that can be used to produce translated outputs at a reduced level of complexity. The original texts for English and Spanish from TICO-19 have been augmented with simplified counterparts. Each sentence has either a simplified version of that sentence that follows the operations we have specified, or a decision has been taken that the sentence is already sufficiently simple.

4. Dataset Analysis

We examine the characteristics of Simple TICO-19 in terms of improvements in readability achieved by the manual simplification process, as well as the nature of the simplification operations performed. Since we simplified the entire development and test splits of TICO-19 (all available data) for English and Spanish, the number of instances in the original and simplified portions of the corpus are equivalent. Where annotators deemed a sentence to be sufficiently simple, the original sentence was copied as the simplification. These cases have a special tag in the dataset.

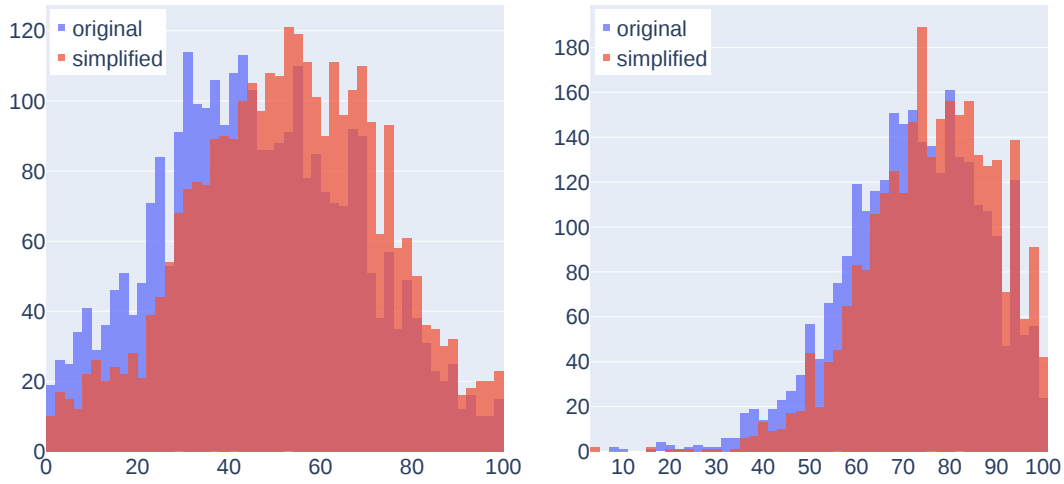
4.1. Readability Improvements

Table 4 shows some high level statistics of both the original and simplified sentences, indicating that simplification took place during the annotation process. The average sentence length in the English and Spanish portions goes down from 23.015 to 21.838 words per sentence for English and from 27.623 down to 24.749 for Spanish. Whilst the word length also goes down (6.444→6.308 for English and 6.287→6.271 for Spanish), the effect is less pronounced. To analyse the readability of the texts, we report Flesch Reading Ease (FRE) (Flesch, 1948) and Szigriszt-Pazos (S-P: a Spanish equivalent of FRE) (Szigriszt Pazos, 2001) for English and Spanish, respectively. Both of these language-specific metrics return a value between 0 and 100, with higher values indicating better readability. Both metrics take into account syllable count, which helps us understand the character of the words that are being used. Typically, simpler words have fewer syllables. In both languages, the readability formulae show that the simplified data portions are easier to read than their original counterparts. The English data shows an increase in FRE from 45.69 to 52.70, whereas the Spanish data improves from 75.17 to 79.49 in S-P.

It can also be observed that the original dev and test sets in English are slightly uneven in terms of sentence length. With the test set containing almost 2 words extra per sentence than the dev set. This is carried forwards into the simplified portions of dev and test. The word length is similar between the English dev and test sets. For the Spanish data, there is also a 2-word discrepancy between the average sentence length for dev and test in both the original and simplified versions, with test being higher. This is expected as the sentences were translated from English into Spanish and as such are likely to retain the same properties.

Analysis of the reading ease of the dev and test portions shows that simplification has some form of normalisation effect on the data. For English, dev (37.44) and dest (45.59) data are initially separated by 8.15 points according to FRE. After simplification, however, the metric demonstrates that the reading ease for each portion is now similar at 53.21 for dev and 52.49 for test, giving a difference of 0.72 points. The Spanish data behaves in the same way when we analyse the S-P metric, although this is less pronounced. Initially the dev and test portions are 3.6 points apart, which comes down to 2.36 points in the simplified portions.

We also analysed the distribution of FRE and S-P scores in the English and Spanish portions of the dataset, respectively. We calculated these metrics at sentence level across our entire corpus. Whilst these are designed for document level (i.e. multi-sentence),



(a) Histogram of FRE at sentence level in original (blue) and simplified (red) English sentences

(b) Histogram of P-S at sentence level in original (blue) and simplified (red) Spanish sentences

Figure 1: Histograms of reading ease in the Simple TICO-19 dataset.

calculating them at sentence level gives an indication of the contribution of each sentence to the overall score, and allows us to understand how the make-up of the corpus has changed through simplification. The results of these are displayed as histograms in Figure 1. In both histograms, the readability of the original sentences are displayed as blue, whereas the readability of the simplified sentences are displayed as red. Whilst the distributions are overlapping, it is clear in both cases that the simplified sentences have a higher mean distribution, indicating that they are generally easier to read. There seems to be more of a simplification effect taking place for English than for Spanish, but the Spanish sentences also appear to begin with a higher readability, so there is less ground to be gained. This may be the effect of transformations that have taken place during the translation process, leading the resulting Spanish texts to be more readable than the original English texts (inherent simplification during the translation process was recently leveraged to develop new Text Simplification systems (Lu et al., 2021)).

4.2. Simplification Operations

We also attempt to quantify the rewriting transformations that were performed to generate the manual simplifications in Simple TICO-19.

4.2.1. Labelling of Operations

Since we did not instruct the annotators to record the simplification operations they applied when simplifying, we used the annotation algorithms available in EASSE (Alva-Manchego et al., 2019) to automatically recognise the operations in all simplification instances in the dataset. These algorithms leverage word alignments between an original sentence and its simplification to identify replacements, deletions and additions. A *replacement* is when two words are aligned and they are not an exact match; a *deletion* is when a word in the original sentence is not aligned to any in the simplification; and an *addition* is when a word in the simplification is not aligned to any in the original sentence. These word-level annotations are further exploited to generate sentence-level operations labels: if at least one word was labelled with a particular simplification operation, then that operation is registered for the whole sentence. Furthermore, the algorithms compute the number of sentences in the original and simplification using NLTK (Bird et al., 2009),[4] and register a *splitting* if the number of sentences in the simplification is higher than in the original sentence. We configured EASSE to use SimAlign (Jalili Sabet et al., 2020) for extracting the word alignments required by the algorithms, since this setting achieved F1 scores ≥ 0.7 for the previously mentioned operations (Alva-Manchego, 2020, chap 3). In addition, because SimAlign relies on Multilingual BERT[5] to compute word similarity, this allows us to exploit the annotation algorithms for both the English and Spanish portions of the dataset.

4.2.2. Analysis

Figure 2 shows the percentages of simplification instances where a particular simplification operation was executed, for each dataset split and language of Simple TICO-19.[6] In the plot, instances where none of the four operations under study were automatically anno-

[4] https://www.nltk.org/api/nltk.tokenize.html

[5] https://github.com/google-research/bert/blob/master/multilingual.md

[6] For the plot, *replacements* and *deletions* are renamed to paraphrasing and compression, respectively, in order to maintain the terminology that has been used in the paper so far.

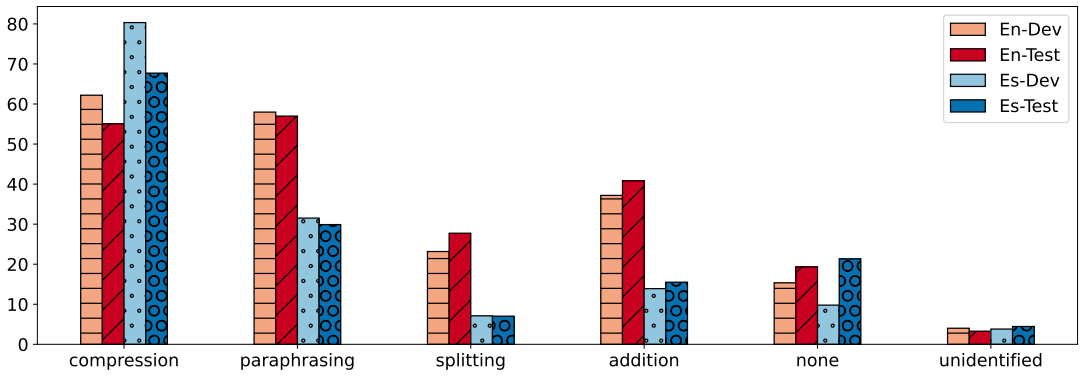


Figure 2: Percentage of simplified sentences that contain a particular simplification operation (automatically identified) in the development and test splits of Simple TICO-19 for both English and Spanish portions.

tated are either considered as: (1) **none** if the original sentence and its simplification are exactly the same (i.e. where the annotators decided not to simplify the original sentence), or (2) **unidentified** in any other case. Manual examination of the latter showed that they include instances where the annotators added or removed punctuation marks, as well as when annotators reordered words or phrases.

Compression and Lexical Paraphrasing are comparatively frequent in the English portion of the dataset, for both the development and test sets. This does not happen for Spanish, however, with Compression having a higher frequency than Lexical Paraphrasing (80% vs 30% in the development set, for instance). Through manual examination, we noticed that most of these cases are indeed instances where the annotators deleted content. However, some correspond to paraphrases like the one in Table 1, where a list of words is replaced by a general phrase that groups them. These cases are difficult to be identified automatically, so Compression is overestimated for Spanish. However, the fact that the algorithms identified more paraphrases in the English data, could indicate that English annotators preferred single word or short phrase replacements.

While Splitting is present in all dataset splits and languages, it has a higher frequency in the English portion than in the Spanish one. This could be a consequence of the annotations of the previous two operations: since Spanish annotators preferred to delete content or replace long lists with single grouping terms, they could have felt less inclined to split long sentences.

In the case of Addition, it has a higher percentage in the English data than in the Spanish one. Further analysis showed that this operation almost always appears together with the others, with words being added to preserve the grammatically of sentences, rather than to include explanations or examples.

Finally, almost all sentences in the data were simplified, with at most 20% of instances in a dataset split of a language being kept unchanged.

5. Baseline Models

We release sample results on the original and simplified sets of Simple TICO-19, which could serve as baselines for future work.

5.1. Experimental Settings

Models. Similar to the TICO-19 initiative, we experimented with NMT models trained on the OPUS dataset using MarianNMT. We performed experiments in both language pairs (En→Es and Es→En), making use of pre-trained models for general language (opus-mt-en-es and opus-mt-es-en) through the SimpleTransformers library.[7] We did not fine-tune our models for simplification or domain specificity. This allows us to observe the effects of using the simplified portions of the dataset alone. At inference time, we used a beam size of 12 and a max output length of 1,000.

Evaluation Metrics. We computed BLEU using SacreBLEU (Post, 2018)[8] with a maximum n-gram order of 4. We also calculated BERTScore (Zhang et al., 2020)[9] with its default parameters.

5.2. Results

We report automatic metrics' scores on the whole test set (*All*) and sub-genres therein (n.b. Wikivoyage is only present in the dev set, not test set). For each language pair (En → Es and Es → En), we experimented with using Original and Simplified as source and target "languages". However, we omit Simplified-to-Original pairs as this is not our aim (i.e. there is little societal benefit in developing a system that makes a text more complex than it was before).

Original as Source. Table 5 shows results when using the Original sentences as input to the models. We attained the highest BLEU and BERTScore values when using the original texts in the other language as references (i.e. the original setting of TICO-19). Scores

[7] https://github.com/ThilinaRajapakse/simpletransformers

[8] https://github.com/mjpost/sacrebleu

[9] https://github.com/Tiiiger/bert_score

	Original-En →				Original-Es →			
	Original-Es		Simplified-Es		Original-En		Simplified-En	
Data Source	BLEU	BERTScore	BLEU	BERTScore	BLEU	BERTScore	BLEU	BERTScore
CMU	33.51	0.678	17.05	0.581	30.82	0.734	29.41	0.683
PubMed	51.63	0.819	42.69	0.757	50.22	0.802	38.32	0.723
Wikinews	55.41	0.826	40.22	0.732	44.43	0.807	32.09	0.717
Wikipedia	52.16	0.875	44.83	0.836	48.91	0.878	38.64	0.833
Wikisource	39.98	0.715	31.85	0.647	42.16	0.775	31.70	0.702
All	51.42	0.841	43.15	0.788	48.66	0.842	38.02	0.783

Table 5: Results per data source of our baseline models on the test set of the Simple TICO-19 dataset, using Original sentences are input to the NMT models.

	Simp-En → Simp-Es		Simp-Es → Simp-En	
Data Source	BLEU	BERTScore	BLEU	BERTScore
CMU	18.34	0.563	22.13	0.640
PubMed	37.79	0.712	35.48	0.685
Wikinews	33.37	0.675	26.52	0.646
Wikipedia	40.46	0.806	36.00	0.805
Wikisource	29.92	0.602	29.05	0.659
All	38.57	0.752	35.09	0.749

Table 6: Results per data source of our baseline models on the test set of the Simple TICO-19 dataset, using Simplified sentences are input to the NMT models.

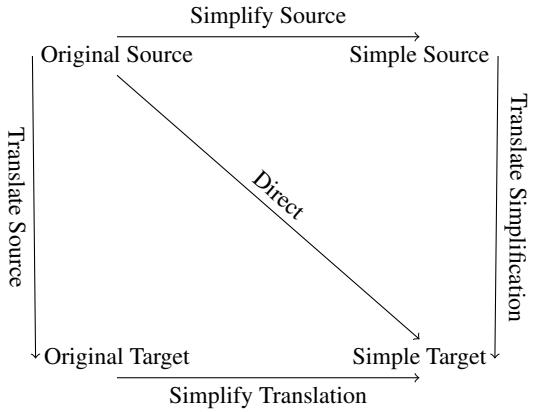


Figure 3: Potential routes to convert from an Original ("complex") source language (top-left) to a Simplified target language (bottom right). The edges of the square represent the pipeline of simplifying and then translating (across, then down) or translating then simplifying (down, then across). The diagonal arrow represents a model which directly produces Simplified target language, skipping the other steps.

drop when translations are compared to the simplified texts in Simple TICO-19. This confirms our hypothesis that the language produced by NMT models is more similar to the original (more "complex") texts than the simplified versions. The sub-genre evaluations show that the model(s) performed consistently better on PubMed and Wikpedia, with lower performance on the more informal texts found in the CMU speech data.

Simplified as Source. A limitation of our approach to simplifying translations is that Simplified-En and Simplified-Es are not direct translations of each other. Despite this, we investigated how well the models could perform when using Simplified sentences as input and as references. Table 6 shows results in these scenarios. Compared to translating from Original sentences, there is a drop in BLEU and BERTScore values for both language pairs and in all sub-genres but CMU.

6. Joint Simplification and Translation: Proposed Methodologies

The purpose of the Simple TICO-19 dataset is to allow research on joint translation and simplification of medical texts related to the COVID-19 pandemic. As shown in Sec. 5, state-of-the-art NMT models are unable to perform this task out-of-the-box. In this section, we propose several methodologies that could be explored in future work in order to implement systems that perform the task proposed by Simple TICO-19.

Joint translation and simplification is clearly two distinct operations on the input data, since texts must be transferred in both language and style. Figure 3 shows the potential routes that could be taken to go from an Original ("complex") language to a Simplified one. We identify three approaches and discuss them below.

Translate then Simplify. An appropriate solution to the task of joint simplification and translation would be to use a pipeline of NLP tools to first translate from the Original source language to the Original target language, and then simplify the resulting Original target language to a Simplified version of that language. This requires firstly an MT model capable of accurately translating from the source language to the target language, as well as a simplification model in the target language. Simplification comes in many forms and it may be possible to apply rule based approaches as post-processing to improve the readability of the final text, rather than using another deep learning approach.

Simplify then Translate. Conversely, we could use a pipeline to first simplify the Original source text and then translate the Simplified source language to the target language. As shown previously, using an MT model trained on Original inputs and outputs will not generate simplified directly. So, it is likely that an MT model trained specifically on simplified texts would be required to produce Simplified target language from Simplified source language.

Direct. A shortcoming of the above two approaches is that they rely on a pipeline of two operations, both of which may introduce errors independently. To mitigate this, it may be possible to train systems that directly translate from Original source language to Simplified target language. The first decision that must be made is whether to train a model from scratch, or to fine-tune an existing model. Whereas existing models are already capable of translating from source to target, they are trained on datasets that do not consider the complexity of the language. Fine-tuning them on appropriate corpora may help to improve the outputs by causing the model to prioritise simpler outputs over more complex ones. Training a model from scratch is always expensive and requires very large high-quality corpora to ensure model reliability. Corpora containing Original source language and Simplified target language are not readily available. It may be possible, however, to auto-generate these corpora by either simplifying an MT corpus or translating a simplification corpus. Either of these routes would lead to a silver standard corpus that could be used to either fine-tune or directly train a model.

7. Conclusions and Future Work

We have presented the Simple TICO-19 dataset, which was developed by manually simplifying the English and Spanish portions of the TICO-19 benchmark. We have reported on the process taken to develop this new dataset, and included statistics on the final version of the data. We also performed baseline experiments showcasing the ability of NMT models to replicate the information type in our simplified corpus, albeit at a reduced capacity compared to reproducing the original language style. We leave the development of models to accomplish this task to future work.

We selected English and Spanish as our language pair and translate in both directions. In the future, we plan to collect simplifications for other languages in the TICO-19 benchmark, considering the same set of original sentences. This would provide multiple language pairs and directions for joint translation and simplification, even enabling research on the task between language pairs that do not necessarily include English.

Figures 1a and 1b show the difference in readability between Original and Simplified English and Spanish in our dataset. It is clear that (a) the original Spanish texts are already reasonably readable; and that (b) there is more simplification effect in the English texts than the Spanish texts. The original Spanish texts were produced by professional translators, and it is likely that during the translation process some innate simplifications of the original texts were performed, leading to the overall increase in readability. Efforts to collect corpora reflecting true-sources rather than professional translation would lead to more complex sources, and a larger gap in readability between source and target.

Our NMT models performed in a similar range to those from the TICO-19 benchmark in the Original-to-Original setting. However, they scored lower when using Simplified texts as references. Future improvements should consider two factors: fine-tuning for simplicity and fine-tuning for genre-specificity. The opus-mt models that we used were trained on general purpose English-to-Spanish texts (and vice versa), so they are unlikely to do well when translating medical content. This is all true for the joint translation and simplification of medical texts, where a general-language model will fail to capture genre-specificities. It may well be the case that fine-tuning for the medical genre would improve the scores in both the Original-to-Original and Original-to-Simplified settings. However, we expect that any improvements gained by fine-tuning for genre specificity are most likely the result of better modelling of the genre of the source language as opposed to modelling its simplicity.

In our work we have taken translated texts and manually simplified them to obtain simplified translations. Another approach could be to take an existing corpus for Text Simplification (e.g. in English) and translate it into a target language (e.g. Spanish). This would have the same effect of providing us with a corpus of parallel simplifications and translations across languages. This would ensure that the translated simplifications are representative of the original source texts (complex or simple). However, in both the setting of simplified translations and translated simplifications, the resulting simple-target texts are the product of two rounds of translation. This means that they are further away from the original-source texts than either the original-target or simple-source texts, and represent a harder problem than either. Even if we employed translators to directly produce easier-to-read texts in a target language, these would still be the result of two tasks being applied.

Simple TICO-19 is the first dataset in the medical domain to contain simplified translations. It can be exploited for both joint translation and simplification, as well as monolingual simplification in two languages. As such, we hope this new dataset benefits the Machine Translation and Text Simplification communities.

8. Acknowledgements

This project was funded by the European Association for Machine Translation (EAMT) under its programme "2021 Sponsorship of Activities", and by the Centre for Advanced Computational Sciences at Manchester Metropolitan University.

9. Bibliographical References

Agrawal, S. and Carpuat, M. (2019). Controlling text complexity in neural machine translation. In *Proceedings of EMNLP-IJCNLP 2019*, pages 1549–1564, Hong Kong, China, November. ACL.

Alva-Manchego, F., Martin, L., Scarton, C., and Specia, L. (2019). EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of EMNLP-IJCNLP 2019: System Demonstrations*, pages 49–54, Hong Kong, China, November. ACL.

Alva-Manchego, F., Martin, L., Bordes, A., Scarton, C., Sagot, B., and Specia, L. (2020). ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of ACL 2020*, pages 4668–4679, Online, July. ACL.

Alva-Manchego, F. (2020). *Automatic Sentence Simplification with Multiple Rewriting Transformations*. Phd thesis, University of Sheffield, Sheffield, UK.

Anastasopoulos, A., Cattelan, A., Dou, Z.-Y., Federico, M., Federmann, C., Genzel, D., Guzmán, F., Hu, J., Hughes, M., Koehn, P., Lazar, R., Lewis, W., Neubig, G., Niu, M., Öktem, A., Paquin, E., Tang, G., and Tur, S. (2020). TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December. Association for Computational Linguistics.

Anastasopoulos, A., Besacier, L., Cross, J., Gallé, M., Koehn, P., Nikoulina, V., et al. (2021). On the evaluation of machine translation for terminology consistency. *arXiv preprint arXiv:2106.11891*.

Alexandra Birch, et al., editors. (2018). *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, Melbourne, Australia, July. Association for Computational Linguistics.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc.

Dong, Y., Li, Z., Rezagholizadeh, M., and Cheung, J. C. K. (2019). EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3393–3402, Florence, Italy, July. Association for Computational Linguistics.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3):221.

Jalili Sabet, M., Dufter, P., Yvon, F., and Schütze, H. (2020). SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online, November. Association for Computational Linguistics.

Junczys-Dowmunt, M., Heafield, K., Hoang, H., Grundkiewicz, R., and Aue, A. (2018). Marian: Cost-effective high-quality neural machine translation in c++. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 129–135.

Kim, Y. and Rush, A. M. (2016). Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas, November. Association for Computational Linguistics.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical Report 8-75, Naval Technical Training Command Millington TN Research Branch, February.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July. Association for Computational Linguistics.

Ko, W.-J., El-Kishky, A., Renduchintala, A., Chaudhary, V., Goyal, N., Guzmán, F., Fung, P., Koehn, P., and Diab, M. (2021). Adapting high-resource nmt models to translate low-resource related languages without parallel data. *arXiv preprint arXiv:2105.15071*.

Lu, X., Qiang, J., Li, Y., Yuan, Y., and Zhu, Y. (2021). An unsupervised method for building sentence simplification corpora in multiple languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 227–237, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Maddela, M., Alva-Manchego, F., and Xu, W. (2021). Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536–3553, Online, June. Association for Computational Linguistics.

Marchisio, K., Guo, J., Lai, C.-I., and Koehn, P. (2019). Controlling the reading level of machine translation output. In *Proceedings of MT Summit XVII*, pages 193–203, Dublin, Ireland, August. EAMT.

Martin, L., de la Clergerie, É. V., Sagot, B., and Bordes, A. (2020a). Controllable sentence simplification. In *LREC 2020-12th Language Resources and Evaluation Conference*.

Martin, L., Fan, A., de la Clergerie, É., Bordes, A., and Sagot, B. (2020b). Multilingual unsupervised sentence simplification. arXiv:2005.00352.

Narayan, S. and Gardent, C. (2014). Hybrid simplification using deep semantics and machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 435–445, Baltimore, Mary-

land, June. Association for Computational Linguistics.

Nisioi, S., Štajner, S., Ponzetto, S. P., and Dinu, L. P. (2017). Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91.

Omelianchuk, K., Raheja, V., and Skurzhanskyi, O. (2021). Text Simplification by Tagging. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–25, Online, April. Association for Computational Linguistics.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October. Association for Computational Linguistics.

Shardlow, M. and Nawaz, R. (2019). Neural text simplification of clinical letters with a domain specific phrase table. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 380–389.

Szigriszt Pazos, F. (2001). *Sistemas predictivos de legilibilidad del mensaje escrito: fórmula de perspicuidad*. Universidad Complutense de Madrid, Servicio de Publicaciones.

Tiedemann, J. and Thottingal, S. (2020). Opus-mt–building open translation services for the world. In *22nd Annual Conference of the European Association for Machine Translation*, page 479.

Vu, T., He, X., Phung, D., and Haffari, G. (2021). Generalised unsupervised domain adaptation of neural machine translation with cross-lingual data selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3335–3346.

Xu, W., Napoles, C., Pavlick, E., Chen, Q., and Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Zhang, B., Xiong, D., and Su, J. (2018). Accelerating neural transformer via an average attention network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1789–1798, Melbourne, Australia, July. Association for Computational Linguistics.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.