Revisiting non-English Text Simplification: A Unified Multilingual Benchmark

Michael J. Ryan, Tarek Naous, Wei Xu
School of Interactive Computing
Georgia Institute of Technology
{michaeljryan, tareknaous}@gatech.edu; wei.xu@cc.gatech.edu

Abstract

Recent advancements in high-quality, large-scale English resources have pushed the frontier of English Automatic Text Simplification (ATS) research. However, less work has been done on multilingual text simplification due to the lack of a diverse evaluation benchmark that covers complex-simple sentence pairs in many languages. This paper introduces the MULTISIM benchmark, a collection of 27 resources in 12 distinct languages containing over 1.7 million complex-simple sentence pairs. This benchmark will encourage research in developing more effective multilingual text simplification models and evaluation metrics. Our experiments using MULTISIM with pre-trained multilingual language models reveal exciting performance improvements from multilingual training in non-English settings. We observe strong performance from Russian in zero-shot cross-lingual transfer to low-resource languages. We further show that few-shot prompting with BLOOM-176b achieves comparable quality to reference simplifications outperforming fine-tuned models in most languages. We validate these findings through human evaluation.[1]

1 Introduction

Automatic text simplification (ATS) is the task of reducing the complexity of a text without changing its original content and meaning (Al-Thanyyan and Azmi, 2021). ATS has many applications, from making a text easier to read for people with reading and cognitive disabilities (Stajner, 2021) and second language learners (Petersen and Ostendorf, 2007) to reducing the complexity of medical texts for easier understanding by the general public (van den Bercken et al., 2019). For better accessibility to diverse communities, this technology should be available without language barriers.

Much of the recent success in English text simplification comes from large parallel corpora of texts with the same content written using both complicated and simple sentences (Xu et al., 2015; Jiang et al., 2020; Alva-Manchego et al., 2020). These resources enable the training of large language models for ATS in English (Scarton and Specia, 2018; Martin et al., 2020; Omelianchuk et al., 2021). ATS research in other languages has received much less attention (Martin et al., 2022). Figure 1 shows that the growth of English text simplification research outpaces progress in other languages.

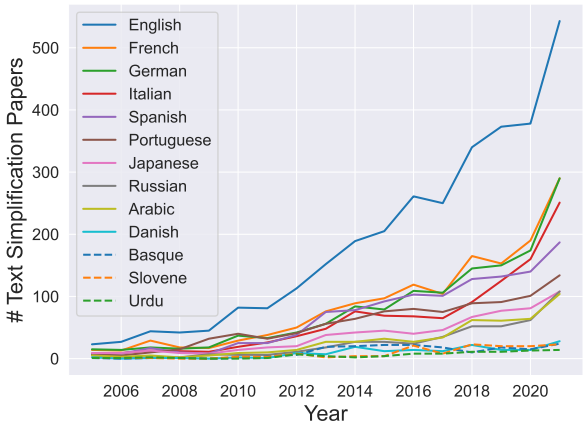

Figure 1: Papers published each year with content related to text simplification and a specific language according to Google Scholar. The quantity of English text simplification work vastly exceeds all other languages.

A diverse multilingual benchmark is essential for a more comprehensive evaluation of multilingual simplification methods, pre-trained models, and evaluation metrics. The lack of a multilingual benchmark that covers a set of high, medium, and low-resource languages belonging to different scripts and language families hinders advancement in multilingual ATS. In this paper, we address this gap in the field by introducing the MULTISIM benchmark that covers 27 text simplification datasets (complex-simple pairs) in 12 different languages. MULTISIM consists of a collection of datasets from the literature that we unify into a single format for easier accessibility to the research

[1]Code and Data available at https://github.com/XenonMolecule/MultiSim

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4898–4927

community. In summary, our main contributions are as follows:

1. We present a comprehensive literature survey of all existing multilingual text simplification corpora, created via several methodologies categorized into four main approaches (§3).

2. We release the MULTISIM benchmark for multilingual text simplification, containing 1,749,056 simple-complex sentence pairs in 12 different languages. To our knowledge, this is the first multilingual benchmark for text simplification. (§4).

3. We run various experiments using pre-trained multilingual language models and analyze their effectiveness in few-shot learning and cross-lingual transfer for challenging cases of low-resource languages or domain-specific simplification (§5). Our results highlight the benefits of domain and language script match for zero-shot transfer. We find that few-shot prompting large language models produces high-quality simplifications in both high and low-resource languages (§6). We validate these findings with human evaluation (§7).

2 Related Works

2.1 Multilingual Benchmarks

Recently researchers have released several multilingual benchmarks to assess that models work well not only in the high-resource settings where they are trained but in all languages. *XTREME-R* (Hu et al., 2020) is a multitask benchmark across 50 languages. The benchmark focuses on classification, question answering, structured prediction, and retrieval. Another text classification benchmark is *XGLUE* (Liang et al., 2020), which covers 11 diverse tasks in 19 languages. Finally, the new *XTREME-UP* benchmark (Ruder et al., 2023) evaluates 88 under-represented languages on 9 tasks from machine translation to OCR to autocomplete.

Single task multilingual benchmarks exist for NLI (Conneau et al., 2018), QA (Lewis et al., 2020; Longpre et al., 2021), causal reasoning (Ponti et al., 2020), semantic similarity (Vulić et al., 2020), style transfer (Briakou et al., 2021), fact checking (Gupta and Srikumar, 2021), fairness (Chalkidis et al., 2022), stance classification (Zheng et al., 2022), text summarization (Giannakopoulos et al., 2015; Ladhak et al., 2020; Scialom et al., 2020), readability (Naous et al., 2023) and more (Gretter, 2014; Meilicke et al., 2012; Li et al., 2020; Raganato et al., 2020). To date, no such benchmarks exist for multilingual text simplification.

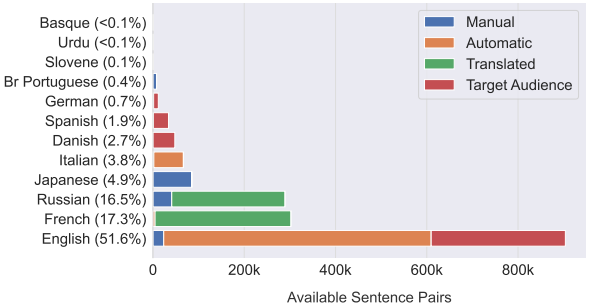

Figure 2: Data availability for text simplification in all languages partitioned on collection strategy. Despite only including three of the most common English datasets, English resources outnumber all other language resources combined.

2.2 Multilingual Text Simplification

The most common approach to automatic text simplification is training a statistical or neural sequence-to-sequence generation model on parallel simplification corpora (§3). Besides in English, researchers have done this in Brazilian Portuguese (Specia, 2010), German (Säuberli et al., 2020; Battisti et al., 2020), Spanish (Štajner et al., 2015; Štajner, 2014), French (Cardon and Grabar, 2020), Japanese (Goto et al., 2015; Maruyama and Yamamoto, 2017), Danish (Klerke and Søgaard, 2013), and Russian (Shatilov and Rey, 2021; Fenogenova, 2021).

Zero-shot and unsupervised learning are promising directions for multilingual text simplification. Mallinson et al. (2020) showed zero-shot simplification in German using a German decoder on a transformer architecture trained on English simplification. Additionally, Martin et al. (2022) showed that mining a massive amount of paraphrases and simplifications as training data was sufficient to achieve state-of-the-art performance in English, Spanish, and French text simplification.

3 Parallel Simplification Corpora

So far, 31 parallel simplification corpora exist in non-English languages. We organize the discussion of these corpora by their creation strategy. Figure 2 shows the amount of data in each language divided by collection strategy. Table 1 summarizes the details of each corpus stratified by language. We provide more detail on each corpus in Appendix A.

Corpus	Source(s)	Simplification Author	Collection Strategy	Alignment Level	Sentence Aligned	Complex Sentences	Simple Sentences	Access
Arabic Corpora								
Saaq al-Bambuu (Khallaf and Sharoff, 2022)	Literature	writer	Target Audience Resource	sentence	auto	2,980	2,980	private
Basque Corpora								
CBST (Gonzalez-Dios et al., 2018)	Science Communications	translator, teacher	Annotator	document	manual	458	591	on request
Brazilian Portuguese Corpora								
PorSimples (Aluísio and Gasperin, 2010)	News, Science Communications	linguist	Annotator	document	manual	7,902	10,174	on request
Danish Corpora								
DSim (Klerke and Søgaard, 2012)	News	journalists	Target Audience Resource	sentence	auto	47,887	60,528	on request
English Corpora†								
ASSET (Alva-Manchego et al., 2020)	Wikipedia	crowdsource	Annotator	sentence	manual	2,359	23,590	open source
Newsela EN (Xu et al., 2015)	News	experts	Target Audience Resource	document	auto	393,798	402,222	on request
Wiki-Auto (Jiang et al., 2020)	Wikipedia	crowdsource	Automatic	document	auto	10,144,476	1,241,671	open source
French Corpora								
Alector (Gala et al., 2020)	Websites, Science Communications	experts	Annotator	document	NA	1,230	1,192	open source
CLEAR (Grabar and Cardon, 2018)	Wikipedia, Medical Documents	crowdsource, experts	Automatic	sentence	auto	4,596	4,596	open source
WikiLarge FR (Cardon and Grabar, 2020)	Wikipedia	crowdsource	Translation	sentence	auto	307,067	308,409	open source
German Corpora								
GEOLinoTest (Mallinson et al., 2020)	News	linguist	Annotator	sentence	manual	1,198	1,198	open source
German News (Säuberli et al., 2020)	News	news agency	Target Audience Resource	document	auto	15,239	14,344	on request
Klexikon (Aumiller and Gertz, 2022)	Wikipedia	crowdsource	Automatic	document	NA	771,059	96,870	open source
Simple Patho (Trienes et al., 2023)	Medical Documents	medical students	Annotator	paragraph	manual	22,191	26,551	private
Simple German (Battisti et al., 2020)	Websites	government	Target Audience Resource	document	auto	12,806	8,400	on request*
TextComplexityDE (Naderi et al., 2019)	Wikipedia	native speaker	Annotator	document	manual	250	250	open source
Italian Corpora								
AdminIT (Miliani et al., 2022)	Government	researchers	Annotator	sentence	manual	777	763	open source
SIMPITIKI Wiki (Tonelli et al., 2016)	Wikipedia	crowdsource	Automatic	sentence	manual	575	575	open source
PaCCSS-IT (Brunato et al., 2016)	Websites	crowdsource	Automatic	sentence	auto	63,006	63,006	open source
Teacher (Brunato et al., 2015)	Websites	teachers	Annotator	document	manual	204	195	open source
Terence (Brunato et al., 2015)	Literature	experts	Annotator	document	manual	1,035	1,060	open source
Japanese Corpora								
EasyJapanese (Maruyama and Yamamoto, 2018)	News, Websites	students	Annotator	sentence	manual	50,000	50,000	open source
EasyJapaneseExtended (Katsuta and Yamamoto, 2018)	News, Websites	crowdsource	Annotator	sentence	manual	34,400	35,000	open source
Japanese News (Goto et al., 2015)	News	journalists, teachers	Target Audience Resource	document	auto	13,356	13,356	private
Russian Corpora								
RuAdapt Encyclopedia (Dmitrieva et al., 2021)	Encyclopedic	researchers	Annotator	document	auto	9,729	10,230	open source
RuAdapt Fairytale (Dmitrieva et al., 2021)	Literature	researchers	Annotator	document	auto	310	404	open source
RuAdapt Lit (Dmitrieva and Tiedemann, 2021)	Literature	writers	Annotator	document	auto	24,152	28,259	on request
RSSE (Sakhovskiy et al., 2021)	Wikipedia	crowdsource	Annotator	sentence	manual	2,000	6,804	open source
RuWikiLarge (Sakhovskiy et al., 2021)	Wikipedia	crowdsource	Translation	sentence	auto	278,499	289,788	on request
Slovene Corpora								
SloTS (Gorenc and Robnik-Šikonja, 2022)	Literature	experts	Target Audience Resource	sentence	manual	1,181	1,287	open source
Spanish Corpora								
FIRST (Orasan et al., 2013)	Literature, News, Medical Documents	experts	Annotator	document	manual	320	332	private
Newsela ES (Xu et al., 2015)	News	experts	Target Audience Resource	document	auto	46,256	45,519	on request
Simplext (Saggion et al., 2015)	News	researchers	Annotator	document	manual	1,108	1,742	on request
Urdu Corpora								
SimplifyUREval (Qasmi et al., 2020)	Literature, News	expert	Annotator	sentence	manual	500	736	open source

Table 1: Important properties of text simplification parallel corpora. †Common English corpora included for comparison. Many other English corpora omitted. *Only scripts to replicate the corpus are available upon request. Simple German results differ from original paper because of changes to availability of online articles. *Sources*: Literature, Science Communications, News, Wikipedia, Websites, Medical Documents, Government, Encyclopedic. *Collection Strategies*: Automatic, Translation, Annotator, Target Audience Resource.

3.1 Manual Simplification

Manual simplification is the most widely-used method for crafting a monolingual parallel simplification corpus. For manual simplification, annotators ranging from experts to crowdsourced workers to the researchers themselves use simplification guidelines (Aluísio et al., 2008b; Siddharthan, 2004; Gonzalez-Dios et al., 2018) to simplify complex documents manually. Researchers have used this methodology to create 19 resources in 10 different languages. Manual simplification gives researchers control over the types of simplification operations in the dataset. However, guiding annotators using rules can result in unnatural simplifications. Stajner (2014) found that relaxing guidelines led to more effective simplifications.

3.2 Automatic Collection

Manual simplification at scale is a costly and time-consuming process. An alternative approach is to use automatic collection methods, which leverage various online knowledge bases or web sources to increase the size of these resources. However, this can come at the cost of sacrificing quality, as automatically paired sentences may not always be an exact complex-simple match. Additionally, it can be challenging to control the level of simplification.

One common source for automatic collection is Wikipedia, which is available in many languages and often has both original and simplified versions with the same content.[2] This has been used to build parallel simplification corpora (Jiang et al., 2020; Tonelli et al., 2016; Cardon and Grabar, 2019), where researchers match articles on the same topic from the regular and simple Wikipedia versions. They then leverage automatic aligners such as a neural CRF aligner (Jiang et al., 2020) or CATS (Štajner et al., 2018) to find matching sentences between the two articles.

Other sources of automatically simplified sentences include web scrapes like Common Crawl (Wenzek et al., 2020). For web scrapes, sentence embeddings (Heffernan et al., 2022) are used to find similar sentences to pair up. Then additional filtering is applied to ensure the sentences are not exact matches. A readability measure can be used to ensure that one sentence is simpler than the other. This was the strategy used to create PaCCSS-IT (Brunato et al., 2016) in Italian and MUSS (Martin et al., 2022) in English, Spanish, and French.

3.3 Machine Translation

Some datasets are machine translations of existing large resources. For example, the French WikiLargeFR (Cardon and Grabar, 2020) and Russian RuWikiLarge (Sakhovskiy et al., 2021) are both machine translations of the English WikiLarge corpus (Zhang and Lapata, 2017). While this allows for large resources in multiple languages, it has two significant drawbacks. Firstly, the final dataset lacks the cultural identity of naturally occurring data in the target language. Secondly, machine translation errors can be introduced in the process, potentially impacting the dataset's quality.

3.4 Target Audience Resources

The final monolingual parallel corpora category is resources created for specific target audiences, such as individuals with lower literacy levels or second language learners. These resources typically have the highest quality, but they can be expensive to produce and therefore are relatively rare. There are currently target audience resources available in seven languages.

One company that specializes in creating high-quality target audience resources in English and Spanish is Newsela[3] (Xu et al., 2015). Founded in 2013, Newsela is a Series D startup that has attracted over $100 million in funding to support its goal of promoting meaningful classroom learning at all levels. The company employs a team of content producers with extensive teaching experience in K-12 education to train and manage a network of freelance writers who create the simplifications.

Several national news agencies have established systems for creating simplified versions of their articles. For instance, News Web Easy[4] in Japan is a division of the Japan Broadcasting Corporation, a public media organization. News Web Easy targets Japanese second language learners and primary and secondary school students (Goto et al., 2015). The German government funds the Austrian Press Agency to produce TopEasy[5] (Säuberli et al., 2020), a simplified version of their news published each weekday. Similarly, the publicly funded Danish Broadcasting Corporation (DR) offers simplified versions of their stories called DR Ligetil[6] (Straightforward) (Klerke and Søgaard, 2012).

4 The MultiSim Benchmark

We release the MULTISIM Benchmark, a collection of 27 parallel simplification corpora in 12 languages and 4 Scripts. 18 of these corpora are open-sourced and are available online. For 9 corpora, permission must be obtained from the original authors. We provide data loaders for these resources.

4.1 Languages

We included all languages with open-source parallel sentence-aligned text simplification corpora. This covers eight languages: English (en), French (fr), German (de), Italian (it), Japanese (ja), Russian (ru), Slovene (sl), and Urdu (ur). Six languages have corpora available on request. We provide data loaders and splits to make these resources compatible with the MULTISIM benchmark. Such resources exist in Basque (eu), Brazilian Portuguese (pt-br), Danish (da), German (de), Russian (ru), and Spanish (es). Some resources are entirely private due to copyright protection or data-sharing permissions. These resources are in Arabic (ar), German (de), Japanese (ja), and Spanish (es).

[2] https://simple.wikipedia.org/wiki/Main_Page
[3] https://newsela.com
[4] https://www3.nhk.or.jp/news/easy/
[5] https://science.apa.at/nachrichten-leicht-verstandlich/
[6] https://www.dr.dk/ligetil

Language	Dataset	#train	#test	#dev
Open Source				
English	WikiAuto	576,126	5,012	5,012
	ASSET*	20,000	3,590	0
French	WikiLargeFR*	296,402	359	992
	CLEAR*	4,196	100	300
German	GEOLino	958	122	118
	TextCompDE	200	25	25
Italian	PaCCSS-IT	60,485	1,267	1,254
	Terence	809	101	102
	AdminIT	588	73	75
	Simpitiki	460	56	59
	Teacher	136	17	17
Japanese	Easy JA	48,000	1,000	1,000
	Easy JA Ext*	34,269	731	0
Russian	RuAdapt Ency	7,782	982	965
	RSSE Corpus*	3,406	3,398	0
	RuAdapt Fairy	248	31	31
Slovene	SloTS*	749	96	94
Urdu	SimplifyUR	594	68	74
	Total	1,055,408	17,028	10,118
Dataloaders Available (Data on Request)				
Basque	CBST	361	46	46
Br Portuguese	PorSimples	6,290	790	784
Danish	DSim Corpus	45,885	997	1,005
English	Newsela EN	291,969	991	1,008
German	German News	8,186	1,024	1,023
Russian	RuWikiLarge*	246,978	365	768
	RuAdapt Lit	22,152	1,000	1,000
Spanish	Newsela ES	30,910	1,001	1,001
	Simplext	737	92	93
	Total	653,468	6,306	6,728

Table 2: MULTISIM splits. *Original splits preserved

4.2 Domains

The MULTISIM benchmark spans 8 domains. *Literature* sources are simplified versions of novels. *Science Communications* are popular science articles already written for public consumption and then rewritten at a simpler level. *News* sources are simplified versions of articles and news stories. *Wikipedia* sources are pulled from original and simple Wikipedia sites. *Website* sources generally come from web scrapes like Common Crawl (Wenzek et al., 2020) or specific target websites with original and simplified texts. *Medical* documents are drug leaflets, clinical notes, and similar texts written for doctors but simplified so that an average patient could understand. *Government* documents are taken from government policies and simplified to use more common vernacular. *Encyclopedic* documents are informational texts like Wikipedia but from other encyclopedic sources.

4.3 Pre-processing and Splitting

For any resource that provided a train, test, dev split, we include the original split of the data in our collection. Otherwise, we randomly divided all sentence pairs into train, test, and dev sets. For resources under 10,000 sentence pairs, we used 80%/10%/10% splits. For resources above 10,000 sentence pairs, we randomly sampled about 1,000 sentences each for the test/dev sets. For resources above 500,000 sentence pairs (WikiAuto), we randomly sampled about 5,000 sentences each for the test/dev sets. We report split sizes in Table 2.

Since several resources in the benchmark come from overlapping domains (i.e., Wikipedia, Web, News), repeat sentences exist between the original datasets. To fix this, we identified overlapping sentences and ensured they fell in the same split by swapping with randomly sampled sentence pairs. We repeated this process until all splits were completely independent.

5 Experiments

5.1 Evaluation Setup

For automatic evaluation, we use SARI (Xu et al., 2016), the average of the F1 score for adding, keeping, and deleting n-grams ($n \in \{1, 2, 3, 4\}$). SARI has been shown to correlate with human judgments of simplicity (Xu et al., 2016). We also report BLEU (Papineni et al., 2002), a common metric in machine translation. Although BLEU scores do not measure simplicity (Sulem et al., 2018), we use them as a check for grammatically and meaning preservation (Xu et al., 2016). We compute all evaluation metrics using `EASSE` evaluation suite (Alva-Manchego et al., 2019).

5.2 Baselines

To put the results of our experiments in perspective, we compare them with two common baselines.

Identity The original sentence is copied and reported as the simplification. This baseline earns high BLEU scores from the high token overlap between original and simple sentences.

Truncation The last 20% of words are cut from the original sentence. This baseline achieves high SARI scores because it balances keeping/deleting tokens, two operations SARI measures.

5.3 Models

For fine-tuning we used mT5 Base (Xue et al., 2021) (580M Parameters). We calculated the S-BLEU score between the original and simple sentences in

			BLEU					SARI				
			Baseline		Finetune			Baseline		Finetune		
Lang	Dataset	Size	Identity	Trunc	Single	Lang	All	Identity	Trunc	Single	Lang	All
eu	CBST	218	72.02	57.87	—	—	66.75	23.46	32.58	—	—	**32.83**
ur	SimplifyUR	470	58.85	41.11	—	—	56.23	24.84	31.30	—	—	**51.74**
sl	SloTS	188	7.76	6.09	—	—	7.63	5.93	19.03	—	—	**30.52**
pt-br	PorSimples	1,949	73.67	51.93	—	—	63.85	28.21	31.25	—	—	**44.27**
de	TextCompDE	144	26.77	19.98	—	—	24.53	15.42	26.81	—	—	**41.15**
	GEOLino	437	69.86	50.03	—	—	71.90	27.45	30.70	—	—	**50.75**
	GermanNews	1,748	7.29	7.13	—	—	6.57	5.61	17.69	—	—	**31.58**
es	Simplext	157	13.91	13.15	—	14.42	12.25	7.94	20.27	—	19.91	**32.68**
	NewselaES	17,022	58.18	43.06	51.78	53.12	48.94	24.21	31.64	29.89	28.56	**35.36**
da	DSim	25,524	31.39	28.85	33.66	33.66	27.25	16.25	26.10	31.40	31.40	**38.44**
it	Simpitiki	24	95.23	74.48	—	24.40	36.28	32.45	32.00	—	20.10	**24.27**
	Teacher	83	34.49	29.05	—	32.21	29.76	17.41	27.75	—	29.98	**30.97**
	AdminIT	114	52.50	45.63	—	40.09	43.80	20.89	28.22	—	34.72	**36.21**
	Terence	394	67.24	49.72	—	59.33	50.65	26.83	32.82	—	**37.77**	36.92
	PaCCSS-IT	55,274	36.76	28.77	49.57	48.31	42.87	18.14	28.26	**57.30**	55.98	54.43
ja	EasyJA	27,600	58.09	8.43	65.83	68.12	66.04	24.64	24.28	67.36	**70.95**	70.11
	EasyJAExt	32,248	20.23	0.00	33.07	35.67	31.50	9.00	35.32	43.15	50.26	**53.49**
ru	RuAdaptFairy	97	12.56	8.03	—	13.11	11.01	10.63	24.84	—	23.77	**26.55**
	RuAdapt Ency	1,450	84.15	59.66	—	76.06	61.83	29.90	31.09	—	**34.73**	34.40
	RSSE	1,477	38.23	34.69	—	36.94	31.78	10.91	22.72	—	29.49	**35.08**
	RuAdapt Lit	10,515	51.22	41.64	49.94	53.74	48.54	22.66	31.94	41.75	**42.03**	42.01
	RuWikiLarge	135,191	57.82	44.38	55.03	51.97	40.82	24.24	31.87	32.01	34.95	**37.59**
fr	CLEAR	3,179	55.00	45.10	25.45	53.72	48.57	23.73	32.17	34.86	30.85	**35.37**
	WikiLargeFR	148,276	58.51	46.67	52.43	51.16	43.57	24.44	32.23	35.20	38.22	**39.23**
en	ASSET	14,814	92.81	88.11	88.26	81.20	85.90	20.73	29.66	35.98	**42.77**	41.56
	NewselaEN	129,387	68.71	52.30	62.78	51.51	55.68	26.17	32.90	38.60	**40.18**	38.80
	WikiAuto	315,018	45.40	41.31	37.95	35.30	36.91	20.93	31.45	42.46	**42.48**	42.00

Table 3: BLEU and SARI scores of mT5 fine-tuning experiments. Size refers to the total train sentence pairs after BLEU filtering. Best fine-tuned SARI score in bold. Results on training sets smaller than 3,000 pairs were omitted since this was not enough data to unlearn the pretraining objective.

the training set and filtered out all sentences outside of the range [10,70] as done by Maddela et al. (2021) to remove identical pairs (high BLEU) and misalignments (low BLEU). We also added four control tokens to the input sentence with information about the character-length compression, Levenshtein similarity, word rank, and dependency tree depth of the output following Martin et al. (2020). We used a grid search of control tokens on the dev set to find the combination that yielded the highest SARI for evaluation. We used BLOOM (Scao et al., 2022) (176B Parameters) for a few-shot. We report hyperparameters, prompts, and details for both models in Appendix C.

6 Results

6.1 Fine-tuning Language Models

We evaluated the mT5 models on all 27 datasets and fine-tuned them in 3 different settings. `Single`: On the training set of the dataset we are testing. `Language`: On the joint training set of all data in the same language. `All`: On the joint training set of all data across all languages. We remove results for training sets with fewer than 3,000 sentence pairs after S-BLEU filtering as we found this was not enough training data to unlearn the pre-training objective. We report the results of these experiments in Table 3.

Joint training improves performance in non-English languages. Joint all training improves SARI scores across every language besides English. English already has a wealth of in-language data so it performs best with joint-language training. There are specific datasets where joint-all does not achieve the highest SARI in other languages. Typically these are within one SARI point. Notably, PaCCSS-IT decreases in performance with more data. This may be due to the automatic collection approach to PaCCSS-IT which is prone to collect slightly noisy data. The similar BLEU scores to the identity baseline for all results suggests consistently high fluency.

6.2 Zero-shot Cross-lingual Transfer

We assess zero-shot cross-lingual and cross-domain transfer by training on one dataset and evaluating on another. We experiment with transfer to a small, domain-specific Italian dataset: Terence,

Scr	Fam	Lang	Dom	Dataset	BLEU	SARI
Transfer to Italian: Terence						
				Finetuned	7.96	23.92
				RuAdaptLit (ru)	57.07	35.19
			W	RuWikiLarge (ru)	10.76	25.85
✓			W	WikiAuto (en)	21.90	30.89
✓	✓		W	WikiLargeFR (fr)	8.66	25.20
✓	✓	✓	W	PaCCSS-IT (it)	40.59	36.99
Transfer to Basque: CBST						
Scr	Fam	Lang	Dom	Dataset	BLEU	SARI
				Finetuned	2.31	24.26
				EasyJA (ja)	1.87	26.67
				RuAdaptLit (ru)	47.31	37.89
✓				NewselaEN (en)	24.37	31.09
✓				PorSimples (pt-br)	7.97	29.44
Transfer to Urdu: SimplifyUR						
Scr	Fam	Lang	Dom	Dataset	BLEU	SARI
				Finetuned	5.20	33.18
			W	WikiAuto (en)	8.34	26.09
				EasyJA (ja)	0.00	17.87
				NewselaEN (en)	0.25	18.03
				RuAdaptLit (ru)	34.78	32.61
				PorSimples (pt-br)	5.24	32.87

Table 4: Transfer experiments to a domain specific small dataset (Terence) and two low resource language datasets (CBST, SimplifyUR). We find that matching Script, Family, Language, and Domain help improve transfer performance.

and two low-resource language datasets: CBST in Basque and SimplifyUR in Urdu. The transfer experiment results are shown in Table 4.

Matching script and language improve transfer performance. In transfer experiments to Italian and Basque, we see a notable improvement in BLEU and SARI scores with datasets in matching scripts (Latin, in this case). The best transfer results in Italian come from another dataset in the same language, demonstrating that in-language transfer learning trumps cross-lingual transfer if the data is available. Transferring across scripts typically corresponds to lower performance except when domains match.

Domain match can help regardless of script. In Urdu, we find that the best cross-lingual transfer results come from datasets in the same domain. This is true even though none of the transfer resources are in the Arabic script. In the transfer to Terence (Italian literature corpus), the Russian dataset with the matching domain, RuAdaptLit, outperforms RuWikiLarge, another Russian dataset from a different domain. Still, the domain alone does not guarantee strong transfer performance. EasyJA and NewselaEN performed poorly in Urdu transfer despite matching in domain.

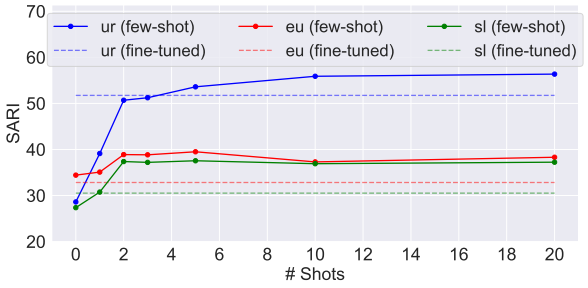

Figure 3: Semantic similarity fewshot performance in low-resource languages. Fewshot prompting achieves higher SARI than mt5 finetuned.

Russian is a good candidate language for cross-lingual transfer. For every test setting, the Russian corpus, RuAdaptLit, transfers well to the target dataset. Additionally, RuWikiLarge transfers better to Terence than the comparable WikiLargeFR, even though both datasets are machine translations of the same English WikiLarge corpus. This suggests that Russian is a good candidate language for cross-lingual transfer, which is in line with the findings of Turc et al. (2021) that Russian is a better choice than English as a pivot language for zero-shot cross-lingual transfer.

6.3 Prompting Multilingual Language Models

We assess few-shot performance in two settings. `Semantic Similarity`: We computed LASER sentence embeddings (Schwenk and Douze, 2017) for all sentences in the train, test, and validation set. During evaluation, we used the k nearest neighbors in the train set by cosine distance as examples. `Random Sampling`: We choose k random sentences from the train set as examples during evaluation. We highlight interesting findings of our few-shot experiments here. Fewshot results for all datasets are available in Table 8.

Few-shot prompting is promising for low-resource languages. Figure 3 shows semantic similarity sampling few-shot performance for the three low-resource languages in our study: Urdu, Basque, and Slovene. In all low-resource languages tested, few-shot outperforms fine-tuned mT5 trained on all data (+4.62, +5.48, +6.72 SARI, respectively). Within five examples few-shot exceeds fine-tuned performance. With limited resources, few-shot prompting is a good alternative to fine-tuning.

	English (ASSET)			Russian (RuAdaptLit)			Italian (Terence)			Urdu (SimplifyUR)		
	Adequacy	Fluency	Simplicity	Adequacy	Fluency	Simplicity	Adequacy	Fluency	Simplicity	Adequacy	Fluency	Simplicity
Reference	4.60±0.22	4.85±0.11	4.13±0.26	3.70±0.47	4.45±0.32	2.50±0.24	4.73±0.55	4.88±0.33	2.98±1.37	4.83±0.38	5.00±0.00	4.25±1.17
mT5 Single	4.45±0.25	4.95±0.07	3.00±0.34	4.50±0.31	4.78±0.19	2.25±0.15	1.23±0.73	1.15±0.66	1.15±0.66	2.38±1.37	2.10±1.32	1.10±0.30
mT5 Joint Language	4.65±0.19	4.98±0.05	3.38±0.25	4.78±0.26	5.00±0.00	2.48±0.24	4.78±0.48	4.83±0.55	2.55±1.01	—	—	—
mT5 Joint All	4.64±0.18	4.93±0.08	2.94±0.21	4.23±0.38	4.85±0.21	2.75±0.32	4.18±1.20	4.65±0.70	2.50±1.09	4.25±1.13	4.88±0.65	2.95±1.30
mT5 English Transfer	—	—	—	2.18±0.49	1.70±0.40	1.18±0.12	1.73±1.45	1.88±1.54	1.25±0.59	1.83±1.08	1.40±0.63	1.00±0.00
mT5 Russian Transfer	4.25±0.33	3.93±0.16	2.63±0.30	—	—	—	4.53±1.01	4.70±0.72	2.35±0.92	3.70±1.73	3.60±1.85	1.60±0.50
BLOOM 5 Shot (Rand)	4.63±0.25	4.75±0.18	3.10±0.40	4.65±0.27	4.78±0.22	2.33±0.20	4.33±1.02	4.53±0.85	2.45±1.11	4.95±0.22	4.93±0.35	3.28±1.69
BLOOM 5 Shot (Sim)	4.63±0.22	4.80±0.13	2.88±0.32	4.63±0.28	4.95±0.07	2.43±0.24	4.00±1.47	4.58±1.01	2.38±1.19	4.83±0.68	4.85±0.70	3.28±1.71

Table 5: Human evaluation adequacy, fluency, and simplicity scores for mT5 fine-tuned, mt5 zero-shot cross-lingual transfer, and BLOOM few-shot in English, Russian, Italian, and Urdu. Scores are averaged over 20 ratings per system with 95% confidence intervals.

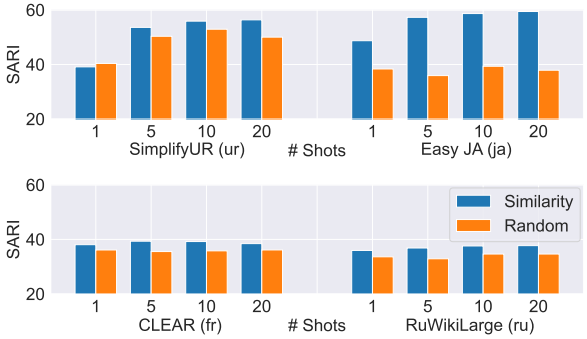

Figure 4: Semantic similarity vs random sampling few-shot performance on four diverse datasets. Semantic similarity consistently scores above random sampling.

Semantic similarity outperforms random sampling. Figure 4 shows semantic similarity vs. random sampling for few-shot evaluation on four diverse datasets: SimplifyUR (low resource language), EasyJapanese (manually simplified), CLEAR (medical domain), and RuWikiLarge (machine translated). In all cases prompting with semantic search outperformed random examples. This trend persists across languages and domains. Typically more examples improved performance, but any past five had marginal benefits.

7 Human Evaluation

For manual evaluation, we enlisted eight volunteers to annotate system outputs in English, Russian, Italian, and Urdu (two in each language) for three properties: Adequacy (is the meaning preserved?), Fluency (is the simplification eloquent/grammatical?), and Simplicity (is the output simpler?). This follows the standard annotation methodology in text simplification research (Martin et al., 2022; Xu et al., 2016). We asked annotators to rate 20 sentences for each model using a Likert scale from 1-5. We report the manual evaluation key in Table 9. Since our ratings are ordinal data, we measure annotator agreement using Krippendorff's alpha (Krippendorff, 2011) through the Fast Krippendorff Library (Castro, 2017). We achieve a high-reliability coefficient in all languages suggesting good annotator agreement. Specifically we calculate $\alpha = 0.80$ in English, $\alpha = 0.79$ in Russian, $\alpha = 0.75$ in Italian, and $\alpha = 0.86$ in Urdu.

Table 5 shows the manual evaluation results. Single dataset fine-tuned performance was deficient in Italian and Urdu because these datasets were small resources (1,012 pairs and 736 pairs, respectively). The Joint-all model performed consistently well across all datasets but did not outperform English language training for English. This aligned with our earlier findings (§6.1) and suggests that very large-scale in-language data is better than multilingual training if such data is available. Russian transfer using RuAdaptLit outperformed English transfer to both Italian and Urdu, reinforcing our observation that Russian is a strong choice for cross-lingual transfer (§6.2). We observe the best model scores from five-shot BLOOM to be on par with the reference simplifications in Italian/Russian and scoring slightly below reference simplifications in English/Urdu. This finding suggests that Few-shot prompting is effective for text simplification in both high and low-resource languages. In the low-resource Urdu setting, few-shot prompting yielded the best results, further substantiating our observations from few-shot prompting experiments (§6.3).

8 Conclusion

We release MULTISIM, the first multilingual text simplification benchmark, a collection of 27 sentence-aligned parallel corpora in 12 diverse languages. We collected these resources by surveying the literature for all existing text simplification resources in non-English languages, which were created via distinct methodologies that we categorize

into four main approaches. Using MULTISIM, we perform fine-tuning, few-shot, and zero-shot cross-lingual transfer experiments with generative multilingual language models (mT5, BLOOM), which revealed new insights in multilingual text simplification. Our results demonstrate the value of domain and script match for zero-shot cross-lingual transfer. We show that Russian is a good candidate pivot language, outperforming transfer from English in two of our case studies on low-resource and out-of-script languages. Further, we show that few-shot prompting BLOOM with examples obtained via semantic similarity outperforms fine-tuned models for low-resource languages. By releasing this benchmark, we hope to encourage and enable the development and evaluation of multilingual models and evaluation metrics for text simplification.

Limitations

This benchmark compiled and analyzed existing resources collected from diverse methods and domains. Although we demonstrated how careful use of these resources could transfer well to other resources, along with a manual analysis of a varied set of corpora, we cannot guarantee the quality of each resource or validate the methods that the original authors used to create them. We explore each dataset's linguistic properties in Appendix B. However, we encourage a deeper exploration of the quality of individual resources by researchers that speak the 12 languages included in this benchmark and corresponding data loaders.

Additionally, the human evaluation performed in this study was limited in scope and served primarily to validate the findings by automatic metrics. A more extensive evaluation with more annotators evaluating more sentences would be beneficial in order to draw further conclusions.

Furthermore, some of the resources discussed in this paper were automatically aligned. Although Neural CRF models in English have been shown to yield high-quality alignments (Jiang et al., 2020), other alignment algorithms such as TF-IDF scoring (Nelken and Shieber, 2006) have been shown to result in a high number of false positives (Xu et al., 2015). Future work could include realigning automatically aligned corpora using an embedding-based sentence alignment model trained on manually annotated alignment data (Jiang et al., 2020). We will continue updating this benchmark as updates are made to the underlying datasets, and new multilingual resources are released.

Acknowledgments

We thank Yang Chen and Yao Dou as well as three anonymous reviewers for their helpful feedback on this work. We also thank Govind Ramesh, Nour Allah El Senary, Luca Castagna, Lory O'Brien, Franco Paglione, Livia Paglione, Anton Lavrouk, Leah Levin, Irina Levin, Muhammad Hassan Maqsood, and Talha Ahmad Khan for their help with human evaluation. Furthermore, we thank Mounica Madella for sharing her control token scripts. This research is supported in part by the NSF awards IIS-2144493 and IIS-2112633, ODNI and IARPA via the BETTER program (contract 2019-19051600004) and the HIATUS program (contract 2022-22072200004). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF, ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Saud Al-Sanousi. 2016. *Saud al-Sanousi's Saaq al-Bambuu: The Authorized Abridged Edition for Students of Arabic*. Georgetown University Press.

Suha S. Al-Thanyyan and Aqil M. Azmi. 2021. Automated text simplification: A survey. *ACM Computing Surveys*, 54(2).

Sandra Aluísio and Caroline Gasperin. 2010. Fostering digital inclusion and accessibility: The PorSimples project for simplification of Portuguese texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 46–53, Los Angeles, California. Association for Computational Linguistics.

Sandra M. Aluísio, Lucia Specia, Thiago A. S. Pardo, Erick G. Maziero, Helena M. Caseli, and Renata P. M. Fortes. 2008a. A corpus analysis of simple account texts and the proposal of simplification strategies: First steps towards text simplification systems. In *Proceedings of the 26th Annual ACM International Conference on Design of Communication*, SIGDOC '08, page 15–22, New York, NY, USA. Association for Computing Machinery.

Sandra M. Aluísio, Lucia Specia, Thiago A.S. Pardo, Erick G. Maziero, and Renata P.M. Fortes. 2008b. Towards brazilian portuguese automatic text simplification systems. In *Proceedings of the Eighth ACM*

Symposium on Document Engineering, DocEng '08, page 240–248, New York, NY, USA. Association for Computing Machinery.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics.

Alberto Anula. 2011. Pautas básicas de simplificación textual y diseño del corpus simplext. Technical report, Technical report, Grupo DILES. Madrid, Spain: Universidad Autónoma de Madrid.

Dennis Aumiller and Michael Gertz. 2022. Klexikon: A German dataset for joint summarization and simplification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2693–2701, Marseille, France. European Language Resources Association.

Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 25–32.

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

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021. Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216, Online. Association for Computational Linguistics.

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, Austin, Texas. Association for Computational Linguistics.

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

Rémi Cardon and Natalia Grabar. 2019. Parallel sentence retrieval from comparable corpora for biomedical text simplification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 168–177, Varna, Bulgaria. INCOMA Ltd.

Rémi Cardon and Natalia Grabar. 2020. French biomedical text simplification: When small and precise helps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 710–716, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Helena M Caseli, Tiago F Pereira, Lucia Specia, Thiago AS Pardo, Caroline Gasperin, and Sandra Maria Aluísio. 2009. Building a brazilian portuguese parallel corpus of original and simplified texts. *Advances in Computational Linguistics, Research in Computer Science*, 41:59–70.

Santiago Castro. 2017. Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. `https://github.com/pln-fing-udelar/fast-krippendorff`.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Schwemer, and Anders Søgaard. 2022. FairLex: A multilingual benchmark for evaluating fairness in legal text processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4389–4406, Dublin, Ireland. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jon Dehdari. 2014. *A Neurophysiologically-Inspired Statistical Language Model*. Ph.D. thesis, The Ohio State University.

Yasuharu Den, Junpei Nakamura, Toshinobu Ogiso, and Hideki Ogura. 2008. A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. In *Proceedings of the Sixth International*

Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA).

Anna Dmitrieva, Antonina Laposhina, and Maria Lebedeva. 2021. A quantitative study of simplification strategies in adapted texts for l2 learners of russian. In *Proceedings of the International Conference "Dialogue*, pages 191–203.

Anna Dmitrieva and Jörg Tiedemann. 2021. Creating an aligned Russian text simplification dataset from language learner data. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 73–79, Kiyv, Ukraine. Association for Computational Linguistics.

Alena Fenogenova. 2021. Text simplification with autoregressive models. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue"*.

Núria Gala, Anaïs Tack, Ludivine Javourey-Drevet, Thomas François, and Johannes C. Ziegler. 2020. Alector: A parallel corpus of simplified French texts with alignments of misreadings by poor and dyslexic readers. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1353–1361, Marseille, France. European Language Resources Association.

George Giannakopoulos, Jeff Kubina, John Conroy, Josef Steinberger, Benoit Favre, Mijail Kabadjov, Udo Kruschwitz, and Massimo Poesio. 2015. MultiLing 2015: Multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 270–274, Prague, Czech Republic. Association for Computational Linguistics.

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

Sabina Gorenc and Marko Robnik-Šikonja. 2022. Slovene text simplification dataset SloTS. Slovenian language resource repository CLARIN.SI.

Isao Goto, Hideki Tanaka, and Tadashi Kumano. 2015. Japanese news simplification: tak design, data set construction, and analysis of simplified text. In *Proceedings of Machine Translation Summit XV: Papers*, Miami, USA.

Natalia Grabar and Rémi Cardon. 2018. CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Tilburg, the Netherlands. Association for Computational Linguistics.

Roberto Gretter. 2014. Euronews: a multilingual speech corpus for ASR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2635–2638, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682, Online. Association for Computational Linguistics.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. *arXiv preprint arXiv:2205.12654*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online. Association for Computational Linguistics.

Akihiro Katsuta and Kazuhide Yamamoto. 2018. Crowdsourced corpus of sentence simplification with core vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Nouran Khallaf and Serge Sharoff. 2022. Towards arabic sentence simplification via classification and generative approaches. *arXiv preprint arXiv:2204.09292*.

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, Sofia, Bulgaria. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Sigrid Klerke and Anders Søgaard. 2012. DSim, a Danish parallel corpus for text simplification. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 4015–4018, Istanbul, Turkey. European Language Resources Association (ELRA).

Sigrid Klerke and Anders Søgaard. 2013. Simple, readable sub-sentences. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 142–149, Sofia, Bulgaria. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Vladimir I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2020. Mtop: A comprehensive multilingual task-oriented semantic parsing benchmark. *arXiv preprint arXiv:2008.09335*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 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. Association for Computational Linguistics.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2020. Zero-shot crosslingual sentence simplification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5109–5126, Online. Association for Computational Linguistics.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.

Takumi Maruyama and Kazuhide Yamamoto. 2017. Sentence simplification with core vocabulary. In *2017 International Conference on Asian Language Processing (IALP)*, pages 363–366.

Takumi Maruyama and Kazuhide Yamamoto. 2018. Simplified corpus with core vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Paul McCann. 2020. fugashi, a tool for tokenizing Japanese in python. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 44–51, Online. Association for Computational Linguistics.

Christian Meilicke, Raúl García-Castro, Fred Freitas, Willem Robert van Hage, Elena Montiel-Ponsoda, Ryan Ribeiro de Azevedo, Heiner Stuckenschmidt, Ondřej Šváb Zamazal, Vojtěch Svátek, Andrei Tamilin, Cássia Trojahn, and Shenghui Wang. 2012. Multifarm: A benchmark for multilingual ontology matching. *Journal of Web Semantics*, 15:62–68.

Martina Miliani, Serena Auriemma, Fernando Alva-Manchego, and Alessandro Lenci. 2022. Neural readability pairwise ranking for sentences in Italian administrative language. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 849–866, Online only. Association for Computational Linguistics.

Ruslan Mitkov and Sanja Štajner. 2014. The fewer, the better? a contrastive study about ways to simplify. In *Proceedings of the Workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA 2014)*, pages 30–40, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Babak Naderi, Salar Mohtaj, Kaspar Ensikat, and Sebastian Möller. 2019. Subjective assessment of text complexity: A dataset for german language. *arXiv preprint arXiv:1904.07733*.

Tarek Naous, Michael J. Ryan, Mohit Chandra, and Wei Xu. 2023. Towards massively multi-domain multilingual readability assessment.

Rani Nelken and Stuart M. Shieber. 2006. Towards robust context-sensitive sentence alignment for monolingual corpora. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1611–168, Trento, Italy. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Vipul Raheja, and Oleksandr Skurzhanskyi. 2021. Text Simplification by Tagging. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–25, Online. Association for Computational Linguistics.

C Orasan, R Evans, and I Dornescu. 2013. Text simplification for people with autistic spectrum disorders. *Towards Multilingual Europe*, pages 287–312.

Gustavo Paetzold, Fernando Alva-Manchego, and Lucia Specia. 2017. MASSAlign: Alignment and annotation of comparable documents. In *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 1–4, Tapei, Taiwan. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. *arXiv preprint arXiv:2005.00333*.

Namoos Hayat Qasmi, Haris Bin Zia, Awais Athar, and Agha Ali Raza. 2020. SimplifyUR: Unsupervised lexical text simplification for Urdu. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3484–3489, Marseille, France. European Language Resources Association.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. Xl-wic: A multilingual benchmark for evaluating semantic contextualization. *arXiv preprint arXiv:2010.06478*.

Sebastian Ruder, Jonathan H Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel A Sarr, Xinyi Wang, et al. 2023. Xtreme-up: A user-centric scarce-data benchmark for under-represented languages. *arXiv preprint arXiv:2305.11938*.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing*, 6(4).

Andrey Sakhovskiy, Alexandra Izhevskaya, Alena Pestova, Elena Tutubalina, Valentin Malykh, Ivana Smurov, and Ekaterina Artemova. 2021. RuSimpleSentEval-2021 shared task: evaluating sentence simplification for russian. In *Proceedings of the International Conference "Dialogue*, pages 607–617.

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

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

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)*, pages 712–718, Melbourne, Australia. Association for Computational Linguistics.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. *arXiv preprint arXiv:1704.04154*.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Rico Sennrich and Martin Volk. 2010. MT-based sentence alignment for OCR-generated parallel texts. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*, Denver, Colorado, USA. Association for Machine Translation in the Americas.

AA Shatilov and AI Rey. 2021. Sentence simplification with rugpt3. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue"*, pages 1–13.

Advaith Siddharthan. 2004. Syntactic simplification and text cohesion. *Research on Language & Computation*, 4.

Lucia Specia. 2010. Translating from complex to simplified sentences. In *International Conference on Computational Processing of the Portuguese Language*, pages 30–39. Springer.

Lucia Specia, Sandra Maria Aluísio, and Thiago A Salgueiro Pardo. 2008. Manual de simplificação sintática para o português. *Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional (NILC-TR-08-06)*.

Sanja Stajner. 2014. Translating sentences from 'original' to 'simplified' spanish. *Procesamiento del Lenguaje Natural*, 53:61–68.

Sanja Stajner. 2021. Automatic text simplification for social good: Progress and challenges. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652, Online. 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, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Sanja Štajner, Marc Franco-Salvador, Simone Paolo Ponzetto, Paolo Rosso, and Heiner Stuckenschmidt. 2017. Sentence alignment methods for improving text simplification systems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 97–102, Vancouver, Canada. Association for Computational Linguistics.

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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Yasuhito Tanaka. 2001. Compilation of a multilingual parallel corpus. *Proceedings of PACLING 2001*, pages 265–268.

Sara Tonelli, Alessio Palmero Aprosio, and Francesca Saltori. 2016. Simpitiki: a simplification corpus for italian. In *CLiC-it/EVALITA*.

Jan Trienes, Jörg Schlötterer, Hans-Ulrich Schildhaus, and Christin Seifert. 2023. Patient-friendly clinical notes: Towards a new text simplification dataset.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of english in zero-shot cross-lingual transfer. *arXiv preprint arXiv:2106.16171*.

Matej Ulčar and Marko Robnik-Šikonja. 2022. Sequence to sequence pretraining for a less-resourced slovenian language. *arXiv preprint arXiv:2207.13988*.

Laurens van den Bercken, Robert-Jan Sips, and Christoph Lofi. 2019. Evaluating neural text simplification in the medical domain. In *The World Wide Web Conference*, WWW '19, page 3286–3292, New York, NY, USA. Association for Computing Machinery.

Laura Vásquez-Rodríguez, Matthew Shardlow, Piotr Przybyła, and Sophia Ananiadou. 2021. Investigating text simplification evaluation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 876–882, Online. Association for Computational Linguistics.

Sanja Štajner. 2014. Translating sentences from 'original' to 'simplified' spanish. *Procesamiento de Lenguaje Natural*, 53:61–68.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. 2020. Multi-SimLex: A Large-Scale Evaluation of Multilingual and Crosslingual Lexical Semantic Similarity. *Computational Linguistics*, 46(4):847–897.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages

4003–4012, Marseille, France. European Language Resources Association.

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

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

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

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

Jonathan Zheng, Ashutosh Baheti, Tarek Naous, Wei Xu, and Alan Ritter. 2022. Stanceosaurus: Classifying stance towards multilingual misinformation. *arXiv preprint arXiv:2210.15954*.

A Resource Summary

Here we provide a brief summary of the 34 monolingual text simplification parallel corpora surveyed in this work. Our summary focuses on the domain, target audience, and collection strategy of each resource.

A.1 English

ASSET (Alva-Manchego et al., 2020) is a high quality collection of 2,390 original sentences from the TurkCorpus (Xu et al., 2016) which sampled from Wikipedia. ASSET contains 10 manually written simplifications for each of the original sentences using a variety of rewrite operations.

The **Newsela English** corpus (Xu et al., 2015) was produced by professional writers at Newsela, a U.S. company dedicated to providing high-quality simplifications of informational content for schools. Each article was written at 5 levels, including the original article (level 0) and 4 levels of simplification. For this paper we use the neural CRF alignments provided by Jiang et al. (2020).

WikiAuto (Jiang et al., 2020) is a neural CRF aligned corpus of original and simple wikipedia documents. This is currently the largest text simplification resource available with over ten million original sentences. Once aligned, the number of sentence pairs reduces to just under 600,000. This is because most of the original and simple wikipedia articles are not exact rewrites.

A.2 Spanish

The **FIRST** corpus (Orasan et al., 2013) was collected as a part of the EU-funded FIRST project to develop a tool to assist people with autism spectrum disorder (ASD) in reading and understanding written documents. The corpus contains 25 original and simplified documents from literature, news, health, general culture, and instructions. The simplifications were performed manually by experts with experience working with individuals with ASD.

The **Simplext** corpus (Saggion et al., 2015) consists of 193 articles from the news outlet, Servimedia, which were manually simplified based on specific simplification recommendations (Anula, 2011). The articles span four domains: national news, international news, society, and culture.

The **Newsela Spanish** corpus (Xu et al., 2015) was created alongside the Newsela English corpus for the same audience of students. For analysis in this work the corpus was aligned between adjacent levels (i.e., 0-1, 1-2, etc.) using CATS aligner (Štajner et al., 2017).

A.3 Italian

The **Terence and Teacher** corpora (Brunato et al., 2015) were the first two Italian parallel corpora for ATS. The **Terence** corpus consists of 32 simplified short stories for children and the **Teacher** corpus consists of 18 texts from educational websites. The Terence corpus was simplified by experts with target rules, while the Teacher corpus was simplified by teachers targeting second language learners.

SIMPITIKI (Tonelli et al., 2016) is comprised of a Wikipedia-based sub corpus and a government-document-based sub corpus. **SIMPITIKI Wiki** used crowdsourced Wikipedia simplifications by extracting edits with keywords such as "simplified" from the edit history of an Italian Wikipedia dump. Prior work has shown that the quality of Wikipedia simplifications is not guaranteed (Xu et al., 2015), however, the simplifications were manually selected to ensure quality. This is the primary corpus studied as Simpitiki throughout this paper. SIMPITIKI PA was produced by the authors for comparing simplification operations and was later absorbed as a subset of the **AdminIT** corpus (Miliani et al., 2022). The authors manually simplified public administration documents about building permits and kindergarten admittance.

PaCCSS-IT (Brunato et al., 2016) instead extracted simplification pairs from a large corpus of text. The researchers assumed that in a large enough text dataset (i.e., scraped from the internet), both complex and simple sentences that have similar meanings were bound to exist. Brunato et al. (2016) found over 63,000 such matches using cosine similarity and an SVM using lexical, morpho-syntactic, and syntactic features. Upon manual analysis by the authors, it turned out that about 85% of the pairs were aligned correctly, while 74% of those correct pairs were actually simplifications.

A.4 French

CLEAR (Cardon and Grabar, 2019) is a parallel corpus of biomedical texts written in French and automatically aligned using a random forests classifier of 10 textual features. In a manual assessment of 30 documents, the authors found that 98.75% of alignments were correct suggesting a high-precision sentence alignment.

WikiLargeFR (Cardon and Grabar, 2020) is a machine-translated version of the English Wik-

iLarge corpus (Zhang and Lapata, 2017) using OpenNMT-py (Klein et al., 2017). It has 297,753 sentence pairs but different exact counts of complex and simple sentences when accounting for sentence splitting. It was created as a comparison with CLEAR for biomedical text simplification.

Alector (Gala et al., 2020) contains expert simplified versions of 79 texts selected at a 2-4th grade reading level. The authors showed that simplifying the texts reduced misreadings in dyslexic and low literacy readers.

A.5 Japanese

Japanese News (Goto et al., 2015) is created from Japan Broadcasting Corporation (NHK)'s online service NEWS WEB EASY which provides original news articles rewritten by Japanese instructors for simplicity. Goto et al. (2015) used a dynamic programming aligner to align 10,651 sentences. They also manually aligned 2,735 sentences.

EasyJapanese (Maruyama and Yamamoto, 2018) was created for the purpose of improving Japanese resources for foreign citizens. Since Japanese language learners usually have a limited vocabulary, the authors decided to produce a parallel corpus using a vocabulary of 2,000 common Japanese words. 5 students in the lab manually simplified 50,000 sentences with S-BLEU (Papineni et al., 2002) scores between simplifications of the same sentence ranging from 0.58 to 0.63. The corpus was built from a previous bilingual web crawl of Japanese and English news articles called the Tanaka corpus (Tanaka, 2001).

EasyJapaneseExtended (Katsuta and Yamamoto, 2018) included 34,400 more sentences from the Tanaka corpus (Tanaka, 2001) with simplifications crowdsourced from the CrowdWorks[7] platform. The authors measured the S-BLEU scores on 100 sentences that each of the 7 workers simplified and found that for 70% of the workers, S-BLEU scores exceeded 0.4.

A.6 Brazilian Portuguese

PorSimples (Aluísio and Gasperin, 2010) was one of the first ATS projects. It had the express purpose of simplifying texts for individuals with reading difficulties. For this project, Caseli et al. (2009) collected a parallel corpus of 104 news articles. The articles were simplified at 2 levels by a linguist specializing in text simplification. In "natural" simplifications, the linguist could choose how to simplify the text. In "strong" simplifications, the linguist had to follow very specific rules (Specia et al., 2008; Aluísio et al., 2008a).

A.7 German

Simple German (Battisti et al., 2020) started from Klaper et al. (2013)'s work that scraped texts from the internet and aligned them using a monolingual sentence alignment algorithm of Barzilay and Elhadad (2003). Battisti et al. (2020) further improved upon this with much more data and better alignment algorithms of CATS (Štajner et al., 2018) and MASSAlign (Paetzold et al., 2017). The original paper reports 378 documents with 17,121 original and 21,072 simple sentences. Note that, these numbers differ from those in Table 1 as the availability of online articles has changed since the original publication.

TextComplexityDE (Naderi et al., 2019) was created to measure text complexity in German. 1000 sentences were taken from German Wikipedia and 100 sentences from Simple German (Klaper et al., 2013). German second language learners rated the sentences on a 7-point Likert scale for complexity. The 250 most complex sentences were manually simplified by native speakers.

GEOLinoTest (Mallinson et al., 2020) was built as an evaluation dataset for a zero-shot ATS model. Mallinson et al. (2020) extracted 20 articles about nature, physics, and people from GeoLino, a children's magazine. A German linguist simplified them to a five to seven-year-old reading level.

German News (Säuberli et al., 2020) contains 3,616 sentences simplified by the Austrian Press Agency on politics, economy, culture, and sports. The authors trained neural simplification models on the corpus and found success, including a simple sentence matched to itself during training (Palmero Aprosio et al., 2019).

The **Klexikon** corpus (Aumiller and Gertz, 2022) is a mapping of documents from German Wikipedia to the children's encyclopedia site: Klexikon [8]. Klexikon targets German readers aged six to twelve. Like WikiAutoEN, Klexikon is a large-scale alignment of documents, however, unlike similar resources, Klexikon does not yet have a gold-standard automatic sentence alignment. The authors of Klexikon are working to release this alignment which will make Klexikon a very large

[7] https://crowdworks.jp/

[8] https://klexikon.zum.de

German text simplification resource.

Simple Patho (Trienes et al., 2023) is an upcoming biomedical text simplification corpus for German of 851 clinical reports simplified by nine medical students. Due to privacy concerns, the dataset is not yet available. When it is released, it will serve as a large and high-quality medical text simplification corpus for the community.

A.8 Basque

The Corpus of Basque Simplified Texts (CBST) (Gonzalez-Dios et al., 2018) contains 227 sentences from 3 science popularization documents simplified to two distinct levels. Two people simplified the documents: a translator without simplification experience who focused on simplification guidelines (Mitkov and Štajner, 2014) and a language teacher who focused on intuitive transformations.

A.9 Danish

DSim (Klerke and Søgaard, 2012) extracted 3,701 pairs of news telegrams from the Danish Broadcasting Corporation's news and educational services. The corpus was simplified by journalists to help reading-impaired adults and adult learners of Danish. The documents were automatically sentence-aligned using TF-IDF scores.

A.10 Urdu

SimplifyUREval (Qasmi et al., 2020) was a corpus made for evaluating the ATS model SimplifyUR. The model used word substitutions to propose simplifications to Urdu. The evaluation corpus contains 500 sentences from newspapers, magazines, books, and literary journals manually simplified by a linguist with a doctorate in Urdu. Two additional native Urdu speakers manually verified 50 sentences and had an inter-annotator agreement of 0.9, measured by Cohen's Kappa.

A.11 Russian

RuWikiLarge (Sakhovskiy et al., 2021) is a machine translated version of EnWikiLarge (Zhang and Lapata, 2017). It has 248,111 sentence pairs but different exact counts of complex and simple sentences when accounting for sentence splitting. It was created as a resource for the RuSimpleSentEval shared task (Sakhovskiy et al., 2021).

RuSimpleSentEval (RSSE) (Sakhovskiy et al., 2021) was mined from Russian Wikipedia's most popular pages also for the RuSimpleSentEval. Crowdsource workers on Yandex Toloka were asked to simplify the sentences.

RuAdapt (Dmitrieva and Tiedemann, 2021) was created from 6 collections of several novels each and 16 individual classical and modern Russian literature books. The simplified books were prepared by Russian-as-a-Foreign-Language (RaaFL) teachers. The corpus was aligned with Bleualign (Sennrich and Volk, 2010) and CATS (Štajner et al., 2017). In addition, researchers contributed both encyclopedic simplifications and fairytale simplifications (Dmitrieva et al., 2021).

A.12 Slovene

The **SloTS** corpus (Gorenc and Robnik-Šikonja, 2022) pulls from 10 existing texts simplified by the RISA Institute. The RISA Institute is an organization that publishes easy-to-read Slovenian novels and news. SloTS is a collection of about 1,000 sentence pairs sampled from 10 novels and manually aligned between the original and simplified versions. This dataset was used to train a Slovene text simplification model built on SloT5 (Ulčar and Robnik-Šikonja, 2022).

A.13 Arabic

Saaq al-Bambuu (Al-Sanousi, 2016) is an internationally acclaimed Arabic novel that has been rewritten for Arabic-as-a-second-language learners. Khallaf and Sharoff (2022) sampled 2,980 parallel sentences from the original and simplified books at two different levels. Unfortunately, due to copyright restrictions, the corpus is not available publicly.

B Corpus Analysis

This section provides some key statistics of the corpora introduced in Section 3.4. In performing this analysis, we hope to highlight the differences between various corpora and offer greater insight into their quality and composition to facilitate future research.

B.1 Basic Statistics

The Basic Statistics computed were vocab size, token count, average tokens per sentence, average characters per token, and average sentences per doc. All of the statistics besides average sentences per document depend heavily on the word tokenization of the corpus. For space-delimited languages, we used the Toktok tokenizer (Dehdari, 2014) from the

natural language toolkit (NLTK) (Bird et al., 2009). For our work, this included all languages besides Urdu and Japanese. For Urdu tokenization, we used UrduHack[9], a Python library built for academic researchers and professional developers working on Urdu NLP projects. For Japanese, we used fugashi (McCann, 2020), a tool for Japanese tokenization in Python. We used the unidic dictionary (Den et al., 2008) to define the Japanese vocabulary for tokenization.

Basic statistics results are reported in Table 6. In general, the trend between original and simplified texts was reduced vocab size, reduced token count, reduced sentence length, and reduced word length. This corresponds well with the expected simplification operations of replacing longer, more complicated words with shorter, more common words. It also aligns with a common simplification strategy of splitting longer sentences into two or more short sentences. This explains why the average sentence length decreased but in many document-aligned corpora, the average sentences per document increased. There were some exceptions to this. Notably, the Japanese corpora had a higher token count and higher average tokens/sentence. This could be due to the limited vocabulary used when creating these two corpora. As a part of the Easy Japanese corpus creation authors were limited to just 2,000 Japanese words. The fugashi tokenizer identified more than 2,000, but still reported a large drop in vocab size. The authors may have needed to be creative with how they chose to rewrite sentences using the limited vocabulary. This could've led to many edits where the author explained a complex idea in several simple words instead of using one more complicated word. Another easily explained set of outliers is SimplifyUR and RSSE having larger vocab sizes from simple to complex. Both of these corpora allow multiple translations of the same original sentence. This means for a given sentence pair with the same original sentence the original vocab size will remain the same while the simple vocab size might increase.

B.2 Document-level Compression

In order to measure the editing levels on a document scale, we investigated document-level compression. Document-level compression is the ratio of the number of characters in the simple document to the number of characters in the original document (Xu et al., 2015). A low document compression ratio indicates a lot of deletions between the original and simplified text, while a high document compression ratio suggests lengthier operations like sentence splitting or rephrasing.

We report the compression ratios of all document-aligned corpora in Figure 6. Most of the compression ratios are approximately normally distributed. For many of the corpora, the compression ratio is centered around one, meaning the original and simplified documents are about the same length. This closely matches the low edit ratios that many of the corpora have (see section B.4). A few of the corpora (Simplext, Newsela, and German News) have lower means instead suggesting more significant document-level edits, such as deletion of entire sentences.

B.3 Sentence-level Edit Operations

Edit operations describe the types of simplifications that were performed to transform from an original sentence to a simplified sentence. These can only be computed for corpora that are sentence aligned. There are 6 edit operations we tracked from the alignments. Each operation corresponds to a mapping (x:y) of x original sentences to y simple sentences. The operations are deletion (1:0), split (1:n), same (1:1), change (1:1), merge (n:1), and insert (0:1). To determine the difference between "same" and "change", the Levenshtein distance (Levenshtein, 1965) was measured between the original and simplified sentence. This distance was divided by the length of the longer sentence. If the difference was greater than 5% then the sentences were marked as changed, otherwise, they were considered the same. Levenshtein distance was calculated using the *fuzzywuzzy*[10] library.

Table 7 shows the distribution of edit operations. Sentence-level edit operations were reported for both document-aligned corpora as well as sentence-aligned corpora that used sentence splitting (1:n mapping). The most common edit operation across corpora was changing the original sentence, followed by keeping the same sentence, then splitting, deleting, and merging. Interestingly, Spanish corpora, like English ones (Xu et al., 2015; Jiang et al., 2020), had more deletion operations than most of the other languages. For the sentence-only aligned corpora, this was because original sentences without simplifications were not included, but this was

[9] https://docs.urduhack.com/en/stable/

[10] https://github.com/seatgeek/thefuzz

Corpus	Lang	Vocab Size		Token Count		Avg Tok/Sent		Avg Char/Tok		Avg Sent/Doc	
		orig ↑	simp ↓	orig ↑	simp ↓	orig ↑	simp ↓	orig ↑	simp ↓	orig	simp
DSim	da	57,308	40,220	953,201	796,201	19.91	13.15	5.57	5.36	—	—
GEOLino	de	4,467	4,266	19,185	17,889	16.01	14.93	**5.68**	**5.74**	—	—
German News A2	de	23,542	7,764	147,905	78,946	20.30	11.23	6.40	5.79	4.03	3.89
German News B2	de	25,039	10,473	160,188	93,283	20.14	12.75	6.43	5.99	4.96	4.56
Klexikon	de	706,243	55,868	15,240,505	1,239,694	19.77	12.80	6.42	5.53	266.53	33.48
Simple German	de	35,763	18,753	313,622	199,861	24.50	23.79	6.60	6.38	57.16	37.50
TextComplexityDE	de	3,068	2,760	7,485	7,092	29.94	28.37	6.78	6.62	10.87	10.87
ASSET	en	**11,998**	**19,320**	521,940	448,376	22.13	19.01	5.28	5.18	—	—
Newsela EN 0-1	en	68,972	61,115	2,187,046	1,881,631	23.95	19.83	5.08	5.06	48.52	50.41
Newsela EN 1-2	en	61,115	53,673	1,881,631	1,733,011	19.83	16.84	5.06	4.98	50.41	54.67
Newsela EN 2-3	en	53,673	42,879	1,733,011	1,458,744	16.84	13.93	4.98	4.86	54.67	55.65
Newsela EN 3-4	en	42,879	34,104	1,458,744	1,144,534	13.93	11.48	4.86	4.75	55.65	53.00
WikiAuto	en	2,009,681	419,496	265,352,569	22,170,411	26.16	17.86	5.19	4.96	—	—
Newsela ES 0-1	es	27,950	23,452	323,034	257,905	28.28	23.31	5.37	5.36	47.01	45.52
Newsela ES 1-2	es	23,452	20,582	257,905	225,659	23.31	19.35	5.36	5.31	45.52	48.00
Newsela ES 2-3	es	20,582	16,148	225,659	178,117	19.35	14.72	5.31	5.21	48.00	49.79
Newsela ES 3-4	es	16,148	11,695	178,117	122,064	14.72	11.42	5.21	5.13	49.79	43.98
Simplext	es	8,071	3,191	38,731	25,409	34.96	14.59	5.47	5.34	5.74	9.03
CBST Intuitive	eu	1,697	1,586	4,575	4,447	19.98	14.53	**6.34**	**6.43**	76.33	102.00
CBST Structural	eu	1,697	1,654	**4,575**	**4,793**	19.98	16.82	6.34	6.29	76.33	95.00
Alector	fr	5,728	5,024	28,283	26,179	22.99	21.96	4.78	4.73	15.57	15.09
CLEAR	fr	11,743	11,205	119,465	118,212	25.99	25.72	**5.72**	**5.73**	—	—
WikiLargeFR	fr	205,933	173,827	8,763,745	6,384,020	28.54	20.70	5.03	4.94	—	—
AdminIT	it	3,420	3,394	29,581	28,784	38.07	37.72	6.08	5.90	—	—
PaCCSS-IT	it	10,478	9,853	580,389	519,211	9.21	8.24	**4.75**	**4.79**	—	—
SimpitikiWiki	it	9,188	9,175	41,899	41,375	72.87	71.96	5.60	5.60	—	—
Teacher	it	1,485	1,061	4,225	3,367	20.71	17.27	4.89	4.76	11.33	10.83
Terence	it	3,681	3,219	19,455	18,881	18.80	17.81	5.13	5.04	32.34	33.12
Easy Japanese	ja	10,331	3,401	**489,302**	**517,651**	**9.79**	**10.35**	1.51	1.49	—	—
Easy Japanese Ext	ja	18,888	5,305	**433,341**	**503,035**	**12.38**	**14.37**	1.55	1.49	—	—
PorSimples Natural	pt-br	9,983	9,527	**64,610**	**65,174**	20.97	13.52	**5.39**	**5.53**	20.01	31.31
PorSimples Strong	pt-br	**9,527**	**9,601**	**65,174**	**65,552**	13.52	12.25	**5.53**	**5.57**	31.31	34.76
RSSE Corpus	ru	**16,467**	**24,307**	138,319	95,067	20.33	13.97	6.73	6.54	—	—
RuAdapt Ency A-B	ru	4,609	3,842	11,085	9,804	12.58	9.96	6.08	5.84	14.44	16.13
RuAdapt Ency A-C	ru	4,927	3,844	11,931	9,809	13.59	9.96	6.05	5.84	14.16	15.89
RuAdapt Ency B-C	ru	27,268	26,200	113,817	110,463	14.28	13.37	6.10	6.08	29.96	31.06
RuAdapt Fairytales	ru	1,688	1,512	4,391	4,289	14.16	10.62	**5.08**	**5.32**	34.44	44.89
RuAdapt Literature	ru	55,321	42,655	368,499	327,228	15.26	11.58	5.14	5.08	168.90	197.62
RuWikiLarge	ru	331,063	275,644	5,760,207	4,540,009	20.70	15.68	5.95	5.79	—	—
SloTS	sl	5,871	2,723	21,804	10,646	18.46	8.27	4.72	4.44	—	—
SimplifyUR	ur	**1,469**	**1,475**	6,580	6,561	8.94	8.91	4.25	4.22	—	—

Table 6: Basic statistics about all of the corpora we analyzed. Typically the vocab size, token count, average tokens per sentence, and average characters per token all decrease from original to simplified texts. Outliers of this trend are highlighted in bold.

Corpus	Deleted 1:0 (%)	Split 1:n (%)	Same 1:1 (%)	Changed 1:1 (%)	Merged n:1 (%)	Inserted 0:1 (%)
English						
Newsela EN 0-1	22.7	12.6	39.5	25.2	0.0	12.2
Newsela EN 1-2	17.4	13.4	36.4	32.8	0.0	10.7
Newsela EN 2-3	27.2	11.8	23.5	37.5	0.0	15.9
Newsela EN 3-4	33.0	10.4	21.2	35.4	0.0	17.8
WikiAuto	94.2	0.9	1.2	3.7	0.0	44.8
Russian						
RuAdapt Fairytales	0.0	22.6	4.2	73.2	0.0	0.0
RuAdapt Ency B-C	0.0	3.5	79.3	17.2	0.0	0.0
RuAdapt Ency A-C	0.0	10.8	31.8	57.4	0.0	0.0
RuAdapt Ency A-B	0.0	10.7	35.4	53.9	0.0	0.0
RuAdapt Literature	0.0	11.8	36.6	51.6	0.0	0.0
Italian						
Terence	0.7	4.1	35.4	57.0	2.9	0.4
Teacher	6.9	9.3	8.3	59.3	16.2	1.5
Spanish						
Newsela ES 0-1	29.1	20.0	19.1	31.6	0.2	0.7
Newsela ES 1-2	18.3	19.8	24.2	37.7	0.1	0.5
Newsela ES 2-3	27.5	22.9	13.3	36.0	0.3	0.4
Newsela ES 3-4	38.5	19.1	11.2	31.1	0.2	0.3
Simplext	16.2	32.2	3.5	47.4	0.7	19.3
German						
TextComplexityDE	0.0	0.0	0.4	99.6	0.0	0.0
German News A2	0.0	22.5	0.8	37.6	39.1	0.0
German News B2	0.0	23.0	1.4	33.2	42.4	0.0
Brazilian Portuguese						
PorSimples Natural	0.6	38.7	22.1	38.3	0.3	0.8
PorSimples Strong	0.2	9.9	73.7	16.0	0.1	0.1
Basque						
Structural CBST	0.0	22.3	24.0	51.5	2.2	0.0
Intuitive CBST	0.0	25.8	27.1	45.9	1.3	0.0

Table 7: Edit operations for all document-level corpora with sentence alignment. All operations are reported as a percentage of original sentences besides "inserted" which is reported as a percentage of simplified sentences.

true even amongst document-aligned corpora.

B.4 Character-level Edit Distances

We analyzed the edit distance distribution (Vásquez-Rodríguez et al., 2021) on all corpora with sentence-level alignments to understand the strength of the edits at a sentence scale. Low edit distances indicate smaller simplifications while high edit distances indicate big changes. We again used character-based Levenshtein distance to measure edit distance. The Levenshtein distance was divided by the length of the longer sentence to obtain a ratio from zero to one. Zero meant the sentence wasn't edited at all, while one meant the sentence was completely different.

The edit distance ratios can be found in Figure 5. For about half of the corpora, the mean edit distance fell below 20%. For the other half mean edit distance ratios ranged from 0.2 to 0.6. Corpora with an approximately normal distribution and higher variance demonstrate a wide variety of both minor and major sentence edits. Corpora with low means and a high concentration of low edit ratios primarily consist of slight modifications.

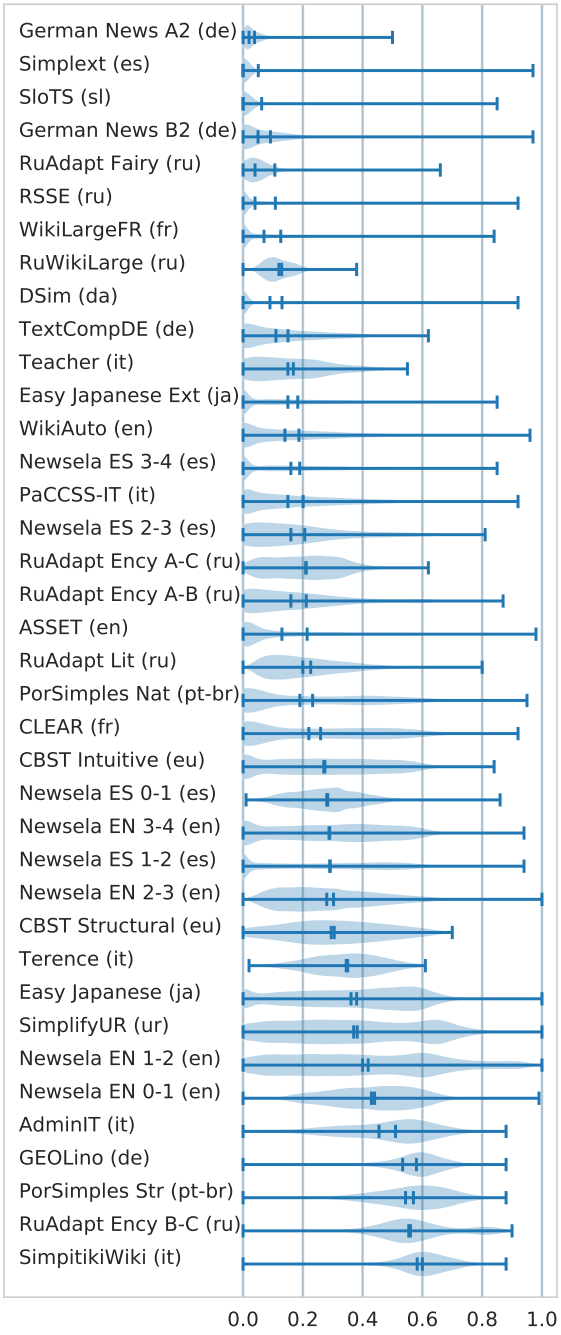

Figure 5: Violin Plots showing the minimum, maximum, mean, and median values for edit distances for all of the sentences in each corpus. Distributions estimated using Gaussian kernel density estimation.

C Experimental Details

Fine-tuning For fine-tuning experiments we used the mT5 Base (Xue et al., 2021) architecture (580M Parameters). We used the sentence piece tokenizer (Kudo and Richardson, 2018) and limited inputs/targets to a length of 128 tokens. We used a learning rate of *5e-5* and the AdamW Optimizer (Loshchilov and Hutter, 2019). Decoding was done using beam search with 4 beams. The train batch size was set to 8. We train for 5 epochs. For any dataset without a development set, we removed

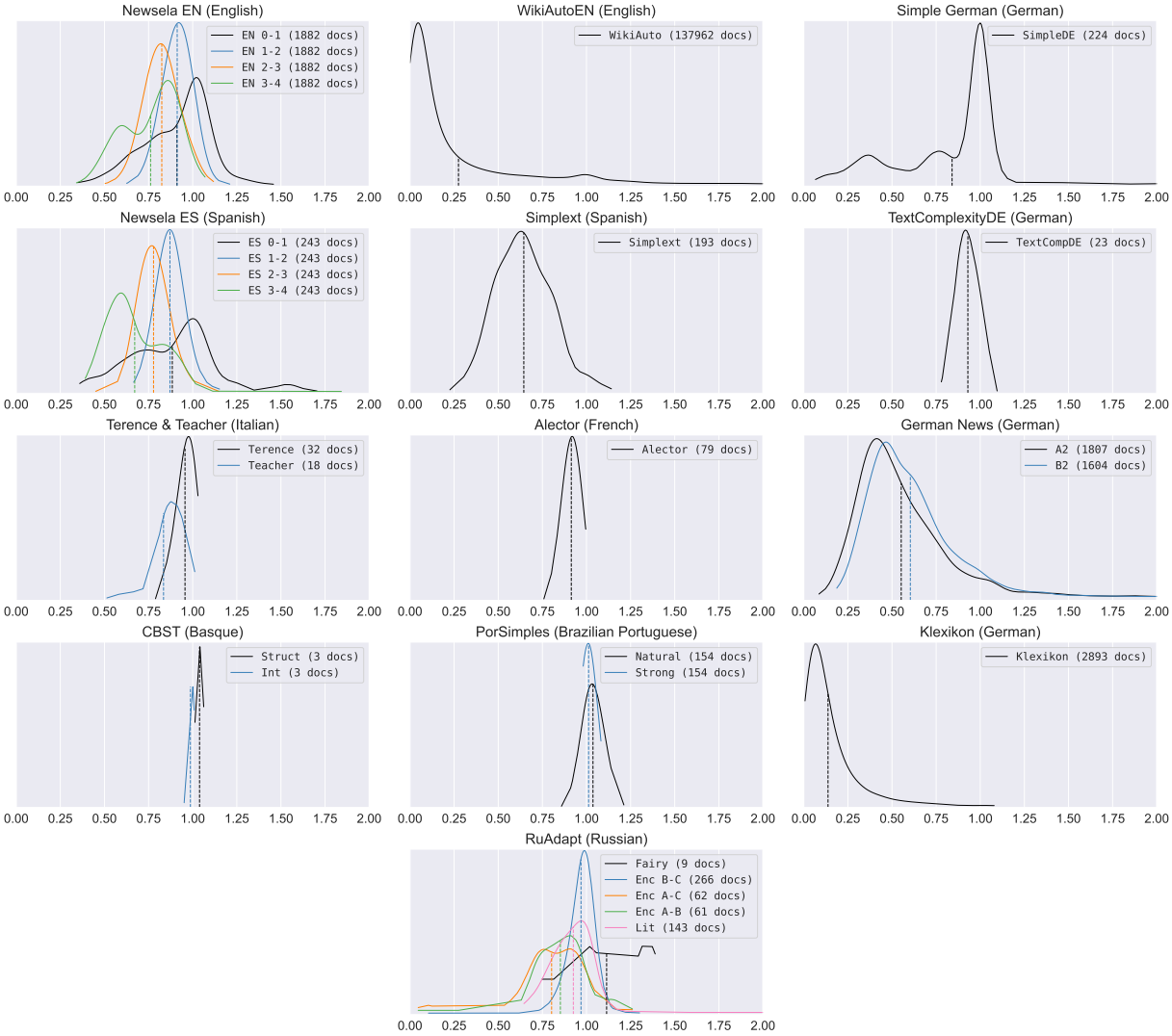

Figure 6: Distribution of document-level compression ratio for document-aligned corpora, smoothed by Gaussian kernel density estimation. Means are marked by dashed lines.

10% of the training set up to 1,000 sentences to create our own dev set. The training was performed on three NVIDIA A40 GPUs.

We also perform preprocessing on the training data inputs. We compute sentence BLEU scores between the original and reference simplifications. For any sentence pairs with an S-BLEU score below 10 or above 70, we remove it from the training set (Maddela et al., 2021). This helps reduce both misaligned and identical pairs. Following Martin et al. (2020) we also add control tokens to the input sentences. We include a character length compression token <NC_[#]>, a Levenshtein similarity token <LS_[#]>, a dependency tree depth ratio token <DR_[#]>, and a word frequency rank token <WR_[#]>. For each token, we compute the respective measure on both the original and simple sentences and include the ratio between [0.05, 2] in increments of 0.05. For example, a prefix might have the form: "<NC_0.9> <LS_0.8> <DR_0.9> <WR_1.05>". More details on computing these tokens can be found in the original paper (Martin et al., 2020). For our dev set grid search to find the optimal token setting we perform a 3x3x3x3 search around the average values from the training set.

Fewshot For fewshot experiments we used BLOOM (Scao et al., 2022) (176B Parameters). We used the HuggingFace inference API[11] to prompt BLOOM with sampling. We used a temperature of 1.0 and a repetition penalty of 0.0. We formatted prompts to BLOOM as:

```
Original: "[EXAMPLE 1 ORIGINAL]"
```

[11] https://huggingface.co/inference-api

```
Simple: "[EXAMPLE 1 SIMPLIFICATION]"
...
Original: "[EXAMPLE N ORIGINAL]"
Simple: "[EXAMPLE N ORIGINAL]"

Original: "[TEST ORIGINAL]"
Simple: "
```

The prefixes "Original" and "Simple" were always in English. Outputs were appended to the prompt and repeated back to the model until the output contained an end quotation followed by a new "`Original:`". This was to prevent half-completed simplifications.

Manual Evaluation Table 9 shows the key provided to the human annotators in each language. Annotators were volunteers with fluency in the target language for annotation. We randomly sampled 20 sentences from the test set of each of the four datasets and we used these 20 sentences to compare across all models. For the "Reference" baseline if any sentence had more than one possible reference simplification we randomly sampled a reference. For any of the system outputs, if the sentence was completely nonsense annotators were instructed to rate the sentence with a score of 1 on all aspects.

All annotators were volunteers that were informed that we were "measuring the quality of various machine learning models that have been trained/prompted to simplify text". All of the volunteers were told they were assessing sentences to be used in evaluation for a research project. For any student employees that volunteered their time for evaluation, they were paid at their normal hourly rate of $18 per hour. Other colleagues that volunteered their time did so on a strictly voluntary basis.

Lang →		en			ru					da
Approach	# shots	ASSET	Newsela EN	WikiAuto EN	RuWikiLarge	RSSE	RuAdapt Ency	RuAdapt Fairy	RuAdapt Lit	DSim
Fine-Tuned	NA	35.98	38.60	42.46	32.01	31.66	26.42	34.79	41.75	31.40
Zero Shot	0	35.52	33.22	34.73	31.47	20.09	33.20	12.74	30.95	35.84
Semantic Similarity	1	36.21	35.29	40.24	35.95	30.18	40.21	33.49	37.05	38.27
	2	36.37	37.70	40.97	36.93	29.75	42.22	35.51	37.61	38.84
	3	36.53	37.71	41.66	37.22	29.79	40.76	37.07	39.06	38.09
	5	36.79	38.82	42.25	36.84	31.08	41.01	38.43	39.89	37.57
	10	36.19	38.65	42.66	37.59	31.33	39.45	40.93	40.44	31.71
	20	36.80	39.26	42.83	37.71	31.22	39.54	38.95	40.32	29.88
Random Sampling	1	35.21	34.02	35.16	33.61	28.23	33.45	20.89	32.26	34.96
	2	35.77	34.87	36.40	34.20	28.44	34.07	24.24	33.14	35.29
	3	35.37	34.31	36.21	34.19	29.40	34.24	22.84	33.41	35.39
	5	35.89	34.39	36.00	32.89	29.30	33.17	25.91	33.45	34.81
	10	36.10	35.35	36.86	34.63	28.69	33.19	29.16	33.92	30.15
	20	36.21	34.73	37.14	34.63	29.60	33.71	31.83	33.89	27.98

Lang →		de			it					pt-br
Approach	# shots	German News	TextCompDE	GEOLino	PaCCSS-IT	Terence	AdminIT	Simpitiki	Teacher	PorSimples
Fine-Tuned	NA	36.04	30.26	26.44	57.30	23.92	23.42	4.11	29.84	31.54
Zero Shot	0	32.48	32.26	29.59	35.42	35.91	32.43	18.43	28.75	35.38
Semantic Similarity	1	36.19	37.63	38.16	51.42	34.95	37.06	27.73	33.97	36.72
	2	36.68	38.60	39.65	49.15	37.25	36.69	26.42	34.14	37.46
	3	36.78	41.03	39.44	48.00	34.95	35.67	27.06	29.41	38.85
	5	37.79	38.81	39.5	45.48	35.94	38.16	26.94	39.10	38.96
	10	37.69	38.93	39.7	37.31	35.39	35.21	27.20	32.62	41.34
	20	36.76	38.93	39.44	33.45	35.17	35.21	27.73	33.46	39.94
Random Sampling	1	32.58	34.94	35.89	38.84	33.60	33.31	21.96	33.94	37.68
	2	34.09	35.37	36.11	39.11	34.15	34.77	23.79	25.01	36.82
	3	34.92	34.13	35.22	38.51	33.96	35.98	25.22	31.41	36.57
	5	34.71	36.68	34.5	37.41	32.01	34.24	25.01	32.30	36.53
	10	35.58	38.07	35.42	35.01	31.60	35.67	25.04	30.82	35.93
	20	35.53	38.07	34.62	30.29	34.38	35.67	25.04	34.39	35.31

Lang →		fr		sl	ja		es		ur	eu
Approach	# shots	WikiLargeFR	CLEAR	SloTS	Easy JA	Easy JA Ext	NewselaES	Simplext	SimplifyUR	CBST
Fine-Tuned	NA	35.20	34.86	36.56	67.36	43.15	29.89	35.62	33.18	24.26
Zero Shot	0	35.71	35.75	27.37	41.71	30.53	34.15	25.36	28.61	34.43
Semantic Similarity	1	36.29	38.06	30.75	48.71	46.08	37.07	32.50	39.12	35.09
	2	35.22	39.03	37.38	54.89	49.39	36.90	38.05	50.69	38.89
	3	36.40	40.16	37.20	55.38	47.01	38.18	39.28	51.23	38.84
	5	36.75	39.34	37.55	57.29	49.30	38.12	40.26	53.60	39.50
	10	36.33	39.21	36.91	58.67	47.50	38.42	39.90	55.89	37.30
	20	37.72	38.45	37.24	59.42	46.55	38.42	39.75	56.36	38.31
Random Sampling	1	35.64	36.12	27.43	38.35	40.70	34.25	29.57	40.38	36.40
	2	36.67	34.64	31.83	37.85	41.38	34.55	33.51	49.00	36.00
	3	36.07	36.92	33.95	35.90	40.04	34.16	35.56	49.28	36.10
	5	36.17	35.54	33.60	35.92	42.55	33.64	36.81	50.34	37.39
	10	36.70	35.80	34.18	39.34	42.50	34.27	37.40	52.93	36.83
	20	35.80	36.13	35.30	37.89	42.11	33.91	39.27	50.02	38.86

Table 8: SARI Scores for BLOOM Fewshot Experiments

	Adequacy (is the meaning preserved?)
1:	The subject of the sentence has changed entirely and is entirely unrelated
2:	The meaning has been seriously altered (negated or changed)
3:	**Two or more important pieces** of information have been added or removed
4:	Meaning is similar but **one** piece of information has been added or removed
5:	Meaning is preserved aside from minor unimportant information
	Fluency (is the simplification eloquent/grammatical?)
1:	The simplification is completely unreadable
2:	The simplification suffers from many serious grammar issues (nearly unreadable)
3:	The simplification has **two or more** grammatical mistakes
4:	The simplification has a minor grammatical issue or is written strangely in one place
5:	The simplification is perfectly eloquent as if written by a human
	Simplicity (is the simplification actually simpler?)
1:	The simplification is actually harder to understand (ex. more complex terms used)
2:	The simplification is about the same difficulty as the original
3:	The simplification is mildly simpler, but this simplification does not help readability
4:	The simplification is actually simpler
5:	The simplification is vastly simpler and could help someone better understand

Table 9: Manual evaluation key provided to annotators

	Sentence	Translated
	AdminIT (Italian)	
Original	E' presente anche il personale esecutivo che provvede allo sporzionamento delle portate.	There is also the executive staff who arrange portioning of the courses.
Simple	È presente anche il personale che divide le portate in porzioni.	There is also the staff who divide the courses into portions.
	ASSET (English)	
Original	The Apostolic Tradition, attributed to the theologian Hippolytus, attests the singing of Hallel psalms with Alleluia as the refrain in early Christian agape feasts.	
Simple	The Apostolic Tradition was created by the religion expert Hippolytus. It shows the singing of Hallel psalms with Alleluia as the refrain in early Christian feasts.	
	CBST (Basque)	
Original	Horrekin batera, gure planeta eta eguzki-sistema gainerakoekin alderatu nahi dituzte, eta ikusi nahi dute ea horrelakoak fenomeno bakanak diren edo oso arruntak diren unibertsoan.	At the same time, they want to compare our planet and the solar system with the rest, and they want to see if such phenomena are rare or very common in the universe.
Simple	Gainera, gure planeta eta eguzki-sistema gainerakoekin alderatu nahi dituzte; fenomeno horiek bakanak edo arruntak dira unibertsoan? hori ikusi nahi dute.	They also want to compare our planet and the solar system with the rest; Are these phenomena rare or common in the universe? they want to see that.
	CLEAR (French)	
Original	une étude concernant les entretiens motivationnels suggérait que cette intervention était bénéfique contre la consommation de cannabis	a study on motivational interviewing suggested that this intervention was beneficial against cannabis use
Simple	l' une des deux études concernant des entretiens motivationnels suggérait que cette intervention était bénéfique sur la consommation de cannabis signalée	one of two studies involving motivational interviewing suggested that this intervention was beneficial on reported cannabis use
	DSim (Danish)	
Original	Stigende vandstand i floderne i det østlige Tjekkiet forvandlede i aftes hundredvis af boliger i området til dødsfælder .	Rising water levels in the rivers in the eastern Czech Republic last night turned hundreds of homes in the area into death traps.
Simple	I det østlige Tjekkiet stiger vandstanden i floderne .	In the eastern Czech Republic, the water level in the rivers is rising.
	EasyJA (Japanese)	
Original	君が言ったことで、僕はびっくりした。	What you said surprised me.
Simple	あなたが言ったことで、私は驚いた。	What you said surprised me.
	EasyJAExt (Japanese)	
Original	彼の不注意にはあきれてしまった。	I was appalled at his carelessness.
Simple	彼の不注意には言葉を失う。	His carelessness leaves me speechless.
	GEOLino (German)	
Original	Denn sie sind zwar mutig, aber durchaus nicht lebensmüde.	Because they are courageous, but by no means tired of life.
Simple	Denn sie sind zwar mutig, aber nicht lebensmüde.	Because they are courageous, but not tired of life.
	GermanNews (German)	
Original	Jedes Kalb erhält spätestens sieben Tage nach der Geburt eine eindeutig identifizierbare Lebensnummer, die in Form von Ohrmarken beidseitig eingezogen wird.	Each calf receives a clearly identifiable life number no later than seven days after birth, which is recorded on both sides in the form of ear tags.
Simple	In Österreich bekommt jedes Kalb kurz nach der Geburt eine Nummer	In Austria, every calf is given a number shortly after birth.

Table 10 continued from previous page

	Sentence	Translated
	NewselaEN (English)	
Original	Putting these parts into jet engines is just what the advanced manufacturing industry has been waiting for: evidence that shows that mainstream manufacturers have figured out how to make the materials and the process work.	
Simple	Putting these parts into jet engines shows that companies have figured out how to make the materials and the process work.	
	NewselaES (Spanish)	
Original	Para el proyecto de Apple, Taylor-Young tomó fotos de paisajes urbanos bajo la lluvia con su iPhone 6.	For the Apple project, Taylor-Young took photos of cityscapes in the rain with her iPhone 6.
Simple	Para el proyecto de Apple, utilizó su iPhone 6 para tomar fotografías de las calles lluviosas de la ciudad.	For the Apple project, she used her iPhone 6 to take pictures of the rainy streets of the city.
	PaCCSS-IT (Italian)	
Original	Anche per questa si chiede l' immediata eseguibilità : Chi è favorevole ?	For this too , immediate execution is requested : Who is in favor ?
Simple	Chiedo l' immediata eseguibilità : Chi è favorevole ?	I ask for immediate execution : Who is in favor ?
	PorSimples (Brazilian Portuguese)	
Original	No Eldorado do Sul poderá ser construído um estádio provisório.	In Eldorado do Sul, a provisional stadium could be built.
Simple	No Eldorado do Sul talvez construam um estádio provisório.	In Eldorado do Sul, perhaps they will build a temporary stadium.
	RSSE (Russian)	
Original	В природном очаге заражение обычно происходит через укус блохи, ранее питавшейся на больном грызуне.	In a natural focus, infection usually occurs through the bite of a flea that previously fed on a sick rodent.
Simple	Блоха может заразить укусом, если ранее она кусала больного грызуна.	A flea can infect with a bite if it has previously bitten a sick rodent.
	RuAdaptEncy (Russian)	
Original	Достоевский женился на стенографистке Анне Григорьевне Сниткиной, которая стала ему близким другом и помощником.	Dostoevsky married the stenographer Anna Grigorievna Snitkina, who became his close friend and assistant.
Simple	Достоевский женился на Анне Григорьевне Сниткиной.	Dostoevsky married Anna Grigoryevna Snitkina.
	RuAdaptFairytales (Russian)	
Original	Пустил стрелу средний брат — полетела стрела к богатому купцу во двор.	The middle brother fired an arrow - an arrow flew to the rich merchant in the yard.
Simple	Стрела среднего брата прилетела на богатый купеческий двор.	The arrow of the middle brother flew to the rich merchant's yard.
	RuAdaptLit (Russian)	
Original	Попала бы моя книжка в лапки какой-нибудь девочке в зеленом платьице... Села бы она у камина с моим сочинением, читала бы, перелистывала бы и улыбалась.	My book would fall into the paws of some girl in a green dress ... She would sit by the fireplace with my essay, read, leaf through and smile.
Simple	И какая-нибудь девочка сидела бы у камина с моей книжкой, читала бы и улыбалась.	And some girl would sit by the fireplace with my book, read and smile.
	RuWikiLarge (Russian)	
Original	Он служил во французском флоте, а в 1889 и 1890 годах служил в команде фрегата Iphig nie и несколько лет провел в Кочинчине.	He served in the French Navy and in 1889 and 1890 was in command of the frigate Iphig nie and spent several years in Cochinchina.

Table 10 continued from previous page

	Sentence	Translated
Simple	Некоторое время он провел во французском флоте. В 1889 и 1890 годах он служил в команде фрегата Iphig nie.	He spent some time in the French Navy. In 1889 and 1890 he served in command of the frigate Iphig nie.
	Simpitiki Wiki (Italian)	
Original	Mesero (Mésar nella variante locale del dialetto milanese) è un comune di 3.716 abitanti della provincia di Milano.	Mesero (Mésar in the local variant of the Milanese dialect) is a town of 3,716 inhabitants in the province of Milan.
Simple	Mesero (Mésar nel locale dialetto milanese) è un comune di 3.716 abitanti della provincia di Milano.	Mesero (Mésar in the local Milanese dialect) is a town of 3,716 inhabitants in the province of Milan.
	Simplext (Spanish)	
Original	Oxfam señaló que las bajas temperaturas de este invierno han aumentado el número de infecciones respiratorias , como la gripe y la neumonía , con más de 200.000 casos notificados en la segunda semana de este mes de enero , y grandes extensiones de tierra de el sur de Pakistán continúan bajo el agua contaminada .	Oxfam said this winter's low temperatures have increased the number of respiratory infections, including influenza and pneumonia, with more than 200,000 cases reported in the second week of January, across large swaths of southern Pakistan. they continue under polluted water.
Simple	Debido a el frío de el invierno , las enfermedades han aumentado entre las personas de Pakistán.	Due to the cold of winter, diseases have increased among the people of Pakistan.
	SimplifyUR (Urdu)	
Original	اسي بولني مں دشواری ہو رہی تہی	He was having trouble speaking
Simple	اسي بولني مں مشکل ہو رہی تہی	He was having difficulty speaking
	SloTS (Slovene)	
Original	Komaj sta bila v stolpu, je Hubert priskočil in kmetico udaril v obraz. Nato jo je še sunil v trebuh s svojim težkim škornjem.	As soon as they were in the tower, Hubert jumped up and punched the peasant in the face. Then he pushed her in the stomach with his heavy boot.
Simple	Ko jo je Hubert pripeljal v stolp, jo je udaril v obraz in brcnil v trebuh.	When Hubert brought her to the tower, he punched her in the face and kicked her in the stomach.
	Teacher (Italian)	
Original	Sebbene sia umido, credo che ad Amsterdam non abbiamo mai costruito niente di più comodo per chi ha bisogno di nascondersi.	Although it is humid, I believe that in Amsterdam we have never built anything more comfortable for those who need to hide.
Simple	E' umido ma è comodo come nascondiglio.	It's humid but it's comfortable as a hiding place.
	Terence (Italian)	
Original	Tutti si precipitarono verso il tendone e si ammassarono dentro per trovare riparo, perché nessuno si voleva infradiciare.	Everyone rushed to the tent and crowded inside for shelter, because no one wanted to get soaked.
Simple	Tutti si misero a correre verso la tenda, e ben presto la tenda fu piena di gente, perché nessuno si voleva bagnare.	Everyone ran towards the tent, and soon the tent was full of people, because nobody wanted to get wet.
	TextComplexityDE (German)	
Original	Die Geschichte der Europäischen Union ist durch ein Geflecht konkurrierender Motive und Entwicklungstendenzen charakterisiert, die zu unterschiedlichen Zeitpunkten jeweils richtungsgebend auf die Entwicklung der Gemeinschaft eingewirkt haben.	The history of the European Union is characterized by a web of competing motives and development tendencies, each of which has had a directional impact on the development of the community at different points in time.
Simple	Die Geschichte der Europäischen Union ist durch große Unterschiede von Motiven und Entwicklungen gekennzeichnet. Zu unterschiedlichen Zeitpunkten haben diese Unterschiede auf die Entwicklung der Gesellschaft Einfluss gehabt.	The history of the European Union is marked by great differences in motives and developments. At different points in time, these differences have had an impact on the development of society.

Table 10 continued from previous page

	Sentence	Translated
	WikiAuto (English)	
Original	The news of Kalākaua's death did not reach Hawaii until January 29 when the "Charleston" returned to Honolulu with the king's remains.	
Simple	Kalākaua's remains were sent to Honolulu aboard the American cruiser USS "Charleston".	
	WikiLargeFR (French)	
Original	La couleur du corps varie du brun moyen au doré à blanc beige et, à l'occasion, elle est marquée de taches brun foncé, surtout sur les membres.	Body color ranges from medium brown to golden to tan-white, and occasionally marked with dark brown spots, especially on the limbs.
Simple	La couleur du corps varie de brun moyen à doré à blanc beige et parfois marquée de taches brun foncé.	Body color ranges from medium brown to golden to tan-white and sometimes marked with dark brown spots.

Table 10: Example sentences sampled from all of the datasets in the MULTISIM benchmark

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 9; Page 9

☐ A2. Did you discuss any potential risks of your work?
Not applicable. The primary contribution of our work is a benchmark for multilingual text simplification consisting of parallel complex-simple sentences. We provide a collection of existing resources for text simplification and empirical analysis of their utility in fine-tuning and few-shot experiments. We do not release any models, and our work is a collection of existing datasets. In effect, we are not releasing any potentially harmful models or resources that merit a discussion of risk.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1; Page 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3; Pages 2-4; Section 4; Pages 4-5

☑ B1. Did you cite the creators of artifacts you used?
Section 3; Table 1; Page 3

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 3; Table 1; Page 3 and Section 4.1; Page 4

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Appendix A; Pages 15-17

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Our data is a collection of simple-complex sentence pairs drawing primarily from wikipedia, news articles, and similar informational domains. None of our data has personally identifiable information and none of it was drawn from social media or similar user-driven platforms. Also as our data is informational in nature (ie. wikipedia, news, public policy) it was not drawn from sources of offensive content.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3; Table 1; Page 3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3; Table 1; Page 3 and Section 4; Table 2; Page 4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Section 5; Pages 5-6; Section 6; Pages 6-8

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix C; Pages 21-22

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5; Pages 5-6 and Appendix C; Pages 21-22

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 6; Pages 6-8 and Section 7; Page 8

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 5; Page 5 and Appendix B; Page 17-18

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 7; Page 8

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix C; Table 9; Page 24

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 7; Page 8 and Appendix C; Page 23

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix C; Page 23

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No. This study does not meet the definition of "human subjects" research by our institutional IRB review board, as we did not: "(1) Interact with, intervene with, or obtain/access private, identifiable information or data about, a living individual (includes online surveys)? or (2) Conduct research on a drug, biologic, or medical device?". The annotations are linguistic judgements provided by volunteer in-house employees and colleagues.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix C; Page 23