OasisSimp: An Open-source Asian-English Sentence Simplification Dataset

Hannah Liu[1*], Muxin Tian[1*], Iqra Ali[2], Haonan Gao[3], Qiaoyiwen Wu[1], Blair Yang[4], Uthayasanker Thayasivam[5], En-Shiun Annie Lee[6,1], Pakawat Nakwijit[2], Surangika Ranathunga[7], Ravi Shekhar[8]

[1]University of Toronto, [2]Queen Mary University of London, [3]Yale University, [4]CoolWei AI Lab
[5]University of Moratuwa, [6]Ontario Tech University, [7]Massey University, [8]University of Essex
{hannahhere.liu, murphy.tian, qiaoyiwen.wu}@mail.utoronto.ca,
{iqra.ali, p.nakwijit}@qmul.ac.uk, eric.gao@yale.edu, yangliwei@coolwei.com,
rtuthaya@cse.mrt.ac.lk, annie.lee@ontariotechu.ca,
s.ranathunga@massey.ac.nz, r.shekhar@essex.ac.uk

Abstract

Sentence simplification aims to make complex text more accessible by reducing linguistic complexity while preserving the original meaning. However, progress in this area remains limited for mid-resource and low-resource languages due to the scarcity of high-quality data. To address this gap, we introduce the *OasisSimp* dataset, a multilingual dataset for sentence-level simplification covering five languages: English, Sinhala, Tamil, Pashto, and Thai. Among these, no prior sentence simplification datasets exist for Thai, Pashto, and Tamil, while limited data is available for Sinhala. Each language simplification dataset was created by trained annotators who followed detailed guidelines to simplify sentences while maintaining meaning, fluency, and grammatical correctness. We evaluate eight open-weight multilingual Large Language Models (LLMs) on the OasisSimp dataset and observe substantial performance disparities between high-resource and low-resource languages, highlighting the simplification challenges in multilingual settings. The OasisSimp dataset thus provides both a valuable multilingual resource and a challenging benchmark, revealing the limitations of current LLM-based simplification methods and paving the way for future research in low-resource sentence simplification. The dataset is available at https://OasisSimpDataset.github.io/.

Keywords: Sentence simplification, low-resource languages, LLM

1. Introduction

Reading is a critical daily activity in our text-rich society, yet it can be challenging for some due to complex words, sentences, or meanings. Organizations such as the Plain Language Association International (PLAIN)[1] and the Easy-to-Read movement (de Normalización, 2018) emphasize using simple texts to make information accessible to all. Ensuring written content is clear and readable is crucial for equitable knowledge access.

Sentence simplification is the task of transforming complex text into an easier-to-understand format by reducing its linguistic complexity while preserving the original meaning and factual information. This process is essential for enhancing information accessibility for a wide range of individuals, including language learners, people with cognitive or reading disabilities such as dyslexia or aphasia, and younger audiences (Paetzold and Specia, 2017; Espinosa-Zaragoza et al., 2023; Al-Thanyyan and Azmi, 2021). High-quality simplified text can also be a valuable preprocessing step for various downstream Natural Language Processing (NLP) tasks, including Machine Translation, text summarization, and question answering (Dadu et al., 2021).

In the last two decades, the NLP community has made significant progress in addressing this challenge by developing various datasets and tools for sentence simplification. Early efforts primarily focused on high-resource languages such as English (Shardlow, 2014; Alva-Manchego et al., 2020b) and Spanish (Ferrés and Saggion, 2022). In particular, the availability of English datasets such as Newsela (Xu et al., 2015) and WikiLarge (Zhang and Lapata, 2017) has driven substantial advances in simplification model development. However, this focus has created a pronounced resource imbalance, leaving many other languages, especially low-resource languages such as Sinhala and Pashto, significantly underserved. The scarcity of high-quality, parallel complex-simple sentence pairs for these languages presents a major obstacle to developing, evaluating, and advancing simplification technologies across linguistic boundaries.

Recent efforts have sought to expand language coverage. For example, Ryan et al. (2023) introduced the MULTISIM benchmark, covering 12 languages using data from 27 sources. Similarly, multilingual datasets for domain-specific applications, particularly in the medical domain (Basu et al., 2023), have emerged, includ-

*These authors contributed equally.
[1]https://plainlanguagenetwork.org/

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6103–6115
11–16 May 2026.

ing MultiCochrane (Joseph et al., 2023) and MultiMSD (Horiguchi et al., 2025). However, many of these datasets are domain-specific and offer limited general language representation.

We introduce the *OasisSimp* dataset (An Open-source Asian-English Sentence Simplification Dataset), a publicly available dataset for sentence-level simplification to address this critical resource gap. OasisSimp covers five languages - English, Sinhala, Tamil, Thai, and Pashto - with all parallel sentence pairs created through expert human simplification. Trained annotators followed consistent, detailed guidelines to manually rewrite complex sentences into simpler forms, ensuring that the core meaning, grammatical correctness, and fluency were preserved. Beyond releasing this resource, we leverage OasisSimp to evaluate LLM-based text simplification approaches proposed in Kew et al. (2023). We report performance of various various LLMs in zero- and few-shot settings and report SARI (Xu et al., 2016) and BERTScore (Zhang et al., 2020).

By providing both the *OasisSimp* dataset and associated benchmarking results, we aim to expand the resources available for multilingual sentence simplification research significantly. We anticipate that this work will stimulate further research, support the development of more inclusive and effective simplification systems, and ultimately make information more accessible to diverse global audiences.

2. Related Work

Sentence Simplification makes written language easier to read and understand while keeping the original meaning intact. It plays an important role in improving the readability and accessibility of text for diverse audiences. Although research in sentence simplification has a long and rich history, it has mainly focused on high-resource languages, particularly English. Simplification generally occurs at three levels: lexical, syntactic, and conceptual or discourse. Lexical simplification involves replacing complex words with simpler alternatives (Venugopal et al., 2022), while syntactic simplification reduces grammatical complexity to make sentences clearer and easier to follow (Shardlow, 2014; Alva-Manchego et al., 2020b). On the other hand, conceptual simplification aims to make the underlying ideas easier to understand, not just the words or sentence structures used to express them (Eschenbrücher, 2021). Most studies have concentrated on lexical and syntactic simplification, and this work focuses explicitly on sentence-level syntactic simplification.

Sentence Simplification Datasets Early and influential datasets for English sentence simplification are Simple English Wikipedia (Coster and Kauchak, 2011), which was constructed by automatically aligning complex sentences with their simplified counterparts, with previous research on automatically constructing parallel data (Kajiwara and Komachi, 2016). Other prominent English simplification datasets include ASSET (Alva-Manchego et al., 2020a), developed through crowdsourcing, and WikiLarge (Zhang and Lapata, 2017), created by combining multiple existing resources. These corpora have been instrumental in advancing the field from early rule-based and statistical methods to more recent neural approaches to sentence simplification (e.g., Nisioi et al., 2017; Martin et al., 2020).

In recent years, research in sentence simplification has gradually expanded beyond English to include other languages. For Italian, Tonelli et al. (2016) and Brunato et al. (2016) introduced sentence-level simplification datasets, with the former focusing on manual simplification and the latter employing automatic methods. For Danish, the DSim corpus (Klerke and Søgaard, 2012) was created using journalists producing simplifications, while Grabar and Cardon (2018) developed a French medical simplification dataset combining crowdsourced and expert annotations. For Russian, the RuSimpleSentEval-2021 Shared Task (Sakhovskiy et al., 2021) provided a crowdsourced dataset with both automatic and manual sentence alignments. A notable advancement in multilingual simplification is the MultiSim benchmark (Saggion et al., 2022), which includes English, Spanish, and Portuguese data. Although this benchmark marked an important step towards multilingual evaluation, it remains centered on high-resource languages. Similarly, Ryan et al. (2023) proposed a non-English benchmark by combining existing resources, but coverage for genuinely low-resource languages remains limited.

The availability of sentence simplification corpora for low-resource languages is still scarce. The SiTSE corpus (Ranathunga et al., 2025) includes 1,000 complex Sinhala sentences simplified by three annotators, yielding 3,000 complex–simple pairs. Likewise, Mondal et al. (2024) created Bengali and Marathi datasets, each containing 500 pairs, while Anees and Abdul-Rauf (2021) developed an Urdu corpus comprising 610 sentences simplified by two annotators. Although these initiatives represent valuable progress, they are relatively small compared to English datasets. To address this gap, our work introduces the OasisSimp dataset, a new multilingual sentence simplification dataset covering mid- and low-resource languages.

Simplification Methodology Supervised learning has long been the dominant paradigm in modern sentence simplification (Al-Thanyyan and Azmi, 2021; Alva-Manchego et al., 2020c). However, supervised approaches are often impractical for low-resource languages, where parallel simplification training data is limited or unavailable. To address this issue, researchers have often turned to machine translation (MT) to create pseudo data, for example, by translating complex English sentences into another language and simplifying them, or by translating simplified English sentences into the target language (Palmero Aprosio et al., 2019; Lu et al., 2021; Sheang and Saggion, 2021; Ki and Carpuat, 2025). Although these methods are creative, they can introduce translation noise and fail to capture language-specific nuances of simplification (Hasan et al., 2021). Further improvements were achieved by mining paraphrases from large-scale web-crawled data (Martin et al., 2022). More recently, the emergence of LLMs has enabled simplification through zero-shot and few-shot prompting, reducing reliance on explicit training data. In this work, we adopt an LLM-based simplification approach, leveraging the strong generalization capabilities of multilingual LLMs to handle sentence simplification across diverse languages.

3. Data Creation

3.1. Language Selection

We selected five diverse languages (English, Thai, Tamil, Sinhala, and Pashto)[2] to create a sentence simplification dataset, based on their linguistic, typological, and sociolinguistic diversity. English, a high-resource Indo-European language, provides well-studied benchmarks for comparison. Thai, a tonal Southeast Asian language with a script that lacks spaces, Tamil, a morphologically rich Dravidian language, Sinhala, an Indo-Aryan language with complex syllable structures, and Pashto, an Eastern Iranian language with dialectal variations, are low-resource languages with diverse scripts and grammatical features. This selection covers multiple language families, scripts (Latin, Thai, Brahmi-derived, Arabic-derived), and linguistic characteristics (analytic vs. agglutinative, tonal vs. non-tonal). Apart from that, Thai, Tamil, and Pashto have no sentence simplification datasets. For Sinhala, only a small dataset exists (Ranathunga et al., 2025). Including these languages addresses a significant gap in sentence simplification research. Additionally, all five languages are spoken by sizable populations, meaning that simplification can have a real social impact by improving literacy, education, and access to information. The resulting dataset will support multilingual research and practical applications across languages with very different linguistic and writing systems.

[2]Based on data availability, English: 5, Thai & Tamil: 3, Sinhala & Pashto: 2 (Joshi et al., 2020; Ranathunga and de Silva, 2022), where 5 is high-resources and 1 is low-resource language.

3.2. Data Source and Selection

Due to the lack of a common data source across all five languages, we individually selected sources for each language, ensuring they are publicly available and cover diverse domains such as Wikipedia, newspapers, and government documents. Our goal was to collect sentences with varying lengths and complexity. We applied automatic filtering criteria on the selected source data, including a minimum sentence length. We also ensured coverage across different topics. For example, Wikipedia articles on animals, products, buildings, etc., and newspaper articles on sports, politics, entertainment, and more. In addition to these general steps, we applied language-specific filtering to account for linguistic characteristics unique to each language (see below). This approach provides a wide range of sentence complexity. Finally, after automatically identifying a large set of potential complex sentences, we manually reviewed them to select a number for simplification. Manual inspection ensured the chosen sentences were genuinely complex and suitable for simplification.

English (*OasisSimp-EN*) The complex English sentences were drawn from Canadian general-interest newspapers via the NewsEdits curation workflow, primarily *The Globe and Mail*. We sampled across general-interest sections (e.g., *News*, *Report on Business*, *Opinion*, *Arts & Life/Books*) and excluded templated material such as stock tables, listings, photo captions, corrections, and headlines. Candidate sentences were restricted to 100–300 characters and excluded if they contained special characters (e.g., “@”, “&”). We also limited proper nouns to reduce world-knowledge dependency and annotation ambiguity, thereby encouraging structural/lexical simplification over entity substitution. To maximize sentence-level simplifiability, we prioritized sentences whose complexity stems from syntax (e.g., coordination and subordination, apposition, heavy noun phrases) and lexical density rather than from named-entity load. This filtering produced a large pool of candidate complex sentences, from which we manually curated 2,500 context-independent complex sentences for simplification to form the *OasisSimp-EN*.

Sinhala (*OasisSimp-SI*) & Tamil (*OasisSimp-TA*) Sinhala and Tamil complex sentences were selected from the SiTa trilingual parallel corpus (Fernando et al., 2020; Ranathunga et al., 2018). This corpus has about 50k unique sentence pairs coming from official government documents, meticulously cleaned and fixed by language experts. These government documents are more complex than other sources from Sri Lanka, such as news or social media text. Government documents often feature longer sentences, domain-specific terminology from fields such as accounting and law, and a highly formal register (Ranathunga et al., 2025). We opted to select sentences containing rare words. Initially, rare word based filtering was done for Sinhala (we selected rare words that have a frequency of less than 5-50[3] in 140k sentences of the common crawl corpus). This resulted in 5859 sentences. Manual observation revealed that some sentences are just lists of items, and some are near duplicates. Therefore, these sentences were manually filtered. Finally, we selected 2500 long sentences. In certain cases, there were technical terms, and within brackets, the English meaning was given. These English terms were removed. A random 500 were selected out of these sentences, and the corresponding Tamil sentences were retrieved from the SiTa trilingual corpus. The Tamil dataset had to be restricted to 520 due to the unavailability of human annotators. Each complex sentence was simplified by 5 annotators, for both Sinhala and Tamil, to form the *OasisSimp-SI* and *OasisSimp-TA* datasets.

Thai (*OasisSimp-TH*) We used the ThaiSum dataset as our initial corpus (Chumpolsathien, 2020). This dataset was collected from Thai news websites: Thairath, ThaiPBS, Prachatai, and The Standard. A Thai-specific CRF segmentator trained on the TED dataset (Phatthiyaphaibun et al., 2024) pre-segmented each document into sentences. We first randomly sampled sentences from the corpus and performed a filtering based on the rare words, length, and topic diversity. We selected sentences containing rare and non-rare words to avoid rare word bias. All sentences were also subjected to a length constraint, with only those more than 15 words being retained. To ensure topic diversity, we re-sampled the pre-filtered sentences from 10 distinct news categories: Politics, Local News, Economics, Society, Foreign Affairs, Quality of Life, Human Rights, Lifestyle, Culture, and Education, resulting in a pool of 3,815 sentences. In addition, due to the ambiguity of sentence and word boundaries and the limitations of the automatic sentence segmentation, one of the authors manually picked the correct sentence boundary. We finally obtained 1499 Thai sentences to form the *OasisSimp-TH* dataset.

Pashto (*OasisSimp-PS*) We used Wikipedia as our primary source for collecting complex Pashto sentences. Wikipedia is a reliable, publicly available, and diverse domain resource for Pashto, providing naturally occurring complex sentences with authentic vocabulary, syntax, and named entities. We initially collected 10,000 Pashto sentences, distributed equally across ten semantic categories like animals, products, buildings, locations/places, events, food, drinks, hobbies, works of art, and organizations *(1,000 sentences per category)*. Human annotators manually extracted these sentences from Wikipedia to ensure relevance and domain coverage within each category. Out of the 1,000 sentences per category, the top *250* most complex sentences were selected based on four key points: sentence length, syntactic complexity, vocabulary richness, and semantic depth. Longer (25/30 words), multi-clause sentences with structured and diverse vocabulary are preferred, ensuring linguistic richness and contextual nuance. This filtering resulted in a final set of *2,500* sentences and formed the *OasisSimp-PS* dataset, ensuring sufficient complexity for evaluating simplification systems. All selected sentences were reviewed by a native Pashto linguistic expert to ensure their cultural and linguistic appropriateness before annotation. The evaluation mainly depends on cultural relevance, naturalness, grammatical correctness, vocabulary suitability, and semantic clarity.

Table 1: Final statistics for each language in the *OasisSimp* dataset. Comp - complex, Simp - simplified.

Lang	# Comp Sentences	Avg. Simp Sentences	Avg.Comp Length	Avg. Simp Length	Source Domain
English	2500	2.86	24.35	17.23	News
Sinhala	2500	5.00	30.12	28.78	Govt
Thai	1499	5.06	48.24	37.77	News
Tamil	520	4.66	23.22	17.65	Govt
Pashto	2500	3.00	28.81	20.31	Wiki

3.3. Data Annotation Workflow

We used 3-6 native language speakers with at least Bachelor's degrees for annotation. We recruited the annotators based on our contacts from the respective countries to have greater quality control over the annotation process, instead of going for an online platform. We instructed annotators to simplify sentences while preserving meaning, fluency, and grammaticality. We provided detailed instructions, adapted from Xu et al. (2016); Alva-Manchego et al. (2020a). Before annotation, anno-

[3]words that had a frequency less than 5 were mostly misspelled words.

Table 2: Sample Examples from the *OasisSimp* Dataset for each language.

Complex	Simplified
English: Alarmed at the escalating bill for weather-related disaster relief, Public Safety Minister Ralph Goodale is in talks with insurers and the provinces.	Public Safety Minister Ralph Goodale is discussing with insurers and provinces about the rising costs of disaster relief. **Operation:** Rewording
Pashto: د نومړي ولسوالۍ د کرنې مخکې لږې دي، خو ډېرې کرهڼيزې دي. د دغې ولسوالۍ په اهمو کښتونو کې جوار، غنم، اوربشې، کچالو او وريجې راځي. **Translation:** The agricultural land in the district is small but very fertile. Major crops include maize, wheat, barley, potatoes, and rice.	په ناماره ولسوالۍ کې لږې کرنيزې ځمکې شته، چې د دې ولسوالۍ عمده حاصلات جوار، غنم، وربشې، کچالو او وريجې دي. **Translation:** There is little agricultural land in Namara district; major crops include maize, wheat, barley, potatoes, and rice. **Operation:** Deletion
Sinhala: අල්ලස් පනත යටතේ සහ මැතිවරණ නීති යටතේ ඡන්ද විමසීමවල දී සිදු කළ වරදවල් සහ වනෙත් හේතු මත සිර දඬුවම් විඳීම හේතුවෙන් 2015 වර්ෂයේ දී මැතිවරණ දෙපාර්තමේන්තුව වෙත වාර්තා වූ පුද්ගලයන් දෙදෙනෙකුගේ නම් ඡන්දහිමි නාමලේඛනයෙන් ඉවත් කිරීමට පියවර ගැනිණි. **Translation:** Steps were taken to remove the names of two individuals from the voters' register, who were reported to the Elections Department in 2015 due to imprisonment under the Bribery Act and the Election Laws for offences committed during elections and other reasons.	අල්ලස් පනත හා මැතිවරණ නීති අනුව ඡන්දවල දී කළ වැරදි හා වනෙත් හේතු මත සිරගත කිරීම් ගැන 2015 දී මැතිවරණ දෙපාර්තමේන්තුවට වාර්තා වූ මිනිසුන් දෙදෙනෙකු ඡන්දහිමි නම් ලැයිස්තුවෙන් ඉවත් කරෙණි. **Translation:** Two people, who were reported to the Elections Department in 2015 due to imprisonment according to the Bribery Act and Election Laws for offences during elections and other reasons, were removed from the voter list. **Operation:** Deletion, rewording
Thai: ขณะที่ ถนนดินแดง จาก แยกใต้ด่วนดินแดง มุ่งหน้าแยกดินแดง รถมากเคลื่อนตัวได้ช้า ท้ายหยุดนิ่งอยู่ที่แยกใต้ด่วนดินแดง, จาก แยกดินแดง มุ่งหน้า แยกวงเวียนอนุสาวรีย์ชัยฯ รถมากเคลื่อนตัวได้ช้า **Translation:** Meanwhile, on Din Daeng Road, from the Din Daeng Expressway Intersection heading towards the Din Daeng Intersection, there is heavy traffic moving slowly, with the tail end of the queue stationary at the expressway intersection. From the Din Daeng Intersection heading towards the Victory Monument roundabout, traffic is also heavy and moving slowly.	บริเวณถนนดินแดงมีรถติดหนักมาก ทั้งฝั่งที่มุ่งหน้าไปแยกดินแดง และ แยกวงเวียนอนุสาวรีย์ชัยฯ **Translation:** There is very heavy traffic on Din Daeng Road in both directions towards the Din Daeng Intersection and the Victory Monument roundabout. **Operation:** Reordering, Deletion
Tamil: மிகவும் கண்கவர் பிரதேசத்தில் அமையப்பெற்றுள்ள இந்த பயிற்சி நிலையத்தில் ஒரே தடவையில் 250 பேர் வரை பங்கு கொள்ளக் கூடிய மண்டபமும் குழுக் கலந்துரையாடலுக்கான வசதியுடன் 60 பேருக்கான தங்குமிட வசதியும் மற்றும் நூலகமும் அமையப்பெற்றதாகும். **Translation:** Located in a very attractive location, this training center has a hall that can accommodate up to 250 people at a time, with group discussion facilities, accommodation for 60 people and a library are available.	கண்கவர் பிரதேசத்தில் உள்ள இந்த பயிற்சி நிலையத்தில் 250 பேர் வரை பங்கு கொள்ளக் கூடிய மண்டபமும் 60 பேருக்கான தங்குமிட வசதியும் மற்றும் நூலகமும் அமையப்பெற்றுள்ளது. **Translation:** Located in an attractive location, in this training center, a hall where 250 people can participate, accommodation facilities for 60 people and library are available. **Operation:** Deletion, rewording

tators received structured training sessions to familiarize them with the simplification guidelines, operational definitions, and annotation criteria. During the training phase, annotators were encouraged to ask questions to ensure they understood how to retain the core meaning while making sentences easier for lower-literacy speakers. We provided at least 3 rounds of training to each annotator. During training, annotators were provided with illustrative examples to demonstrate the following key simplification operations.

- Replace rare or technical terms with simpler synonyms *(Rewording)*.
- Split long sentences into multiple shorter sentences *(Sentence Splitting)*.
- Remove unnecessary details while preserving meaning *(Deletion)*.
- Restructure sentences for clarity *(Reordering)*.

The annotations were performed in batches of 25-100 sentences, depending on language, com-

plexity, and the annotator's capacity. Initially, one of the authors verified that the annotators followed the instructions. In case of any issue, we provided further instructions to the annotator. After a few rounds of building confidence with the annotator, we randomly sampled some simplified text for verification. In Table 2, we present one example from each language with the corresponding transaction and operation performed.

Table 1 presents the final statistics for each language in the OasisSimp dataset, including the number of complex sentences, the average number of simplified sentences per complex instance, the average sentence lengths for both complex and simplified versions, and their respective sources. The sentence length for Thai is measured by tokens, while for other languages it's measured by the number of words. In most languages, each complex sentence is paired with three simplified versions, except for a few English sentences that were excluded due to quality issues. As shown across all languages, the average length of simplified sentences is consistently shorter than that of complex sentences, indicating that simplification often involves the removal of redundant or less essential information. We divided the dataset into validation and test subsets to support unsupervised model development, allocating 80% of the data to testing and the remaining 20% to validation.

4. Methodology and Evaluation

Our work focuses on advancing less-resource sentence simplification datasets, rather than proposing a new simplification methodology. To this end, we adopt an existing LLM-based simplification framework, following the prompting strategy proposed by Kew et al. (2023), where the model is instructed to "Simplify the given sentence …" to generate simplified outputs. This approach allows us to systematically evaluate the capabilities of different multilingual models under a consistent simplification paradigm, isolating the effect of model design and linguistic coverage on performance.

We evaluated a diverse set of open-weight, multilingual LLMs, including Aya (Aya-Expanse-8B), Cmd-R (c4ai-command-r7b-12-2024), DeepSeek (deepseek-llm-7B-chat), EuroLLM (EuroLLM-9B-Instruct), Gemma (Gemma-3-12B-it), LLaMA (Llama-3.2-3B-Instruct), Mistral (Mistral-7B-Instruct-v0.2), and Qwen (Qwen2.5-7B-Instruct). These models encompass a range of architectural designs, parameter scales, and training paradigms. While not all models explicitly disclose the full list of languages in their training data, we broadly categorize them based on whether a language is known to be included, excluded, or likely included based on an educated guess (e.g., Gemma). We also include models such as EuroLLM, trained primarily on European languages, to examine their ability to generalize to unseen or low-resource languages. Based on the available language information, all models include English. LLaMA includes Tamil and Thai, while Gemma is most likely trained on all evaluated languages except Pashto.

We conduct experiments in both zero-shot and few-shot settings, building upon BLESS (Kew et al., 2023). For zero-shot, we experimented with multiple temperatures (0.1 - 0.9 in steps of 0.1) and report the best performance. In the few-shot configuration, we used 1-shot and 5-shot settings using the best temperature from zero-shot. This design enables a systematic comparison of LLMs with varying capacities and linguistic generalization abilities in the context of less-resource sentence simplification. We used the two standard automatic simplification metrics for the evaluation: SARI (Xu et al., 2016) and BERTScore (Zhang et al., 2020). SARI measures the goodness of words added, deleted, and kept by the simplification system, while BERTScore measures the semantic similarity of the simplified sentence with the reference sentences[4].

5. Results and Discussion

In Table 3, 4, 5, 6, 7, we present Zero, 1, and 5-shot SARI and BERTScore ($\mathbf{F}_{\text{ref}}$) using multiple LLMs from English, Sinhala, Thai, Tamil, and Pashto, respectively.

Zero-Shot Performance: The zero-shot evaluation shows clear differences in how language and model strength affect multilingual sentence simplification. As a high-resource language, English BERTScore has less variation. This reflects the strong baseline ability of LLMs in languages that match their training data. In contrast, low- and mid-resource languages like Pashto and Thai show the largest variation in the BERTScore across models, which suggests that zero-shot is not effective for all LLMs. However, the Gemma model performs surprisingly well, achieving high BERTScore scores in Sinhala (66.77), Tamil (74.55), and Pashto (56.95). This indicates that these models transfer knowledge across languages more effectively and follow instructions better. DeepSeek and Llama's negative scores for Tamil and Pashto, respectively, highlight that some models struggle to generalize in low-resource settings. Further analysis of SARI and its components (ADD, KEEP, DEL) gives more

[4]We didn't use BLEU Score because of its unsuitability (Callison-Burch et al., 2006; Reiter, 2018). However, for completeness, we included BLEU results in Appendix.

Table 3: **Results on English (*OasisSimp-EN*) dataset**: Zero, 1- and 5-shot Results. Highest value in **bold** and lowest underlined. ✔: Language included in LLM training; ✔: Likely included (Educated Guess); ✗: Not included.

Model	0 Shot					1 Shot					5 Shot				
	SARI Comp.			SARI	F_{ref}	SARI Comp.			SARI	F_{ref}	SARI Comp.			SARI	F_{ref}
	ADD	KEEP	DEL			ADD	KEEP	DEL			ADD	KEEP	DEL		
Aya ✔	9.32	44.98	75.23	43.18	54.44	9.68	44.90	72.51	42.36	56.35	10.18	45.91	71.16	42.42	57.20
Cmd-R ✔	**9.69**	44.95	72.89	42.51	55.90	**10.99**	43.71	77.57	44.09	55.03	**11.91**	45.28	77.09	44.76	56.63
DeepSeek ✔	7.03	41.47	76.30	41.60	51.88	7.80	41.12	76.82	41.91	51.92	9.41	42.03	77.22	42.89	54.15
EuroLLM ✔	9.32	45.60	68.36	41.10	**56.98**	**10.99**	**46.98**	69.35	42.44	**57.96**	11.63	46.55	70.93	43.04	**58.10**
Gemma ✔	5.24	44.43	68.54	39.40	51.87	6.55	43.26	74.44	41.41	52.34	9.19	44.67	77.06	43.64	55.27
LLaMA	6.48	43.31	68.34	39.38	54.30	8.11	43.42	72.83	41.45	54.53	9.93	44.75	73.75	42.81	56.00
Mistral ✔	8.56	43.66	**77.46**	**43.23**	52.49	10.31	43.82	**78.43**	44.18	54.55	11.61	44.01	**78.59**	44.74	55.89
Qwen ✔	8.70	**46.07**	73.53	42.77	42.36	9.54	46.40	77.25	**44.39**	53.03	10.88	**47.01**	77.08	**44.99**	55.27

Table 4: **Results on Sinhala (*OasisSimp-SI*) dataset**: Zero, 1- and 5-shot Results. Highest value in **bold** and lowest in underlined. ✔: Language included in LLM training; ✔: Likely included (Educated Guess); ✗: Not included.

Model	0 Shot					1 Shot					5 Shot				
	SARI Comp.			SARI	F_{ref}	SARI Comp.			SARI	F_{ref}	SARI Comp.			SARI	F_{ref}
	ADD	KEEP	DEL			ADD	KEEP	DEL			ADD	KEEP	DEL		
Aya ✗	0.47	23.11	68.49	30.69	54.94	0.40	26.01	67.39	31.27	54.22	0.21	27.62	65.75	31.20	44.78
Cmd-R ✗	0.18	**29.95**	65.05	31.73	58.32	0.19	30.24	65.08	31.84	55.94	0.13	28.87	65.18	31.40	45.34
DeepSeek ✗	0.16	16.92	68.71	28.60	-0.14	0.14	24.99	66.71	30.61	57.68	0.08	20.23	67.60	29.30	38.55
EuroLLM ✗	0.03	19.19	68.14	29.12	47.59	0.04	18.74	68.18	28.99	47.56	0.00	2.03	**70.18**	24.07	-60.47
Gemma ✔	**3.89**	28.64	**71.77**	**34.77**	**66.77**	**4.62**	**36.89**	**71.25**	**37.59**	**70.44**	**5.65**	**44.21**	69.82	**39.89**	**73.89**
LLaMA ✗	0.32	17.77	69.25	29.11	47.10	0.31	19.23	68.91	29.49	45.49	0.30	20.49	68.18	29.66	32.99
Mistral ✗	0.19	11.61	70.00	27.26	42.84	0.22	19.58	68.60	29.47	54.85	0.17	25.28	66.80	30.75	58.47
Qwen ✗	0.46	28.70	66.27	31.81	59.50	0.43	29.17	65.84	31.81	57.44	0.38	29.01	65.66	31.68	46.73

Table 5: **Results on Thai (*OasisSimp-TH*) dataset**: Zero, 1- and 5-shot Results. Highest value in **bold** and lowest in underlined. ✔: Language included in LLM training; ✔: Likely included (Educated Guess); ✗: Not included.

Model	0 Shot					1 Shot					5 Shot				
	SARI Comp.			SARI	F_{ref}	SARI Comp.			SARI	F_{ref}	SARI Comp.			SARI	F_{ref}
	ADD	KEEP	DEL			ADD	KEEP	DEL			ADD	KEEP	DEL		
Aya ✗	0.33	21.01	84.30	35.22	56.49	0.38	24.45	83.53	36.12	59.01	0.43	25.98	82.59	36.33	60.66
Cmd-R ✗	0.47	29.02	82.19	37.23	**62.17**	0.51	26.88	81.85	36.41	61.84	0.45	28.89	81.86	37.07	62.26
DeepSeek ✗	0.12	15.27	84.34	33.24	27.98	0.18	15.80	83.95	33.31	51.68	0.14	15.77	84.09	33.34	40.95
EuroLLM ✗	0.15	26.35	80.43	35.64	52.25	0.23	26.26	80.23	35.57	60.26	0.16	28.05	77.71	35.31	62.57
Gemma ✔	**0.66**	27.78	**85.23**	37.89	38.99	**1.30**	**34.25**	**85.31**	**40.29**	**65.22**	**1.48**	37.53	**84.82**	**41.28**	67.30
LLaMA ✔	0.25	24.83	83.50	36.19	8.85	1.07	33.08	82.13	38.76	63.97	1.43	**39.99**	79.26	40.23	**68.91**
Mistral ✗	0.22	23.59	84.55	36.12	43.13	0.44	23.94	83.65	36.01	58.03	0.60	24.72	82.97	36.09	61.02
Qwen ✗	0.57	**32.25**	84.30	**39.04**	58.20	0.76	33.78	84.11	39.55	60.40	1.06	36.60	84.24	40.64	64.27

insight into how models behave during zero-shot simplification. The English SARI score has less variation across models, compared to other languages, meaning the models align well with the original text structure for English compared to other languages. Among all SARI components, the DEL is the highest, showing that models focus on removing unnecessary or complex information. The ADD component is the lowest, especially in non-English languages, suggesting that models find it difficult to add new or helpful information. Overall, the findings show that while Gemma handles multiple languages well and is sensitive to text structure, low-resource languages still pose challenges to all

Table 6: **Results on Tamil (*OasisSimp-TA*) dataset**: Zero, 1- and 5-shot Results. Highest value in **bold** and lowest in underlined. ✔: Language included in LLM training; ✔: Likely included (Educated Guess); ✗: Not included.

Model	0 Shot					1 Shot					5 Shot				
	SARI Comp.			SARI	F_{ref}	SARI Comp.			SARI	F_{ref}	SARI Comp.			SARI	F_{ref}
	ADD	KEEP	DEL			ADD	KEEP	DEL			ADD	KEEP	DEL		
Aya ✗	2.42	18.48	70.60	30.50	72.42	2.25	27.49	68.73	32.82	74.40	2.00	34.27	66.62	34.30	76.24
Cmd-R ✗	2.78	26.40	69.28	**32.82**	74.14	2.07	30.78	66.61	33.15	74.97	1.44	34.40	66.15	34.00	76.08
DeepSeek ✗	0.33	11.88	70.83	27.68	-0.67	0.49	19.23	**69.54**	29.75	68.27	0.11	8.76	**70.56**	26.48	3.30
EuroLLM ✗	0.19	24.90	68.07	31.05	69.89	0.32	26.56	67.05	31.31	71.21	0.16	25.05	67.73	30.98	65.97
Gemma ✔	**4.22**	23.15	**71.07**	**32.82**	**74.55**	**5.17**	**35.26**	68.65	**36.36**	**77.01**	**5.16**	**47.11**	65.73	**39.34**	**79.70**
LLaMA ✔	0.75	19.17	69.88	29.93	9.71	1.03	28.81	66.64	32.16	64.14	0.51	28.46	67.42	32.13	40.07
Mistral ✗	0.57	16.71	70.29	29.19	27.62	1.19	28.34	67.35	32.29	72.23	0.89	32.65	66.40	33.31	75.04
Qwen ✗	1.76	**26.58**	69.32	32.55	73.53	1.64	29.61	68.07	33.11	74.81	1.36	33.94	66.22	33.84	76.03

Table 7: **Results on Pashto (*OasisSimp-PS*) dataset**: Zero, 1-, and 5-shot Results. Highest value in **bold** and lowest in underlined. ✔: Language included in LLM training; ✔: Likely included (Educated Guess); ✗: Not included.

Model	0 Shot					1 Shot					5 Shot				
	SARI Comp.			SARI	F_{ref}	SARI Comp.			SARI	F_{ref}	SARI Comp.			SARI	F_{ref}
	ADD	KEEP	DEL			ADD	KEEP	DEL			ADD	KEEP	DEL		
Aya ✗	0.62	23.98	67.47	30.69	49.17	1.08	45.60	58.62	35.10	60.83	1.77	53.81	47.17	34.25	68.25
Cmd-R ✗	0.75	50.82	51.73	34.44	61.91	0.93	54.41	44.44	33.26	67.84	0.70	**56.53**	35.62	30.95	**70.52**
DeepSeek ✗	0.52	41.19	60.71	34.14	38.65	0.90	48.83	54.59	34.78	63.65	0.91	50.16	52.51	34.53	66.26
EuroLLM ✗	0.50	**54.28**	44.40	33.06	**67.55**	0.65	**54.87**	43.28	32.93	**69.72**	0.78	55.37	42.09	32.75	70.42
Gemma ✗	**3.84**	25.08	**70.78**	33.23	56.95	**4.47**	34.75	**68.57**	35.93	61.47	**5.39**	46.39	**61.95**	**37.91**	66.04
LLaMA ✗	0.70	18.34	70.28	29.77	-22.40	3.15	46.28	61.67	**37.04**	51.15	1.96	46.53	58.15	35.55	33.03
Mistral ✗	0.94	26.36	68.13	31.81	47.73	1.42	41.20	63.04	35.22	61.31	1.51	45.93	58.60	35.35	64.40
Qwen ✗	2.34	47.48	58.92	**36.25**	58.02	2.81	49.88	55.34	36.01	64.76	2.62	53.79	48.71	35.04	65.57

LLMs.

Few-shot Performance The few-shot evaluation (1-shot and 5-shot) shows clear and consistent improvements compared to zero-shot performance across all languages and models, especially in low-resource settings. Even one example helps the models better understand the structure and style needed for sentence simplification. Across almost all languages, 5-shot improves both SARI and $\mathbf{F}_{ref}$ over 0-shot, especially for Gemma and Qwen. For instance, English (Gemma 39.4 to 43.6 SARI), Sinhala (Gemma 34.8 to 39.9), Thai (Gemma 37.9 to 41.3), Tamil (Gemma 32.8 to 39.3), and Pashto (Gemma 33.3 to 37.9) all show clear gains from 0 to 5 shots. This indicates that in-context examples enhance the model's ability to balance meaning preservation with simplification. These results demonstrate that few-shot examples improve both lexical and semantic alignment in medium-resource settings.

Discussion: The SARI component analysis shows that LLMs handle content retention, deletion, and addition differently across languages. The DEL component is the most stable, with consistently high values and high scores in English (68–78) and Thai (77–85), showing that models remove unnecessary detail. Even in low-resource languages like Pashto (Gemma 70), models delete unnecessary parts for simplification. The KEEP component improves notably with few-shot learning, particularly where zero-shot transfer is weak. For example, LLaMA's KEEP score in Thai increases from 24.83 to 39.99, demonstrating better retention and structural alignment when examples are provided. Gemma performs consistently well across languages, showing a strong ability to remove redundant material while maintaining fluency and coherence.

The ADD component remains the most variable and challenging simplification aspect, with consistently low values across all languages. English achieves modest results (5.24 - 11.91), while scores in low-resource languages remain minimal, including Sinhala (0.00–5.65), Thai (0.12–1.48), Tamil (0.11–5.16), and Pashto (0.50–5.39). This challenge is most evident in low-resource settings.

For Sinhala and Pashto, except for Gemma, the ADD value is below 1 in most cases, indicating limited ability to generate new simple words. Overall, while LLMs are good at deleting unnecessary information and retaining key content, they struggle to generate new and contextually meaningful text. This limitation is especially pronounced in low-resource and morphologically complex languages for sentence simplification.

The SARI scores for English remain consistent across all models, likely because English is included in every model's training data and benefits from extensive available resources. LLaMA shows relatively high SARI scores for Tamil and Thai, indicating that direct language inclusion during training can positively impact performance. However, it is difficult to draw firm conclusions due to the limited transparency regarding the complete set of languages and the training data size for each model.

6. Conclusion

This paper introduced OasisSimp, a new multilingual dataset for sentence-level simplification, covering English, Sinhala, Thai, Tamil, and Pashto. To ensure high-quality data, all simplifications were produced by human annotators. The dataset focuses on asian languages that have been largely overlooked in existing simplification research. We evaluated the performance of eight LLMs on the OasisSimp dataset under zero-shot and few-shot learning settings. The results indicate that all models face challenges in zero-shot scenarios; however, their performance improves substantially when provided with a few examples, demonstrating the effectiveness of few-shot learning for sentence simplification. Gemma exhibited the most consistent performance across languages and conditions among the tested models. Overall, OasisSimp represents a significant step toward broadening the scope of sentence simplification research beyond high-resource languages. This work lays the foundation for developing and evaluating more inclusive and equitable sentence simplification systems by providing high-quality, human-annotated data for less-resourced languages such as Sinhala and Pashto. We hope that OasisSimp will encourage future research on multilingual simplification and contribute to making written information more accessible to diverse global audiences.

7. Ethical Considerations and Limitations

7.1. Limitations

Language selection was purely dependent on the availability of speakers of the corresponding language. Therefore, it is not possible to carry out a systematic study with respect to language classes. The dataset size was limited, at the same time comparable to existing datasets, by the availability of annotators and/or funding. This also affected the number of simplified sentences for some languages and resulted in no human evaluation of the LLM output. There wasn't enough data to train LLMs for the sentence simplification task, and human evaluation of LLM outputs was not conducted due to resource constraints.

Due to the lack of detailed information on the languages included in training and the sizes of the data used for each LLM, it is difficult to draw concrete conclusions about language inclusion in LLMs. A more in-depth analysis of how language-family relationships and cross-linguistic transfer affect model performance would provide valuable insights, but it is beyond the scope of this work.

7.2. Ethical Considerations

All the complex sentences were retrieved from publicly available authentic sources (e.g., government documents, news, and Wikipedia). The simplification guidelines issued for human participants clearly stated the need to retain the original meaning of the sentence. Therefore, we do not expect any undesirable content in the simplified version of the corpus. However, the authors themselves did not individually check all the sentences for undesirable content. The annotators were offered co-authorship or paid at the standard rates for the country in which they reside.

8. Acknowledgement

This work was partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN-2024-06887, the NSERC Discovery Launch Supplement DGECR-2024-00008, and the Digital Research Alliance of Canada (formerly Compute Canada) Grant RRG no. 5397 on "Multilingual multicultural NLP and LLMs". We also thank CoolWei AI Lab for providing GPU resources that enabled this research. This work was partially supported by the ELOQUENCE project (grant number 101070558) funded by the UKRI and the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the UKRI, European Union, or European Commission-EU. Neither the European Union nor the granting authority can be held responsible for them.

9. Bibliographical References

Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

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

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

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

Yusra Anees and Sadaf Abdul-Rauf. 2021. Automatic sentence simplification in low resource settings for Urdu. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 60–70.

Chandrayee Basu, Rosni Vasu, Michihiro Yasunaga, and Qian Yang. 2023. Med-easi: Finely annotated dataset and models for controllable simplification of medical texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

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.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Nakhun Chumpolsathien. 2020. Using knowledge distillation from keyword extraction to improve the informativeness of neural cross-lingual summarization. Master's thesis, Beijing Institute of Technology.

William Coster and David Kauchak. 2011. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA. Association for Computational Linguistics.

Tanvi Dadu, Kartikey Pant, Seema Nagar, Ferdous Ahmed Barbhuiya, and Kuntal Dey. 2021. Text simplification for comprehension-based question-answering. *arXiv preprint arXiv:2109.13984*.

Asociación Española de Normalización. 2018. Lectura fácil. pautas y recomendaciones para la elaboración de documentos. *AENOR INTERNACIONAL SAU, Madrid*.

Anne Eschenbrücher. 2021. What makes a concept complex? measuring conceptual complexity as a precursor for text simplification. In *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 154–160.

Isabel Espinosa-Zaragoza, José Abreu-Salas, Elena Lloret, Paloma Moreda, and Manuel Palomar. 2023. A review of research-based automatic text simplification tools. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 321–330, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Aloka Fernando, Surangika Ranathunga, and Gihan Dias. 2020. Data Augmentation and Terminology Integration for Domain-Specific Sinhala-English-Tamil Statistical Machine Translation. *CoRR*, abs/2011.02821.

Daniel Ferrés and Horacio Saggion. 2022. ALEXSIS: A dataset for lexical simplification in Spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3582–3594, Marseille, France. European Language Resources Association.

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.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Koki Horiguchi, Tomoyuki Kajiwara, Takashi Ninomiya, Shoko Wakamiya, and Eiji Aramaki. 2025. MultiMSD: A corpus for multilingual medical text simplification from online medical references. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9248–9258, Vienna, Austria. Association for Computational Linguistics.

Sebastian Joseph, Kathryn Kazanas, Keziah Reina, Vishnesh Ramanathan, Wei Xu, Byron Wallace, and Junyi Jessy Li. 2023. Multilingual simplification of medical texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16662–16692, Singapore. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Tomoyuki Kajiwara and Mamoru Komachi. 2016. Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1147–1158, Osaka, Japan. The COLING 2016 Organizing Committee.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking large language models on sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. Association for Computational Linguistics.

Dayeon Ki and Marine Carpuat. 2025. Automatic input rewriting improves translation with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10829–10856, Albuquerque, New Mexico. 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).

Xinyu Lu, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2021. An unsupervised method for building sentence simplification corpora in multiple languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 227–237, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the Twelfth 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.

Sneha Mondal, Ritika Ritika, Ashish Agrawal, Preethi Jyothi, and Aravindan Raghuveer. 2024. DIMSIM: Distilled multilingual critics for Indic text simplification. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 16093–16109.

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

Gustavo H Paetzold and Lucia Specia. 2017. A survey on lexical simplification. *Journal of Artificial Intelligence Research*, 60:549–593.

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.

Wannaphong Phatthiyaphaibun, Korakot Chaovavanich, Charin Polpanumas, Arthit Suriyawongkul, Lalita Lowphansirikul, and Pattarawat Chormai. 2024. PyThaiNLP: Thai natural language processing in Python.

Surangika Ranathunga and Nisansa de Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. 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 823–848, Online only. Association for Computational Linguistics.

Surangika Ranathunga, Fathima Farhath, Uthayasanker Thayasivam, Sanath Jayasena, and Gihan Dias. 2018. Si-Ta: Machine translation of Sinhala and Tamil official documents. In *2018 National Information Technology Conference (NITC)*, pages 1–6. IEEE.

Surangika Ranathunga, Rumesh Sirithunga, Himashi Rathnayake, Lahiru De Silva, Thamindu Aluthwala, Saman Peramuna, and Ravi Shekhar. 2025. SiTSE: Sinhala Text Simplification Dataset and Evaluation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(5):1–19.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Michael J Ryan, Tarek Naous, and Wei Xu. 2023. Revisiting non-English text simplification: A unified multilingual benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4898–4927, Toronto, Canada. Association for Computational Linguistics.

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

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

Matthew Shardlow. 2014. A Survey of Automated Text Simplification. *International Journal of Advanced Computer Science and Applications*, 4(1).

Kim Cheng Sheang and Horacio Saggion. 2021. Controllable sentence simplification with a unified text-to-text transfer transformer. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 341–352, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Sara Tonelli, Alessio Palmero Aprosio, and Francesca Saltori. 2016. Simpitiki: a simplification corpus for italian. In *Proceedings of the third Italian conference on computational linguistics CLiC-it*, pages 291–296.

Gayatri Venugopal, Dhanya Pramod, and Ravi Shekhar. 2022. CWID-hi: A dataset for complex word identification in Hindi text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5627–5636, 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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT.

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.

10. Appendix

Other than the SARI and the BERTScore ($\mathbf{F}_{\text{ref}}$) discussed in section 5, the models are also evaluated on OasisSimp using BLEU scores (Papineni et al., 2002) in Table 8, Table 9, Table 10, Table 11, and Table 12.

Performance varies across languages: EuroLLM achieves the best results for English, Gemma performs best for Sinhala, while Cmd-R and LLaMA obtain the strongest results for Thai, depending on the shot setting. For Tamil and Pashto, models Cmd-R and Gemma achieve the best performance on Tamil, while models EuroLLM and Cmd-R perform best on Pashto, depending on the number of shots.

Table 8: BLEU scores on **English (*OasisSimp-EN*)** dataset across different shot settings. Highest value in **bold** and lowest underlined.

	English		
Model	0 Shot	1 Shot	5 Shot
Aya	18.38	20.59	21.72
Cmd-R	21.03	19.58	20.72
DeepSeek	17.19	17.27	18.91
EuroLLM	**23.07**	**23.07**	**22.79**
Gemma	15.48	16.00	18.92
LLaMA	22.25	21.52	21.78
Mistral	17.40	19.67	20.63
Qwen	11.55	17.21	19.09

Table 9: BLEU scores on **Sinhala(*OasisSimp-SI*)** dataset across different shot settings. Highest value in **bold** and lowest underlined.

	Sinhala		
Model	0 Shot	1 Shot	5 Shot
Aya	7.25	9.03	10.80
Cmd-R	11.86	11.90	11.63
DeepSeek	8.04	7.12	3.08
EuroLLM	2.24	2.19	0.00
Gemma	**14.38**	**22.36**	**31.82**
LLaMA	6.45	6.77	7.34
Mistral	2.29	4.55	7.72
Qwen	12.28	12.64	13.02

Table 10: BLEU scores on **Thai(*OasisSimp-TH*)** dataset across different shot settings. Highest value in **bold** and lowest underlined.

	Thai		
Model	0 Shot	1 Shot	5 Shot
Aya	7.70	9.65	11.58
Cmd-R	**13.08**	12.42	13.84
DeepSeek	2.24	5.99	5.71
EuroLLM	11.76	13.94	16.73
Gemma	3.67	11.20	14.43
LLaMA	2.16	**14.71**	**20.36**
Mistral	4.54	8.52	9.79
Qwen	5.35	5.55	9.16

Table 11: BLEU scores on **Tamil(*OasisSimp-TA*)** dataset across different shot settings. Highest value in **bold** and lowest underlined.

	Tamil		
Model	0 Shot	1 Shot	5 Shot
Aya	11.26	20.69	29.97
Cmd-R	**19.36**	**26.17**	30.58
DeepSeek	3.67	10.11	1.48
EuroLLM	17.04	18.98	15.60
Gemma	9.62	21.36	**36.71**
LLaMA	7.94	23.17	22.21
Mistral	7.38	22.21	27.66
Qwen	18.77	23.30	29.67

Table 12: BLEU scores on **Pashto(*OasisSimp-PS*)** dataset across different shot settings. Highest value in **bold** and lowest underlined.

	Pashto		
Model	0 Shot	1 Shot	5 Shot
Aya	7.42	22.04	32.73
Cmd-R	28.39	33.99	**38.49**
DeepSeek	18.37	25.15	26.92
EuroLLM	**33.74**	**34.65**	35.54
Gemma	7.22	12.84	22.65
LLaMA	5.28	22.09	23.30
Mistral	7.32	16.11	20.95
Qwen	21.03	27.00	30.01