

# Controlling Text Complexity in Neural Machine Translation

**Sweta Agrawal**

Department of Computer Science  
University of Maryland  
sweagraw@cs.umd.edu

**Marine Carpuat**

Department of Computer Science  
University of Maryland  
marine@cs.umd.edu

## Abstract

This work introduces a machine translation task where the output is aimed at audiences of different levels of target language proficiency. We collect a high quality dataset of news articles available in English and Spanish, written for diverse grade levels and propose a method to align segments across comparable bilingual articles. The resulting dataset makes it possible to train multi-task sequence-to-sequence models that translate Spanish into English targeted at an easier reading grade level than the original Spanish. We show that these multi-task models outperform pipeline approaches that translate and simplify text independently.

## 1 Introduction

Generating text at the right level of complexity can make machine translation (MT) more useful for a wide range of users. As [Xu et al. \(2015\)](#) note, simplifying text makes it possible to develop reading aids for people with low-literacy ([Watanabe et al., 2009](#); [De Belder and Moens, 2010](#)), for non-native speakers and language learners ([Petersen and Ostendorf, 2007](#); [Allen, 2009](#)), for people who suffer from language impairments ([Carroll et al., 1999](#); [Canning et al., 2000](#); [Inui et al., 2003](#)), and for readers lacking expert knowledge of the topic discussed ([Elhadad and Sutaria, 2007](#); [Siddharthan and Katsos, 2010](#)). Such readers would also benefit from MT output that is better targeted to their needs by being easier to read than the original.

Prior work on text complexity has focused on simplifying input text in one language, primarily English ([Chandrasekar et al., 1996](#); [Coster and Kauchak, 2011](#); [Siddharthan, 2014](#); [Xu et al., 2015](#); [Zhang and Lapata, 2017](#); [Scarton and Specia, 2018](#); [Kriz et al., 2019](#); [Nishihara et al., 2019](#)). Simplification has been used to improve MT by restructuring complex sentences into shorter and

simpler segments that are easier to translate ([Gerber and Hovy, 1998](#); [Štajner and Popovic, 2016](#); [Hasler et al., 2017](#)). Contemporaneously to our work, [Marchisio et al. \(2019\)](#) show that tagging the English side of parallel corpora with automatic readability scores can help translate the same input in a simpler or more complex form. Our work shares the goal of controlling translation complexity, but considers a broader range of reading grade levels and simplification operations grounded in professionally edited text simplification corpora.

Building a model for this task ideally requires rich annotation for evaluation and supervised training that is not available in bilingual parallel corpora typically used in MT. Controlling the complexity of Spanish-English translation ideally requires examples of Spanish sentences paired with several English translations that span a range of complexity levels. We collect such a dataset of English-Spanish segment pairs from the Newsela website, which provides professionally edited simplifications and translations. By contrast with MT parallel corpora, the English and Spanish translations at different grade levels are only comparable. They differ in length and sentence structure, reflecting complex syntactic and lexical simplification operations.

We adopt a multi-task approach to control complexity in neural MT and evaluate it on complexity-controlled Spanish-English translation. Our extensive empirical study shows that multitask models produce better and simpler translations than pipelines of independent translation and simplification models. We then analyze the strengths and weaknesses of multitask models, focusing on the degree to which they match the target complexity, and the impact of training data types and reading grade level annotation.<sup>1</sup>

---

<sup>1</sup>Researchers can request the bilingual Newsela data

## 2 Background

Given corpora of parallel complex-simple segments, text simplification can naturally be framed as a translation task, borrowing and adapting model architectures originally designed for MT. Xu et al. (2016) provide a thorough study of statistical MT techniques for English text simplification, and introduce novel objectives to measure simplification quality. Interestingly, they indirectly make use of parallel translation corpora to derive simplification paraphrasing rules by bilingual pivoting (Callison-Burch, 2007). Zhang and Lapata (2017) train sequence-to-sequence models to translate from complex to simple English using reinforcement learning to directly optimize the metrics that evaluate complexity (SARI) and fluency and adequacy (BLEU). Scarton and Specia (2018) address the task of producing text of varying levels of complexity for different target audiences. They show that neural sequence-to-sequence models informed by target-complexity tokens inserted in the input sequence perform well on this task. While the vast majority of text simplification work has focused on English, Spanish (Štajner et al., 2015), Italian (Brunato et al., 2016; Aprosio et al., 2019) and German (Klaper et al., 2013) have also received some attention.

While most MT approaches only indirectly capture style properties (e.g., via domain adaptation), a growing number of studies share the goal of considering source texts and their translation in their pragmatic context. Mirkin and Meunier (2015) introduce personalized MT. Rabinovich et al. (2016) and Vanmassenhove et al. (2018) suggest that the gender of the author is implicitly marked in the source text and that dedicated statistical and neural systems better preserve gender traits in MT output. Neural MT has enabled more flexible ways to control stylistic properties of MT output. Sennrich et al. (2016) first propose to append a special token to the source that neural MT models can attend to and to select the formal (Sie) or informal (du) version of second person pronouns when translating into German. Niu et al. (2018) show that multi-task models can jointly translate between languages and styles, producing formal and informal translations with broader lexical and phrasal

changes than the local pronoun changes in Sennrich et al. (2016). Closest to our goal, Marchisio et al. (2019) address the task of producing either simple or complex translations of the same input, using automatic readability scoring of parallel corpora. They show that training distinct decoders for simple and complex language allows better complexity control than using the target complexity as a side-constraint. By contrast, our approach exploits text simplification corpora for richer supervision for both training and evaluation.

## 3 A Multi-Task Approach to Complexity Controlled MT

**Task** We define **complexity controlled MT** as a task that takes two inputs: an input language segment  $s_i$  and a target complexity  $c$ . The goal is to produce a translation  $s_o$  in the output language that has complexity  $c$ . For instance, given input Spanish sentences in Table 1, complexity controlled MT aims to produce English translations at a specific level of complexity, which might differ from the complexity of the original Spanish.

**Model** We model  $P(s_o|s_i, c; \theta)$  as a neural encoder-decoder with attention (Bahdanau et al., 2015). This architecture has been used successfully for the two related tasks of text simplification (Wang et al., 2016; Zhang and Lapata, 2017; Nisioi et al., 2017; Scarton and Specia, 2018) and machine translation (Bahdanau et al., 2015). The encoder constructs hidden representation for each word in the input sequence, while the decoder generates the target sequence, conditioned on hidden source representations. We hypothesize that training a single encoder-decoder model to perform the two distinct tasks of machine translation and text simplification will yield a model that can perform complexity controlled MT. We adopt the multi-task framework proposed by Johnson et al. (2016) to train multilingual neural MT systems.

**Representing target complexity** Target complexity  $c$  can be incorporated in sequence-to-sequence models as a special token appended to the beginning of the input sequence, which acts as a side constraint. The encoder encodes this token in its hidden states as any other vocabulary token, and the decoder can attend to this representation to guide the generation of the output sequence. This simple strategy has been used in MT to control second person pronoun forms when translat-

---

at <https://Newsela.com/data/>. Scripts to replicate our model configurations and our cross-lingual segment aligner are available at <https://github.com/sweta20/ComplexityControlledMT>.

$c_i$	Spanish ( $s_i$ )	$c_o$	English ( $s_o$ )	Operation
9	Doug Ratliff, un empresario de 67 años de edad de Richlands, Virginia, dijo que la elección de Trump sería uno de los días más felices de su vida.	3	Doug Ratliff is a businessman from Virginia. Ratliff said Trump’s election would be one of the happiest days of his life.	Splitting; Deletion
12	Incluso antes de haber nacido, Daliyah Marie Arana, según dicen sus padres, estaba aprendiendo a leer.	4	Daliyah Marie Arana has been learning to read since before she was born.	Paraphrasing
9	Kes Gray es el escritor de la serie de cuentos de animales "Oi Frog and Friends". A él no le interesaron mucho los descubrimientos del estudio. Los autores y los ilustradores solo necesitan que los personajes principales animales de sus historias sean adorables, concluyó.	5	Kes Gray is the writer of the rhyming animal series "Oi Frog and Friends." He was not bothered by the study’s findings. Writers and artists just need to keep the main animal characters in their stories cuddly, he said.	Lexical substitution

Table 1: Cross-lingual Newsela examples: the Spanish text  $s_i$  of complexity, or reading grade level,  $c_i$  is automatically aligned to English text  $s_o$  of  $c_o$ . Simplification transformations range from sentence splitting and deletions to paraphrasing and lexical substitution.

ing into German (Sennrich et al., 2016), to indicate the target language in multilingual MT (Johnson et al., 2016), and to obtain formal or informal translations of the same input (Niu et al., 2018). In monolingual text simplification tasks (Scarton and Specia, 2018), the reading grade level has been encoded as such a special token.

**Training Data and Objectives** Fully supervised training would ideally require translation samples with outputs representing different levels of complexity for the same input segment. However, constructing such data at the scale required to train deep neural networks is expensive and unrealistic. Our multi-task training configuration lets us exploit different types of training examples to train shared encoder-decoder parameters  $\theta$ . We use the following samples/tasks:

- Complexity controlled MT samples ( $s_i, c_o, s_o$ ): These are the closest samples to the task at hand, but are hard to obtain. They are used to define the complexity-controlled MT loss

$$\mathcal{L}_{CMT} = \sum_{(s_i, c_o, s_o)} \log P(s_o | s_i, c_o; \theta) \quad (1)$$

- MT samples ( $s_i, s_o$ ): These are sentence

pairs drawn from parallel corpora. They are available in large quantities for many language pairs (Tiedemann, 2012) and are used to define the MT loss

$$\mathcal{L}_{MT} = \sum_{(s_i, s_o)} \log P(s_o | s_i; \theta) \quad (2)$$

- Text simplification samples in the MT target language ( $s_o, c_{s'_o}, s'_o$ ) where  $s'_o$  is a simplified version of complexity  $c_{s'_o}$  for input  $s_o$ , which are likely to be available in much smaller quantities than MT samples.

$$\mathcal{L}_{Simplify} = \sum_{(s_o, c_{s'_o}, s'_o)} \log P(s'_o | s_o, c_{s'_o}; \theta) \quad (3)$$

The multi-task loss is simply obtained by summing the losses from individual tasks:  $\mathcal{L}_{CMT} + \mathcal{L}_{MT} + \mathcal{L}_{Simplify}$ .

## 4 The Newsela Cross-Lingual Simplification Dataset

We build on prior work that used the Newsela dataset for English or Spanish text simplification by automatically aligning English and Spanish segments of different complexity to enable complexity-controlled machine translation.

The Newsela website provides high quality data to study text simplification. Xu et al. (2015) argue that text simplification research should be grounded in texts that are simplified by professional editors for specific target audiences, rather than more general-purpose crowd-sourced simplifications such as those available on Wikipedia. They show that Wikipedia is prone to sentence alignment errors, contains a non-negligible amount of inadequate simplifications, and does not generalize well to other text genres. By contrast, Newsela is an instructional content platform meant to help teachers prepare curriculum that match the language skills required at each grade level. The Newsela corpus consists of English articles in their original form, 4 or 5 different versions rewritten by professionals to suit different grade levels as well as optional translations of original and/or simplified English articles into Spanish resulting in 23,130 English and 5,320 Spanish articles with grade annotations respectively.

This section introduces our method to align English and Spanish segments across complexity levels, and the resulting bilingual dataset.

#### 4.1 Cross-Lingual Segment Alignment

Extracting training examples from this corpus requires aligning segments within documents. This is challenging because text is neither simplified nor translated sentence by sentence, and as a result, equivalent content might move from one sentence to the next. Past work has introduced techniques to align segments of different complexity within documents of the same language (Xu et al., 2015; Paetzold et al., 2017; Štajner et al., 2018).

Complexity controlled MT requires aligning segments of different complexity in English and Spanish. Existing methods for aligning sentences in English and Spanish parallel corpora are not well suited to this task. For instance, the Gale-Church algorithm (Gale and Church, 1993) assumes that aligned sentences should have similar length. This assumption does not hold if the English article is a simplification of the Spanish article. Consider the following Spanish text and its English translation in Newsela:

**Spanish:** LA HAYA, Holanda - Te has tomado alguna vez una selfie?, Hoy en día es muy fácil. Solo necesitas un teléfono inteligente.

**Google Translated English:** THE HAGUE, Netherlands - Have you ever taken a selfie? Today

is very easy. You only need a smart phone.

**Original English Version:** THE HAGUE, Netherlands - All you need is a smartphone to take a selfie. It is that easy.

As a result, we adapt a monolingual text simplification aligner for cross-lingual alignment. MASSAlign (Paetzold et al., 2017) is a Python library designed to align segments of different length within comparable corpora of the same language. It employs a vicinity-driven search approach, based on the assumption that the order in which information appears is roughly constant in simple and complex texts. A similarity matrix is created between the paragraphs/sentences of aligned documents/paragraphs using a standard bag-of-words TF-IDF model. It finds a starting point to begin the search for an alignment path, allowing long-distance alignment skips, capturing 1-N and N-1 alignments. To leverage this alignment flexibility, we apply MASSAlign to English articles and Spanish articles machine translated into English by Google translate.<sup>2</sup> An important property of Google translated articles is that they are aligned 1-1 at the sentence level. This lets us deterministically find the Spanish replacement for the aligned Google translated English version returned by MASSAlign. Translation quality is high for this language pair, and even noisy translated articles contain enough signal to construct the similarity matrix required by MASSAlign.

#### 4.2 Resulting Dataset

We thus create: both samples for complexity controlled MT ( $s_i, c_o, s_o$ ) and traditional monolingual text simplification samples ( $s_o, c_{s'_o}, s'_o$ ) that can be used by the multi-task model (Section 3). Since the properties of Newsela monolingual simplification samples have been studied thoroughly by Xu et al. (2015), we present key statistics for the cross-lingual simplification examples only. Table 2 contrasts Newsela parallel segments with bilingual parallel sentences drawn from the OPUS corpus (Tiedemann, 2009). We use Global Voices and News Commentary from OPUS corpus as it has the most similar domain to the Newsela data. Aligned segments in Newsela are about twice as long as segments in parallel corpora, and contain more than two sentences on each side on average. By contrast, parallel corpora samples align sentences one-to-one on average.

<sup>2</sup><https://translate.google.com/>

Dataset	# tokens/segment	#sents/segment	# of types	# of tokens
<i>Spanish</i>				
Newsela	50.13	2.17	57,361	7,792,285
Global Voices	22.96	1.03	254,111	15,921,948
News	26.73	1.03	80,840	5,587,307
<i>English</i>				
Newsela	43.37	2.65	39,012	7,139,717
Global Voices	21.93	1.06	222,383	15,208,054
News	23.76	1.04	49,589	4,939,085

Table 2: Comparisons of the English-Spanish Newsela corpus with machine translation corpora from OPUS drawn from Global Voices and News Commentary.

Articles are distributed across reading levels spanning grades 2 to 12 for both English and English-Spanish pairs. Table 3 highlights the vocabulary differences among the different grade levels for the Newsela Spanish-English corpus. The vocabulary size of the corpus corresponding to lower grade level is smaller as compared to higher complexity levels. Also, complex sentences have more words per sentence on average but fewer sentences per segment compared to their simplified counterparts. Simple sentences differ from complex sentences in various ways, ranging from sentence splitting and content deletion to paraphrasing and lexical substitutions, as illustrated in Table 1.

## 5 Experiment Settings

We evaluate **complexity controlled MT** using a subset of the 150k Spanish-English segment pairs extracted from Newsela as described in Section 4. We select Spanish and English segments that have different reading grade levels, so that given a Spanish input, the task consists in producing an English translation which is simpler (lower reading grade level) than the Spanish input. The train/development/test split ensures that there is no overlap between articles held out for testing and articles used for training. We refer to the corresponding training examples as **MT+simplify** since it represents the joint task of translation and simplification.

### 5.1 Evaluation Metrics

We evaluate the truecased detokenized output of our models using three automatic evaluation metrics, drawing from both machine translation and text simplification evaluation.

BLEU (Papineni et al., 2002) estimates translation quality based on  $n$ -gram overlap between system output and references. However it does not separate mismatches due to meaning errors and mismatches due to simplification errors.

SARI (Xu et al., 2016)<sup>3</sup> is designed to evaluate text simplification systems by comparing system output against references and against the input sentence. It explicitly measures the goodness of words that are added, deleted and kept by the systems. Xu et al. (2016) showed that BLEU shows high correlation with human scores for grammaticality and meaning preservation and SARI shows high correlation with human scores for simplicity. In the cross-lingual setting, we cannot directly compare the Spanish input with English hypotheses and references, therefore we use the baseline machine translation of Spanish into English as a pseudo-source text. The resulting SARI score directly measures the improvement over baseline machine translation.

In addition to BLEU and SARI, we report Pearson’s correlation coefficient (PCC) to measure the strength of the linear relationship between the complexity of our system outputs and the complexity of reference translations. Heilman et al. (2008) use it to evaluate the performance of reading difficulty prediction. Here we estimate the reading grade level complexity of MT outputs and reference translations using the Automatic Readability Index (ARI)<sup>4</sup> score, which combines evidence from the number of characters per word and number of words per sentence using hand-tuned

<sup>3</sup><https://github.com/cocoxu/simplification>

<sup>4</sup><https://github.com/mmautner/readability>

Grade	Source (Spanish)			Target (English)		
	word types	tokens/segment	sents/segment	word types	tokens/segment	sents/segment
2	-	-	-	3749	34.31	3.76
3	2,628	38.59	3.52	10,615	35.57	3.28
4	8,431	39.33	2.95	10,414	39.38	3.09
5	17,082	40.96	2.59	18,508	42.18	2.87
6	16,945	42.78	2.25	16,613	44.21	2.62
7	22,352	46.65	2.19	23,617	47.54	2.57
8	19,317	47.07	1.96	17,746	46.66	2.24
9	24,846	50.08	1.87	22,230	50.08	2.25
10	482	49.19	1.90	341	38.23	1.70
12	42,355	53.98	1.96	-	-	-

Table 3: Grade level Statistics of the Newsela Spanish-English corpus. Vocabulary size decreases with the reading grade level. Simpler segments contain fewer sentences and are often shorter than complex segments.

weights (Senter and Smith, 1967):

$$ARI = 4.71\left(\frac{chars}{words}\right) + 0.5\left(\frac{words}{sents}\right) - 21.43 \quad (4)$$

## 5.2 Training Data

In addition to the Newsela **MT+Simplify** training examples described above, which are of the form  $(s_i, c_o, s_o)$ , we use monolingual English simplification data, bilingual parallel training data and Spanish simplification data.

**Newsela Simplification** provides training examples of the form  $(s_o, c_{s'_o}, s'_o)$ , where  $s_o$  and  $s'_o$  are in the same language. We refer to this data as **Simplify** data. It is used for training multi-task models and for auxiliary evaluation on English only. Our version of this corpus has 513k English segment pairs extracted using the method by Paetzold and Specia (2016). Similar to Scarton and Specia (2018), an original article 0 can be aligned to up to four simplified versions: 1,2,3 and 4. Here 4 denotes the least simplified level and 0 represents the most simplified level. The train split consists of 460k instance pairs whereas the development and test sets consist of roughly 20K instances, drawn from the same articles as the MT+preserve and MT+simplify test set. For Spanish, we have 110k segment pairs, which will be used to train the Spanish simplification baseline.

**Bilingual Parallel Data (Newsela)** We also extract parallel Spanish-English segments from Newsela based on aligned segments between Spanish and English articles that have the same

reading grade level. We use this dataset to provide in-domain MT training examples which includes roughly 70k instances.

All datasets are pre-processed using Moses tools for normalization, tokenization and true-casing (Koehn et al., 2007). We further segment tokens into subwords using a joint source-target byte pair encoding model with 32,000 operations (Sennrich et al., 2015).

## 5.3 Sequence-to-Sequence Model Configuration

We use the standard encoder-decoder architecture implemented in the Sockeye toolkit (Hieber et al., 2017). Both encoder and decoder have two Long Short Term Memory (LSTM) layers (Bahdanau et al., 2015), hidden states of size 500 and dropout of 0.3 applied to the RNNs of the encoder and decoder which is same as what was used by Scarton and Specia (2018). We observe that dot product based attention works best in our scenario, perhaps indicating that the task of complexity controlled translation requires mostly local changes that do not lead to long distance reorderings across sentences. We train using the Adam (Kingma and Ba, 2014) optimizer with a batch size of 256 segments and checkpoint the model every 1000 updates. Training stops after 8 checkpoints without improvement of validation perplexity. The vocabulary size is limited to 50000. We decode with a beam size of 5. Grade side-constraints are defined using a distinct special token for each grade level (from 2 to 12). The constraint token corresponds to the grade level of the target instance.

## 5.4 Baseline

We contrast the multi-task system with pipeline based approaches, where translation and simplification are treated as independent consecutive steps. We train a neural MT model to perform translation from Spanish to English and other neural MT models to perform monolingual text simplification for Spanish and English respectively. In the first pipeline setup, the output from the translation model is fed as input to an English simplification model while in the other, the output from the Spanish simplification model is fed as input to an translation model. As Scarton and Specia (2018), we simply use grade level tokens as side constraints on English simplification examples to control output complexity.<sup>5</sup>

## 6 Evaluation of Complexity Controlled MT

We compare pipeline and multitask models on the Newsela complexity controlled MT task (Table 4). Overall, results show that compared to pipeline models, multitask models produce complexity controlled translations that better match human references according to BLEU. SARI suggests that multitask translations are simpler than baseline translations, and their resulting complexity correlates better with reference grade levels according to PCC.

The two pipeline models use the same MT system, therefore the difference between them comes from text simplification: using English simplification (first pipeline) outperforms Spanish simplification (second pipeline) according to BLEU and PCC, but not SARI. This can be explained by the smaller amount of Spanish simplification training data, which yields a model that generalizes poorly.

The “All tasks” model highlights the strengths of the multi-task approach: combining training samples from many tasks yields improvements over the “Translate and Simplify” multi-task model which is trained on the exact same data as the pipelines. However, even without additional training data, the multitask “Translate and Simplify” model improves over baselines mainly by simplifying the output more, which suggests that the simplification component of the multitask model benefits from the additional MT training data.

<sup>5</sup>Additional constraints based on simplification operations were also used in that work but did not provide substantial benefits when operations are predicted based on the input.

Qualitative analysis suggests that the multi-task model is capable of distinguishing among different grade levels and the simplification operations performed for different grade levels are gradual. Table 5 illustrates simplification operations observed for a fixed grade 12 Spanish input into English with target grade levels ranging from 9 to 3. When translating to a nearby grade level, for example 9, the model roughly translates the entire input. For lower grade levels such as 7 and 5, lexical simplification and sentence splitting is observed. For the simplest grade level, the model deletes additional content. More examples are provided in the Appendix (Table 13).

Complexity cont. MT	BLEU	SARI	PCC
<i>Pipeline Baselines</i>			
Translate then Simplify	21.98	30.4	0.436
Simplify then Translate	17.09	37.4	0.275
<i>Multitask Models</i>			
Translate and Simplify	22.51	44.8	0.572
All Tasks	<b>22.75</b>	<b>45.0</b>	<b>0.608</b>

Table 4: Compared to pipeline models, multitask models produce complexity controlled translations that better match human references (BLEU), that are simpler (SARI), and whose resulting complexity correlates better with the target grade level (PCC). Pipeline models are trained on Newsela Simplification data and MT parallel data from Newsela and OPUS. “Translate and Simplify” uses the exact same data in a multi-task model. The “All tasks” model uses all data available, including Newsela MT+Simplify examples.

## 7 Analysis

### 7.1 Output Grade Analysis

We aim to better understand to what degree models simplify the input text: how often does the output complexity exactly matches that of the reference? Does this change depend on the distance between input and output complexity levels? Table 6 compiles Adjacency Accuracy scores (Heilman et al., 2008), which represent the percentage of sentences where the system output complexity is within 1 or 2 grades of the reference text. We derive the reading grade levels from ARI (Senter and Smith, 1967) and conduct this analysis for the best pipeline (“Translate then Simplify”) and multitask models (“All Tasks”). These adjacency scores are broken down according to the distance between input and target grade levels.

12	Ahora el museo Mauritshuis está por inaugurar una exposición dedicada a los autorretratos del siglo XVII, que destaca las similitudes y diferencias entre las fotos modernas y las obras de arte históricas.
9	Now the museum Mauritois is launching an exhibition dedicated to the 18th century author-itations, highlighting the similarities and differences between modern photos and historical artworks.
7	The museum is <b>now set to open</b> an exhibition dedicated to the 18th century authoritations, highlighting the similarities and differences between modern photos and historical artworks.
5	The museum is now set to open an exhibit dedicated to the 18th century. <b>It highlights</b> the similarities and differences between modern photos and historical artworks.
3	The museum is now set to open an exhibit dedicated to the 18th century. It <b>shows</b> the similar-ities and differences between modern photos and art works.

Table 5: Example of multi-task model outputs when translating grade 12 Spanish into increasingly simpler English.

Adj. level	Model	Source Grade - Target Grade									
		1	2	3	4	5	6	7	8	9	10
1	Pipeline	<b>0.593</b>	<b>0.629</b>	0.594	0.556	0.524	0.493	0.472	0.457	0.448	0.444
	Multitask	0.59	0.626	0.594	<b>0.561</b>	<b>0.529</b>	<b>0.504</b>	<b>0.482</b>	<b>0.467</b>	<b>0.458</b>	<b>0.453</b>
2	Pipeline	<b>0.717</b>	<b>0.759</b>	<b>0.786</b>	0.747	0.713	0.678	0.654	0.637	0.626	0.621
	Multitask	0.711	0.755	0.784	<b>0.753</b>	<b>0.725</b>	<b>0.696</b>	<b>0.67</b>	<b>0.653</b>	<b>0.642</b>	<b>0.636</b>

Table 6: Adjacency ARI accuracy within grade level given by Adjacency level for the system output with respect to the target grade: Multitask model is able to better capture the target grade than the pipeline model when the difference between the source and the target grade is greater than 3.

When the source and target grades are close, roughly 60% of system outputs that are within a  $\pm 1$  window of the correct grade level. The pipeline model matches the target grade slightly better than the multitask model. However, in the more difficult case where the difference between source and target grades is larger than three, the multitask model outperforms the pipeline. Increasing the adjacency window to  $\pm 2$  pushes the percentage of matches in the 70s.

Overall these results show that multitask and pipeline models are able to translate and simplify, but that they do not yet fully succeed at precisely controlling the complexity of their output to match a specific target reading grade.

## 7.2 Ablation Experiments

Table 7 shows the impact of different training data types on the multitask model using ablation experiments. OPUS improves BLEU and SARI performance across the board. However, using OPUS without any Newsela MT data (Row 4) hurts the correlation score, indicating the importance of in-domain MT data to control complex-

ity. The difference between the performance when using joint translation and simplification (MT+S) examples (Row 2) vs. simplification only (S in Row 3) is small in terms of BLEU (+0.11) and PCC (0.012), indicating that the monolingual simplification dataset can provide simplification supervision when MT+Simplify data is unavailable. The overall best performance for the task is obtained by using all types of training examples.<sup>6</sup>

## 7.3 Evaluation on Auxiliary Tasks

In addition to complexity controlled MT, the multi-task model can be used to simplify English text, and to translate from Spanish-to-English without changing the complexity. For completeness, we evaluate on these two auxiliary tasks.

Table 8 summarizes the results: the multitask model slightly outperforms a dedicated simplification model on English simplification, showing the benefits of the additional training data from other tasks. By contrast, on the resource-rich MT task, the standalone translation system performs better.

<sup>6</sup>A random sample of outputs from the best model configuration are provided in the Appendix (Table 14).

Newsela			OPUS	Evaluation Metrics		
S	MT+S	MT	MT	BLEU	SARI	PCC
✓	✓	✓	✓	22.75	45.0	0.608
	✓	✓	✓	22.62	44.5	0.584
✓		✓	✓	22.51	44.8	0.572
✓			✓	19.16	43.4	0.468
✓	✓	✓		14.65	41.4	0.521

Table 7: Data ablation experiments showing the impact of different types of training examples on multi-task model. The OPUS parallel corpus is essential for good performance. Simplification data (S) can be used for simplification supervision when joint translation and simplification examples (MT+S) are unavailable.

This can be explained by the fact that the standalone system is only responsible for text translation, while the multi-task model is exposed to samples of more diverse complexity levels during training which damage its ability to preserve complexity.

Task	BLEU	SARI	PCC
<i>English Simplification</i>			
Simplify	55.76	41.7	0.736
Translate and Simplify	<b>56.47</b>	41.3	0.730
All Tasks	56.05	<b>42.1</b>	0.736
<i>Machine Translation</i>			
Translate	<b>29.09</b>	-	<b>0.769</b>
Translate and Simplify	27.33	-	0.647
All Tasks	27.63	-	0.658

Table 8: Evaluation on auxiliary tasks: Multitask models trained on both the translation and simplification dataset improves the performance for the task of English Simplification.

#### 7.4 Provenance of Reading Grade Level

Our models control complexity using the gold reading grade level assigned by professional Newsela editors. We investigate the impact of replacing these gold labels by automatic predictions from the ARI metric. ARI can be computed for any English segment, including for MT parallel corpora that are not annotated for complexity.

Table 9 shows that ARI provides an adequate substitute for manually annotated reading grade levels, as BLEU and SARI score remain close when Newsela reading grade levels are replaced by ARI-based tags. However, annotating all data

Complexity cont. MT	BLEU	SARI	PCC
Newsela reading grade	22.51	44.80	0.572
ARI on Newsela data	22.26	<b>45.12</b>	<b>0.581</b>
ARI on all data	20.91	44.75	0.577

Table 9: Complexity controlled MT with automatic vs. manual reading grade level tags: ARI provides an adequate substitute for manually assigned grade levels.

with ARI grades, including the OPUS parallel corpus, hurts BLEU. We attribute this result to the differences in length and number of sentences per segment in OPUS vs. Newsela (Table 2): segments of vastly different lengths can have the same ARI score (Equation 4), thus confusing the multi-task model.

## 8 Conclusion

We introduce a new task that aims to control complexity in machine translation output, as a proxy for producing translations targeted at audiences with different reading proficiency levels. We construct a Spanish-English dataset drawn from the Newsela corpus for training and evaluation, and adopt a sequence-to-sequence model trained in a multitask fashion.

We show that the multitask model improves performance over translation and simplification pipelines, according to both machine translation and simplification metrics. The reading grade level of the multi-task outputs correlate better with target grade levels than with pipeline outputs. Analysis shows that these benefits come from their ability to combine larger training data from different tasks. Manual inspection also shows that the multi-task model successfully produces different translations for increasingly lower grades given the same Spanish input.

However, even when simplifying translations, multitask models are not yet able to exactly match the desired complexity level, and the gap between the complexity achieved and the target complexity increases with the amount of simplification required. Our datasets and models thus provide a foundation to investigate strategies for a tighter control on output complexity in future work.

## References

David Allen. 2009. A study of the role of relative clauses in the simplification of news texts for learn-

- ers of english. *System*, 37(4):585–599.
- Alessio Palmero Aprosio, Sara Tonelli, Marco Turchi, Matteo Negri, and Mattia A Di Gangi. 2019. Neural text simplification in low-resource conditions using weak supervision. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 37–44.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations (ICLR)*.
- 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.
- Chris Callison-Burch. 2007. *Paraphrasing and Translation*. Ph.D. thesis, University of Edinburgh.
- Yvonne Canning, John Tait, Jackie Archibald, and Ros Crawley. 2000. Cohesive generation of syntactically simplified newspaper text. In *International Workshop on Text, Speech and Dialogue*, pages 145–150. Springer.
- John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying text for language-impaired readers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*.
- R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. [Motivations and Methods for Text Simplification](#). In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2, COLING ’96*, pages 1041–1044, Stroudsburg, PA, USA.
- William Coster and David Kauchak. 2011. Simple English Wikipedia: A New Text Simplification Task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2, HLT ’11*, pages 665–669, Stroudsburg, PA, USA.
- Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Proceedings of the SIGIR workshop on accessible search systems*, pages 19–26. ACM Press New York.
- Noemie Elhadad and Komal Sutaria. 2007. Mining a lexicon of technical terms and lay equivalents. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 49–56. Association for Computational Linguistics.
- William A Gale and Kenneth W Church. 1993. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102.
- Laurie Gerber and Eduard Hovy. 1998. Improving Translation Quality by Manipulating Sentence Length. In *Conference of the Association for Machine Translation in the Americas (AMTA)*.
- Eva Hasler, Adrià de Gispert, Felix Stahlberg, Aurelien Waite, and Bill Byrne. 2017. [Source sentence simplification for statistical machine translation](#). *Computer Speech & Language*, 45(Supplement C):221–235.
- Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the third workshop on innovative use of NLP for building educational applications*, pages 71–79. Association for Computational Linguistics.
- Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *arXiv preprint arXiv:1712.05690*.
- Kentaro Inui, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, and Tomoya Iwakura. 2003. [Text simplification for reading assistance: A project note](#). In *Proceedings of the Second International Workshop on Paraphrasing*, pages 9–16, Sapporo, Japan.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. [Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation](#). *arXiv:1611.04558 [cs]*.
- Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.
- 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.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), Demonstration Session*, Prague, Czech Republic.
- Reno Kriz, João Sedoc, Marianna Apidianaki, Carolina Zheng, Gaurav Kumar, Eleni Miltsakaki, and Chris Callison-Burch. 2019. Complexity-weighted loss and diverse reranking for sentence simplification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

- Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3137–3147.
- Kelly Marchisio, Jialiang Guo, Cheng-I Lai, and Philipp Koehn. 2019. Controlling the reading level of machine translation output. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 193–203.
- Shachar Mirkin and Jean-Luc Meunier. 2015. Personalized Machine Translation: Predicting Translational Preferences. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2019–2025, Lisbon, Portugal.
- Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. Controllable text simplification with lexical constraint loss. In *Proceedings of the 57th Conference of the Association for Computational Linguistics: Student Research Workshop*, pages 260–266.
- Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 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, Vancouver, Canada.
- Xing Niu, Sudha Rao, and Marine Carpuat. 2018. [Multi-Task Neural Models for Translating Between Styles Within and Across Languages](#). In *27th International Conference on Computational Linguistics (COLING 2018)*.
- Gustavo Paetzold, Fernando Alva-Manchego, and Lucia Specia. 2017. Massalign: Alignment and annotation of comparable documents. *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 1–4.
- Gustavo Henrique Paetzold and Lucia Specia. 2016. Semeval 2016 task 11: Complex word identification. *Proceedings of SemEval*, pages 560–569.
- Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA.
- Sarah E Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *Workshop on Speech and Language Technology in Education*.
- Ella Rabinovich, Shachar Mirkin, Raj Nath Patel, Lucia Specia, and Shuly Wintner. 2016. [Personalized Machine Translation: Preserving Original Author Traits](#). *arXiv:1610.05461 [cs]*.
- Carolina Scarton and Lucia Specia. 2018. Learning Simplifications for Specific Target Audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 712–718.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Controlling Politeness in Neural Machine Translation via Side Constraints](#). pages 35–40.
- RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, CINCINNATI UNIV OH.
- Advait Siddharthan. 2014. [A survey of research on text simplification](#). *ITL - International Journal of Applied Linguistics*, 165(2):259–298.
- Advait Siddharthan and Napoleon Katsos. 2010. Reformulating discourse connectives for non-expert readers. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1002–1010. Association for Computational Linguistics.
- Sanja Štajner, Iacer Calixto, and Horacio Saggion. 2015. Automatic text simplification for spanish: Comparative evaluation of various simplification strategies. In *Proceedings of the international conference recent advances in natural language processing*, pages 618–626.
- Sanja Štajner, Marc Franco-Salvador, Paolo Rosso, and Simone Paolo Ponzetto. 2018. Cats: A tool for customized alignment of text simplification corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.
- Sanja Štajner and Maja Popovic. 2016. [Can text simplification help machine translation?](#) In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 230–242.
- Jörg Tiedemann. 2009. News from OPUS-A collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume 5, pages 237–248.
- Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.
- Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting Gender Right in Neural Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008.
- Tong Wang, Ping Chen, John Rochford, and Jipeng Qiang. 2016. Text simplification using neural machine translation. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Willian Massami Watanabe, Arnaldo Candido Junior, Vinícius Rodriguez Uzêda, Renata Pontin de Matos Fortes, Thiago Alexandre Salgueiro Pardo, and Sandra Maria Aluísio. 2009. Facilita: reading assistance for low-literacy readers. In *Proceedings of the 27th ACM international conference on Design of communication*, pages 29–36. ACM.

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.

Xingxing Zhang and Mirella Lapata. 2017. [Sentence Simplification with Deep Reinforcement Learning](#). *arXiv:1703.10931 [cs]*.

## A Supplemental Material

Table 10 and 11 provides the statistics of grade pair distribution in the Newsela English and Newsela Spanish-English dataset.

Src / Tgt	2	3	4	5	6	7	8	9	10
3	2652	0	0	0	0	0	0	0	0
4	4984	8212	0	0	0	0	0	0	0
5	2287	19589	23775	0	0	0	0	0	0
6	1914	7625	21022	21380	0	0	0	0	0
7	608	8897	14249	33466	10944	0	0	0	0
8	623	3710	13267	17347	22745	12006	0	0	0
9	130	5058	5031	19834	4684	30929	2144	0	0
10	6	40	224	320	382	289	400	142	0
11	0	0	15	19	11	16	28	0	0
12	1069	6818	18430	34232	28532	41561	29836	31327	97

Table 10: Number of text segments per grade level pair in our Newsela English Corpus

Src / Tgt	2	3	4	5	6	7	8	9	10
3	293	0	0	0	0	0	0	0	0
4	670	1305	0	0	0	0	0	0	0
5	251	3383	1957	0	0	0	0	0	0
6	223	1124	2090	2833	0	0	0	0	0
7	60	1249	926	4986	1244	0	0	0	0
8	96	548	1016	1804	3705	1221	0	0	0
9	16	717	211	3074	189	6135	263	0	0
10	0	3	5	15	26	1	46	0	0
12	189	1288	1902	5312	4708	7796	4995	7077	30

Table 11: Number of text segments per grade level pair in our Newsela English-Spanish Corpus

Model	Bleu	SARI	Flesch
<i>Results copied from <a href="#">Scarton and Specia (2018)</a></i>			
seq2seq w/ side-constraint	62.91	41.01	82.91
<i>Reimplementation evaluated on our Newsela download</i>			
seq2seq w/ side-constraint	58.61	39.81	70.44
seq2seq w/ side-constraint + BPE	61.87	52.78	66.98

Table 12: Comparison with previously published results on Newsela English text simplification. Our implementation yields BLEU and SARI score that are close to those reported in [Scarton and Specia \(2018\)](#). The difference in Flesch score can be attributed to changes in the number and complexity of articles available in newsela at the time the datasets were extracted.

12	Se estima que 75 personas han expresado interés en alojar al menos a un solicitante de asilo, dijo Cronk. Algunas de estas personas viven en las principales áreas metropolitanas de California y Nueva York. Otros son de zonas remotas y rurales de Montana y Dakota del Norte.
8	An estimated 75 people have expressed interest in hosting at least one asylum-seeker said Cronk. Some of these people live in major California and New York area. Others are from remote and rural areas of Montana and North Dakota.
6	An estimated 75 people have expressed interest in hosting at least one asylum-seeker said Cronk. Some of these people live in major California and New York area. Others are from remote and rural areas of Montana and North Dakota.
4	An estimated 75 people have expressed interest in hosting at least one asylum-seeker said Cronk. Some of these people live in the <b>main</b> areas of California and New York. Others are from remote and rural areas of Montana and North Dakota.
2	<b>Many people live in the</b> United States and New York City. Some are from remote areas of Montana and North Dakota.
12	El gobierno federal realizó un contrato con el centro de detención juvenil en Vicennes, Indiana, desde el año 2004 hasta el año 2010, para que alojara a aquellos niños inmigrantes considerados como los más peligrosos.
9	The federal government conducted a contract with the youth detention center at Vicennes, Indiana, since 2004 to 2010 to host those immigrant children considered as the most dangerous ones.
7	The federal government conducted a contract with the youth detention center at Vicennes, Indiana, since 2004 to 2010, <b>so that they hosted</b> those immigrant children considered as the most dangerous ones.
5	The federal government <b>made</b> a contract with the youth detention center in Vice-Year. <b>It is in the</b> United States since 2004 to 2010.
3	The government made a deal with the youth detention center in Vice-Year. It is in the United States since 2004 to 2010. <b>It was to host</b> those immigrant children as the most dangerous in 2010.
12	Poco después de que el dron despegó durante la prueba de Verizon en Cape May, una "aeronave de seguimiento" salió tras él, para garantizar que el dron pudiera evadir otros aviones en caso de volar dentro de un espacio aéreo designado.
9	Shortly after the drone took off from Verizon in Cape May, a "unmanned aircraft following" came out after him, to ensure that the drone could evade other planes in case of flying inside a designated airspace.
6	Shortly after the drone took off in Cape May, a "unmanned aircraft force" came out <b>behind</b> him. <b>They could</b> ensure that the drone could evade other planes in case of flying inside a designated airspace.
2	<b>Not long</b> after the drone took off from Verizon in Cape May, a "unmanned aircraft help" came out after him. <b>It is to make sure</b> the drone could evade other planes in case of flying inside a air space.

Table 13: Examples of simplification operations observed when simplifying from a higher Grade level into different lower grade levels using the multitask model (All tasks).

Src_G	Tgt_G	Source	Target	Predicted
8	5	Los partidarios de la reserva dicen que marca un precedente para muchos países que trabajan conjuntamente para proteger una gran porción del océano que no se halla dentro de la jurisdicción específica de ningún país.	Supporters of the reserve say it sets an example for multiple countries working together to protect a large area of ocean. The area is not controlled by any single nation.	Supporters of the reserve say that marks a precedent for many countries that work together to protect a large portion of the ocean.
5	3	Los padres no respondieron bien al escuchar que sus hijas estaban surfeando porque les preocupaba el peligro y la reputación de las niñas.	The parents did not react well to their daughters surfing. They worried about the danger and what other people would think.	Parents did not want to hear that their daughters were surfing their daughters because they were worried about the danger and their reputation.
12	5	Según dijo el actor Eds Eramiha, esta escena habría sido difícil de imaginar hace tan solo dos décadas, cuando el idioma te reo maorí era ampliamente considerado una lengua moribunda que no valía la pena enseñar.	Eds Eramiha is an actor in the play. He says it would have been difficult to imagine this scene just 20 years ago. Back then, te reo Maori was widely regarded as a dying language not worth teaching.	Actress Eds Eramiha said this scene would have been hard to imagine only two decades ago. The language was widely viewed as a dying tongue that would not be worth teaching.
5	4	Dos astrónomos han planteado una fascinante pregunta: ¿Pueden las lunas tener lunas? La interesante respuesta es que sí, una luna puede tener su propia luna.	Two astronomers asked a question: Can moons have moons? The interesting answer is yes, a moon can have its own moon.	Two astronomers have asked a fascinating question: Can Mondays have Monday? The interesting answer is that yes, a moon can have its own.
9	3	Como hijo de un sastre y modisto de pocos ingresos, consiguió trabajo en el taller de un sastre en Brooklyn a la edad de 15 años, terminó su secundaria estudiando por las noches y luego se dispuso a triunfar en la vida.	Campos is the son of a tailor and a poor dressmaker. He found work in a Brooklyn clothing shop at age 15 and He finished high school at night. Then he set out to succeed.	The son of a saga and a modest age of a few incomes, he got work at the age of 15 years old. He graduated from the evenings and then turned into life in life.

Table 14: Example translations produced by our best multitask model. Refer Table 7 (Row 1).