SiniticMTError: A Machine Translation Dataset with Error Annotations for Sinitic Languages

Hannah Liu[1], Junghyun Min[2], Ethan Yue Heng Cheung[1], Shou-Yi Hung[1], Syed Mekael Wasti[3], Runtong Liang[1], Shiyao Qian[1], Shizhao Zheng [1], Elsie Chan[1], Ka Ieng Charlotte Lo[1], Wing Yu Yip[1], Richard Tzong-Han Tsai[4], En-Shiun Annie Lee[1,3]

[1]University of Toronto [2]Georgetown University

[3]Ontario Tech University [4]National Central University, Taiwan

hannahhere.liu@mail.utoronto.ca

Abstract

Despite major advances in machine translation (MT) in recent years, progress remains limited for many low-resource languages that lack large-scale training data and linguistic resources. Cantonese and Wu Chinese are two Sinitic examples, although each enjoys more than 80 million speakers around the world. In this paper, we introduce SiniticMTError, a novel dataset that builds on existing parallel corpora to provide error span, error type, and error severity annotations in machine-translated examples from English to Mandarin, Cantonese, and Wu Chinese. Our dataset serves as a resource for the MT community to utilize in fine-tuning models with error detection capabilities, supporting research on translation quality estimation, error-aware generation, and low-resource language evaluation. We report our rigorous annotation process by native speakers, with analyses on inter-annotator agreement, iterative feedback, and patterns in error type and severity.

Source: The weather is beautiful today.

MT: 今天 (*Today*) 我觉得 (*I think*) 天气很 (*the weather is*) 漂亮。 (*beautiful*)

Reference: 今天 (*Today*) 天气很 (*the weather is*) 好。 (*beautiful*)

Annotation Spans:

- **Text:** "我觉得"
  Error type: Addition
  Severity: Major
  Start: 2 **End:** 5
- **Text:** "漂亮"
  Error type: Mistranslation
  Severity: Minor
  Start: 8 **End:** 10

Figure 1: Sample Mandarin entry. `mt` looks fluent, but contains subtle semantic errors: an unwarranted subjective phrase (Addition) and a lexical mistranslation (Mistranslation). While 漂亮 directly translates to *beautiful*, it usually describes people or objects and 好 *good* is more natural when used to describe the weather.

1 Introduction

Machine translation (MT) has made significant advancements in recent years as either supervised systems (Luong et al., 2015; Lakew et al., 2018; Liu et al., 2020; Wang et al., 2022; Liu, 2022; Park et al., 2023) or LLMs (Zhu et al., 2024; Freitag et al., 2024). However, such work often focuses on higher-resource languages, and progress remains limited with low-resource languages (Ranathunga et al., 2023), where fine-tuned models suffer from poor performance (Lee et al., 2022; Shliazhko et al., 2024) and LLMs output noise (Iyer et al., 2024; Levine et al., 2025).

We focus on two major yet low-resource Sinitic languages, Cantonese[1] and Wu Chinese[2]. They remain underserved despite having more than 80 million and 83 million speakers respectively, across southern and eastern China and various diasporas across the world (Chappell, 2015; Eberhard et al., 2023). The limited progress is attributable to the dominance of Mandarin as *lingua franca* in these regions (Norman, 1988; Li, 2006), scarcity in publicly available parallel corpora (Xiang et al., 2024), the lack of standardization in writing systems (Pan et al., 1991; Tang et al., 2002; Kwan-hin and Bauer, 2002; Snow, 2008), and their status as primarily vernacular (i.e. spoken rather than written) languages (Pan et al., 1991; Snow, 2004; Li, 2006).

In this paper, we add to the Sinitic MT literature by presenting SiniticMTError, a novel suite of datasets that build on `FLORES+` (Goyal et al., 2022; NLLB Team et al., 2024; Yu et al., 2024) to provide erroneous machine translation examples and

[1]Also known as Yue (Eberhard et al., 2023)

[2]Whose most well-known dialect is Shanghainese (Eberhard et al., 2023). We discuss our terminology for these languages in Appendix A.

Figure 2: Overview of our annotation pipeline. We input English sentences from FLORES+ to generate mt outputs (e.g., from NLLB) into Sinitic languages (Mandarin, Cantonese, Wu).

detailed span-level error annotations including error type and severity for Mandarin, Cantonese, and Wu Chinese (Figure 1).

We report preliminary statistics across 2,000 Mandarin and 1,000 Cantonese sentences, with Wu Chinese annotations still in progress. Early results suggest distinctive error distributions across languages. We describe our annotation process conducted by native speakers (Figure 2; Section 3), discussions on patterns in error types and severity (Section 4), and detailed analyses on translation issues (Section 5), offering both linguistic insights and resources to advance low-resource MT. A sample of our dataset is available at an Anonymous GitHub Repository.

2 Related Work

Annotation frameworks like Multidimensional Quality Metrics (MQM; Burchardt, 2013) and Error Span Annotations (ESA; Chen et al., 2020; Kocmi et al., 2024, *inter alia*) represent important steps toward more equitable multilingual NLP by introducing standardized methods for evaluating translation quality. MQM and ESA facilitate identifying and localizing errors to improve translation pipelines (Chen et al., 2020; Zhang et al., 2025), and explicit fine-tuning or prompting to improve the quality of generative model output (Kocmi and Federmann, 2023; Wang et al., 2024).

While these frameworks offer a promising foundation, existing datasets built on them have largely focused on high-resource languages such as English, French, and Mandarin Chinese (Freitag et al., 2021; Sellam et al., 2021). One effort to adapt these tools to lower-resource languages is by Singh et al. (2024) who focus on low-resource Indic languages.

FLORES-101, FLORES-200, and FLORES+ (Goyal et al., 2022; NLLB Team et al., 2024) are initiatives that have introduced multilingual benchmarks covering over 100 languages, many of which are low-resource, to support equitable evaluation of MT systems. Building upon FLORES-200, AfriCOMET (Wang et al., 2024) is an annotation effort aiming to bridge the gap between low-resource languages and MT systems by providing annotated datasets for underrepresented African languages. Similar work has emerged for Bambara (Dou and Neubig, 2022), Amharic and Tigrinya (Shapiro et al., 2023), and Spanish languages (Perez-Ortiz et al., 2024), offering parallel corpora and baseline models. Within the Sinitic language family, resources in Hokkien[3] (Lu et al., 2022; Chen et al., 2023; Lu et al., 2024) and Hakka (Hung and Huang, 2022; Lai et al., 2024) have been compiled.

Despite related efforts, publicly available resources in Cantonese and Wu Chinese remain scarce. Proprietary LLMs offer commercial service in Cantonese (OpenAI et al., 2024; Team et al.,

[3]Also known as Min Nan (Southern Min; Eberhard et al., 2023)

2024) and several Cantonese–Mandarin or Cantonese–English parallel corpora exist in the form of subtitles (Wong and Zhang, 2017), dictionaries (Mair and DeFrancis, 2003), or government transcripts (Lee, 2011). However, resources are limited in scale or accessibility as surveyed by Xiang et al. (2024). As a result, previous work in Cantonese MT have relied on synthetic data augmentation (Liu, 2022; Hong et al., 2024). To the best of our knowledge, the recent addition to FLORES+ (Yu et al., 2024) represents the sole publicly available MT resource in Wu Chinese.

3 Dataset Construction and Annotation

We follow Han orthographic standardization of Yu et al. (2024) for Wu Chinese, where phonetic transliterations follow the pronunciation of the Chongming dialect. We also follow NLLB Team et al. (2024)'s choice of traditional Han orthography for Cantonese.

Each example in the dataset consists of a triplet of sentences: a source sentence (`src`), a machine-translated sentence (`mt`), and a reference translation (`ref`). The `src` and `ref` pairs are drawn from the FLORES+, which is an extension of the FLORES-200 dataset (NLLB Team et al., 2024). The `mt` sentences are generated using the 600M NLLB-200 (Team et al., 2022) for Mandarin and Cantonese, and Qwen2.5 Max (Team, 2024) for Wu Chinese. We describe our model selection and `mt` generation in detail in Appendix B.

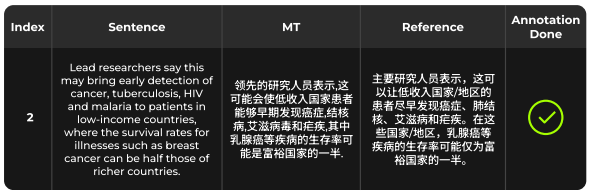

Figure 3: Annotators were presented with the original English source sentence, the machine-translated output (`mt`), and a human reference translation for comparison. The green check marks indicate completed annotations.

The constructed dataset is then provided to annotators for error span annotation. For each language pair (English-Mandarin, English-Cantonese, English-Wu Chinese), we recruit three bilingual annotators who are listed as collaborating authors. The annotators are native speakers of their respective target languages and highly proficient in English. They are familiar with both the linguistic conventions of their target language and the goals of the annotation task.

We use the TranslationCorrect tool (Wasti et al., 2025) as our annotation interface. A sample screen from the interface is illustrated in Figure 3. We discuss the interface in greater detail in Appendix C.

3.1 Annotation Guidelines

Annotators are instructed to examine the `mt` sentence with reference to both `src` and `ref`, and identify any translation errors by highlighting spans directly on the `mt` sentences. For each error span, annotators categorize the **severity** and **error type**, while recording the erroneous **span indices** in the `mt` sentence. The label spaces for **severity** and **type**, adopted from MQM guidelines (Burchardt, 2013) and the AfriCOMET framework (Wang et al., 2024) are described in Appendix D.

Granularity. When identifying errors, the annotators are asked to be as fine-grained as possible. For example, if a sentence contains two words that are each mistranslated, two separate mistranslation error spans should be recorded. If a single text segment contains multiple errors, the leftmost span with highest severity is to be recorded.

Additional info. One error type is omission, where content in the source sentence is missing from the translation. In such case, annotators are instructed to insert the missing information the post-editing box and highlight the inserted span with the appropriate type (omission), span, and severity labels.

3.2 Annotation Workflow

First, annotators were trained on using the annotation tool, then provided with detailed guidelines, including definitions of error categories and severity levels, along with examples. The training was then followed by a multi-stage pilot setup.

Mandarin pilot. Mandarin annotators had two rounds of pilot studies. In the first round, each annotator completed 50 examples, which were then reviewed by a language lead. The lead provided feedback and held group discussions to ensure consistency, before annotators completed a second round of 50 examples. Qualitative analysis showed clear improvement in annotation consistency.

Cantonese pilot. Cantonese annotators followed a more granular setup consisting of 4-rounds. The initial round included 50 examples and the others 10 examples each, completing 80 examples in total. Additional rounds of 10 examples were

added to ensure sufficient agreement before proceeding. We describe the details of adjustments made during the pilot stages in Appendix E.

Main annotation phase. During the main annotation phase, annotators worked in batches of 50 to 200 sentences. With each batch, we employed a similar iterative process where annotations were reviewed, recurring issues discussed, and relevant guidelines fine-tuned. During such iterative process, guidelines on annotation granularity and categorizing missing punctuation as omission were established.

Quality assurance. After completing the first round of annotations (main annotator round), annotators proceed to the Quality Assurance (QA) stage. In this phase, each annotator will be assigned sentences previously annotated by another team member. Each participant reviews these annotations and discusses any discrepancies with the original annotator, should there be any, to improve consistency across annotators. We plan to perform inter-annotator agreement analysis in multiple stages.

4 Preliminary Dataset Statistics

We report our preliminary statistics from the ongoing annotation of Mandarin and Cantonese data. At the time of writing, we have annotated 2009 Mandarin sentences and 996 Cantonese sentences, with Wu Chinese still in its pilot training phase. Table 1 and Table 2 present the distribution of error categories and severity levels in our current annotation data. On average, each Mandarin sentence contains 1.93 annotation spans, while each Cantonese sentence contains 6.82 spans, showing a much higher span density than Mandarin. For Wu Chinese, we will follow the same annotation setup as Mandarin.

In Mandarin annotations, mistranslation, omission, and grammar errors are the most frequent error types. For Cantonese, the most common error types are omission, typography, and mistranslation. This suggests that although these language models are capable of generating well-formed sentences, they often struggle with incomplete representations of semantic and syntactic structure (*c.f.* Berglund et al., 2024; Min et al., 2025). Cantonese outputs also record a much higher number of error spans per sentence compared to Mandarin, which indicates that the model performs worse in Cantonese. This likely represents the limited availability of high-quality Cantonese resources (Xiang et al., 2024). Finally, an analysis of the number of major versus minor errors shows that many errors are minor semantic or felicity issues rather than complete misinterpretations, attesting to the powerful multilingual adaptability of transformer-based language models (e.g. Liu et al., 2020; Zhu et al., 2024). Overall, these early results show both shared and language-specific challenges in Sinitic machine translation.

Type	Count	Proportion	Freq / 1k
Mistranslation	2,306	61.2%	1,148
Omission	534	14.2%	266
Grammar	267	7.1%	133
Unintelligible	198	5.3%	99
Typography	133	3.5%	66
Untranslated	128	3.4%	64
Spelling	114	3.0%	57
Addition	86	2.3%	43

(a) Mandarin (2009 sentences)

Type	Count	Proportion	Freq / 1k
Mistranslation	2,471	36.3%	2,481
Typography	1,723	25.6%	1,730
Omission	1,355	19.9%	1,360
Addition	470	6.9%	472
Grammar	348	5.1%	349
Spelling	244	3.6%	245
Untranslated	180	2.6%	181
Unintelligible	1	0.0%	1

(b) Cantonese (996 sentences)

Table 1: Error Category counts in Mandarin and Cantonese `mt` outputs, sorted by frequency.

Severity	Mandarin	Cantonese
Minor	2,127	5,863
Major	1,639	949

Table 2: Error severity counts in Mandarin and Cantonese `mt` outputs. Minor errors are more frequent in both languages, though major errors remain substantial.

5 Annotation Insights

In this section, we discuss several insights we gathered during span-level our annotation efforts, which highlight cross-lingual differences in syntax and semantics.

Differences in error type distribution. Beyond lexical differences, Cantonese and Mandarin differ in several ways, e.g. word order and function word inventory (Zhang, 1998). The suite of aspect and feature markers and sentence-final particles differ, with Cantonese boasting a richer inventory that encodes more fine-grained nuance in tense, speaker

stance or attitude (Yap and Chor, 2011; Lee, 2019). Cantonese also allows serial verb constructions (e.g. *go buy eat rice* as a sequence) to a greater extent than in Mandarin (Matthews, 2006). Such differences may be reflected in Table 5, where the grammar error type is much more frequent per sentence in Cantonese than in Mandarin. The difference in functional word inventory is likely a major source of such error; we observe erroneous particle uses in Cantonese `mt` sentences, many of which are valid particles in Mandarin yet ungrammatical in Cantonese. Examples include erroneous use of 才 in place of Cantonese particle 先 *only after*, Mandarin possessive 的 in place of Cantonese 嘅; and Mandarin 在 in place of Cantonese copula 喺 *be at*.

Translationese. Compared to English, Chinese languages typically have simpler sentence structure and clause segmentation (Morbiato, 2018), relies much more heavily on particles and pro-drop (Li and Thompson, 1979; Paul, 2014), and has different headedness principles (Levy and Manning, 2003). However, many machine translation outputs were constructed in an English clause structure (i.e. translationese; Riley et al., 2020), which results in unnatural or even ungrammatical phrasing. For example, subordinate clauses were often translated as long embedded segments, instead of being split into multiple short sentences, which is more natural in Chinese.

Lack of standardization. As discussed in Section 1, MT in Cantonese and Wu Chinese face unique difficulties as a primarily a spoken language. Written Cantonese most often appears in informal contexts like texting, where conventions are inconsistent. As a result, both MT systems and Cantonese annotators face the challenge of the lack of an accepted written norm. There are several instances in the annotations where multiple written forms correspond to the same spoken word, such as 嗰樣, 咁樣 *like this, this way*. Annotators accept both variants: 嗰樣 is considered the standard form, while 咁樣 is more widely used in practice.

Language Confusion. We also observe language confusion; some machine translations in Wu or Cantonese unexpectedly output material from other languages. While it may have been unsurprising to observe language confusion within the Sinitic family, we observe one instance where *flu* was erroneously translated into Japanese インフルエンザ, a Katakana adoption of *influenza*. This behavior may resemble code mixing in humans, a behavior noted across bilingual infants (Lanza, 1997) and adults (Muysken, 2000). However, we believe this is closer to language confusion or interference often observed in multilingual NLP systems (Wang et al., 2020; Yu et al., 2024; Lee et al., 2025), which has been attributed to high temperature, language mismatch between representation learning and preference tuning, and under-training (Marchisio et al., 2024).

These behaviors illustrate the challenge of stable and accurate generation in a low-resource language in multilingual machine translation. More broadly, these findings call for more attention and resources for low-resource Sinitic languages, which lack representation in current multilingual NLP research.

6 Conclusion

In this paper, we introduce SiniticMTError, a parallel English-to-Sinitic machine translation error dataset in Mandarin, Cantonese, and Wu Chinese that comprise of erroneous machine translation text, erroneous spans, their error type and severity, and a gold-label translation. The dataset is a work-in-progress; the Mandarin split is complete with 2,000 sentences; Cantonese in progress with 1,000, and Wu in its pilot stages.

SiniticMTError will be one of the first human-annotated span-level resources for Wu Chinese, along with the Shanghainese Universal Dependencies dataset (Yang, 2025) and serve as a significant addition to a small collection of Cantonese human-annotated resources. Beyond the dataset's immediate role in error analysis and MT, we anticipate that the parallel dataset can support a wide range of downstream applications. The dataset may be adapted for several natural language understanding tasks, including language detection and linguistic acceptability judgment. Moreover, the parallel nature of the dataset makes it a promising resource for transfer learning, which allows models trained on high-resource languages to be adapted more effectively to low-resource Sinitic varieties.

Limitations and Future Work

While we present the annotation guidelines and workflow, the annotation is incomplete at the time of submission. We plan to complete the annotation by the time the paper is to be presented, but the reported analysis is preliminary and may not

be representative of our completed work.

In addition, the dataset builds on FLORES+ (Goyal et al., 2022; NLLB Team et al., 2024; Yu et al., 2024), whose source sentences are in English. Although the Cantonese and Wu sentences are parallel, the erroneous machine translations are attempts to translate from English using MT systems and LLM, and may not represent real-world use cases of translating to and from Cantonese and Wu Chinese, whose speakers may often be limited proficiency bilinguals fluent in Mandarin (Li, 2006).

Finally, our work only spans Cantonese and Wu Chinese among over a dozen of Sinitic languages (Tang and van Heuven, 2007; Chappell, 2015), each with varying numbers of native and bilingual speakers (Norman, 1988; Eberhard et al., 2023). Future work may bootstrap our annotation guidelines and tools to expand to other Sinitic languages as well.

In the future, we plan to finish the annotation of about 2000 sentences for both Mandarin and Cantonese, as well as around 1000 sentences for Wu Chinese. We also intend to conduct experiments that showcase the utility of the datset for machine translation evaluation and error analysis. We further plan to release both the dataset and our experimental results to encourage broader adoption and future research in this area.

Acknowledgment

We thank the members of the Lee Language Lab for their valuable comments and feedback. Any errors remain our own.

Responsible Research Statement

Our work introduces a dataset built from a publicly available, non-sensitive source dataset FLORES+ (Goyal et al., 2022; NLLB Team et al., 2024; Yu et al., 2024). The annotators are native speakers of the target language (Cantonese or Wu Chinese), knowledgeable about their work and the dataset's downstream use. We internally review the work to ensure it does not contain any personally identifiable information.

The mt outputs on which error spans were annotated were generated by publicly available LLMs and may contain unintended biases or stereotypes. We encourage responsible and context-aware use of our dataset in downstream applications.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2024. The reversal curse: LLMs trained on "a is b" fail to learn "b is a". In *The Twelfth International Conference on Learning Representations*.

Aljoscha Burchardt. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Hilary Chappell. 2015. *Diversity in Sinitic languages*. Oxford University Press.

Mengyun Chen, Tao Ge, Xingxing Zhang, Furu Wei, and Ming Zhou. 2020. Improving the efficiency of grammatical error correction with erroneous span detection and correction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7162–7169, Online. Association for Computational Linguistics.

Peng-Jen Chen, Kevin Tran, Yilin Yang, Jingfei Du, Justine Kao, Yu-An Chung, Paden Tomasello, Paul-Ambroise Duquenne, Holger Schwenk, Hongyu Gong, Hirofumi Inaguma, Sravya Popuri, Changhan Wang, Juan Pino, Wei-Ning Hsu, and Ann Lee. 2023. Speech-to-speech translation for a real-world unwritten language. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4969–4983, Toronto, Canada. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Zi-Yi Dou and Graham Neubig. 2022. Unsupervised machine translation of low-resource languages: A case study on bambara. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 335–346. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2023. *Ethnologue: Languages of the World*, 26 edition. SIL International, Dallas.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021.

Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kung Hong, Lifeng Han, Riza Batista-Navarro, and Goran Nenadic. 2024. CantonMT: Cantonese to English NMT platform with fine-tuned models using real and synthetic back-translation data. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 590–599, Sheffield, UK. European Association for Machine Translation (EAMT).

Yi-Hsiang Hung and Yi-Chin Huang. 2022. A preliminary study on Mandarin-Hakka neural machine translation using small-sized data. In *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)*, pages 307–315, Taipei, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Vivek Iyer, Bhavitvya Malik, Pavel Stepachev, Pinzhen Chen, Barry Haddow, and Alexandra Birch. 2024. Quality or quantity? on data scale and diversity in adapting large language models for low-resource translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1393–1409, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Cheung Kwan-hin and Robert S Bauer. 2002. The representation of cantonese with chinese characters. *Journal of Chinese Linguistics Monograph Series*, pages i–489.

Yen-Chun Lai, Yi-Jun Zheng, Wen-Han Hsu, Yan-Ming Lin, Cheng-Hsiu Cho, Chlh-Chung Kuo, Chao-Shih Huang, and Yuan-Fu Liao. 2024. Construction of large language models for taigi and hakka using transfer learning. In *2024 27th Conference of the Oriental COCOSDA International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6.

Surafel M. Lakew, Aliia Erofeeva, Matteo Negri, Marcello Federico, and Marco Turchi. 2018. Transfer learning in multilingual neural machine translation with dynamic vocabulary. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 54–61, Brussels. International Conference on Spoken Language Translation.

Elizabeth Lanza. 1997. *Language mixing in infant bilingualism: A sociolinguistic perspective*. Oxford University Press.

En-Shiun Annie Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Ifeoluwa Adelani, Ruisi Su, and Arya D. McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics.

Nahyun Lee, Yeongseo Woo, Hyunwoo Ko, and Guijin Son. 2025. Controlling language confusion in multilingual LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1026–1035, Vienna, Austria. Association for Computational Linguistics.

Peppina Po-lun Lee. 2019. *Focus manifestation in Mandarin Chinese and Cantonese: A comparative perspective*. Routledge.

Thomas H. C. Lee. 2011. A bilingual corpus of legislative texts in hong kong: The hong kong hansard corpus. *Language Resources and Evaluation*, 45(2):123–139.

Lauren Levine, Junghyun Min, and Amir Zeldes. 2025. Building UD cairo for Old English in the classroom. In *Proceedings of the Eighth Workshop on Universal*

Dependencies (UDW, SyntaxFest 2025), pages 97–104, Ljubljana, Slovenia. Association for Computational Linguistics.

Roger Levy and Christopher D. Manning. 2003. Is it harder to parse Chinese, or the Chinese treebank? In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 439–446, Sapporo, Japan. Association for Computational Linguistics.

Charles N Li and Sandra A Thompson. 1979. Third-person pronouns and zero-anaphora in chinese discourse. *Syntax and semantics*, 12(01).

David C. S. Li. 2006. Chinese as a lingua franca in greater china. *Annual Review of Applied Linguistics*, 26:149–176.

Evelyn Kai-Yan Liu. 2022. Low-resource neural machine translation: A case study of Cantonese. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 28–40, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Bo-Han Lu, Yi-Hsuan Lin, Annie Lee, and Richard Tzong-Han Tsai. 2024. Enhancing Taiwanese hokkien dual translation by exploring and standardizing of four writing systems. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6077–6090, Torino, Italia. ELRA and ICCL.

Sin-En Lu, Bo-Han Lu, Chao-Yi Lu, and Richard Tzong-Han Tsai. 2022. Exploring methods for building dialects-Mandarin code-mixing corpora: A case study in Taiwanese hokkien. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6287–6305, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Victor H. Mair and John DeFrancis. 2003. Abc cantonese-english dictionary corpus. Data derived from the ABC Cantonese-English dictionary.

Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. 2024. Understanding and mitigating language confusion in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6653–6677, Miami, Florida, USA. Association for Computational Linguistics.

Stephen Matthews. 2006. On serial verb constructions in cantonese. *Serial verb constructions: A cross-linguistic typology*, 2.

Stephen Matthews and Virginia Yip. 2011. *Cantonese: A Comprehensive Grammar*, 2nd edition. Routledge, London. EBook published 23 May 2013.

Junghyun Min, Minho Lee, Woochul Lee, and Yeonsoo Lee. 2025. Punctuation restoration improves structure understanding without supervision. In *Proceedings of the 10th Workshop on Representation Learning for NLP (RepL4NLP-2025)*, pages 120–130, Albuquerque, NM. Association for Computational Linguistics.

Anna Morbiato. 2018. Word order and sentence structure in mandarin chinese: new perspectives.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Pieter Muysken. 2000. *Bilingual speech: A typology of code-mixing*. Cambridge university press.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Jerry Norman. 1988. *Chinese*. Cambridge University Press.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Wuyun Pan, S.F. Zhengzhang, R.J. You, and Lien Chinfa. 1991. An introduction to the wu dialects. *Journal of Chinese Linguistics Monograph Series*, (3):235–291.

Geon Woo Park, Junghwa Lee, Meiying Ren, Allison Shindell, and Yeonsoo Lee. 2023. VARCO-MT: NCSOFT's WMT'23 terminology shared task submission. In *Proceedings of the Eighth Conference on Machine Translation*, pages 919–925, Singapore. Association for Computational Linguistics.

Waltraud Paul. 2014. Why particles are not particular: Sentence-final particles in chinese as heads of a split cp. *Studia linguistica*, 68(1):77–115.

Juan Antonio Perez-Ortiz, Felipe Sánchez-Martínez, Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis, Aaron Galiano Jimenez, Antoni Oliver, Claudi Aventín-Boya, Alejandro Pardos, Cristina Valdés, Jusèp Loís Sans Socasau, and Juan Pablo Martínez. 2024. Expanding the FLORES+ multilingual benchmark with translations for Aragonese, aranese, Asturian, and Valencian. In *Proceedings of the Ninth Conference on Machine Translation*, pages 547–555, Miami, Florida, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

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

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*, 55(11).

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746, Online. Association for Computational Linguistics.

Thibault Sellam, Surafel M. Lakew, Marcos Zampieri, Barry Haddow, and Alexandra Birch. 2021. The multilingual evaluation landscape: A survey of datasets for multilingual machine translation. In *Proceedings of the 2021 Conference on Machine Translation (WMT)*, pages 958–973. Association for Computational Linguistics.

Sam Shapiro, Fesseha Ghidey, Shammur Chowdhury, and 1 others. 2023. Lesan: A multilingual and multimodal platform for low-resource languages in ethiopia. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mgpt: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

Anushka Singh, Ananya Sai, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, and Mitesh Khapra. 2024. How good is zero-shot MT evaluation for low resource Indian languages? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 640–649, Bangkok, Thailand. Association for Computational Linguistics.

Don Snow. 2004. *Cantonese as written language: The growth of a written Chinese vernacular*, volume 1. Hong Kong University Press.

Don Snow. 2008. Cantonese as written standard? *Journal of Asian Pacific Communication*, 18(2):190–208.

Chaoju Tang and Vincent J van Heuven. 2007. Mutual intelligibility and similarity of chinese dialects: Predicting judgments from objective measures. *Linguistics in the Netherlands*, 24(1):223–234.

Sze-Wing Tang, Fan Kwok, Thomas Hun-Tak Lee, Caesar Lun, Kang Kwong Luke, Peter Tung, and Kwan Hin Cheung. 2002. Guide to lshk cantonese romanization of chinese characters. *Hong Kong: Linguistic Society of Hong Kong*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2022. Progress in machine translation. *Engineering*, 18:143–153.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, and 39 others. 2024. AfriMTE and AfriCOMET: Enhancing COMET to

embrace under-resourced African languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5997–6023, Mexico City, Mexico. Association for Computational Linguistics.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Syed Mekael Wasti, Shou-Yi Hung, Christopher Collins, and En-Shiun Annie Lee. 2025. Translationcorrect: A unified framework for machine translation post-editing with predictive error assistance. *Preprint*, arXiv:2506.18337.

Kam-Fai Wong and Xiaodong Zhang. 2017. Building a parallel corpus for english-cantonese machine translation. In *Proceedings of the 2nd Workshop on Asian Translation (WAT)*, pages 85–90. Association for Computational Linguistics.

Rong Xiang, Ming Liao, and Jing Li. 2024. Cantonese natural language processing in the transformers era. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 69–79, Bangkok, Thailand. Association for Computational Linguistics.

Qizhen Yang. 2025. ShUD: the first shanghainese Universal Dependency treebank. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 186–193, Ljubljana, Slovenia. Association for Computational Linguistics.

Foong Ha Yap and Winnie Chor. 2011. Asymmetry in grammaticalization –the case of directional particles in cantonese. 2011 The 5th Conference on Language, Discourse and Cognition (CLDC-5) ; Conference date: 29-04-2011 Through 01-05-2011.

Hongjian Yu, Yiming Shi, Zherui Zhou, and Christopher Haberland. 2024. Machine translation evaluation benchmark for Wu Chinese: Workflow and analysis. In *Proceedings of the Ninth Conference on Machine Translation*, pages 600–605, Miami, Florida, USA. Association for Computational Linguistics.

Lily H Zhang, Hamid Dadkhahi, Mara Finkelstein, Firas Trabelsi, Jiaming Luo, and Markus Freitag. 2025. Learning from others' mistakes: Finetuning machine translation models with span-level error annotations. In *Forty-second International Conference on Machine Learning*.

Xiaoheng Zhang. 1998. Dialect MT: A case study between Cantonese and Mandarin. In *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A Language Terminology

Names and boundaries between languages in China are often fuzzy (Chappell, 2015), with varying conventions across fields. We clarify that this work discusses annotations in the language rather than dialects specific to the city or a region, although the annotations may reflect the prestige dialect that is spoken in Shanghai and Pearl Delta Region, respectively (Chappell, 2015; Eberhard et al., 2023).

Yue or Cantonese? The distinction between Yue and Cantonese can be unclear. While Eberhard et al. (2023) describes Cantonese as an alternate name for Yue, some use it to describe the Guangzhou dialect of Yue (e.g. Matthews and Yip, 2011). In this work, we follow Ethnologue (Eberhard et al., 2023) and previous work in Cantonese MT (Liu, 2022; Hong et al., 2024) to refer to the entire Yue language as Cantonese. We note that NLLB Team et al. (2024) use "Yue Chinese" to describe what we describe as Cantonese in this paper.

Wu or Shanghainese? The distinction between Wu and Shanghainese is much clearer– Shanghainese is a dialect of Wu (Eberhard et al., 2023). In MT, the sole work in Wu Chinese (Yu et al., 2024) does not use the term Shanghainese. We follow such prior work to discuss our annotations in Wu Chinese, rather than Shanghainese.

B Model Selection and `mt` Generation

B.1 Cantonese

The model selection process determined the model family used for the machine translations and the particular parameter size used. Models in the 1 billion parameter range and 6 billion parameter range were investigated, as models with around 1 billion parameters can be run locally on mobile devices, while models in the 6 billion parameter range can be run locally with a mid-range GPU such as an NVIDIA RTX 9060. Improving models within these parameter ranges can increase accessibility

of high quality multilingual models while preserving privacy, as users would be able to prevent third-parties from gathering user data. To compare performance, the first 10 sentences from the English subset of `FLORES+` (Yu et al., 2024) were translated with each of the models into Cantonese. For instruct or chat models, the instruction was left as the model default, as changing the instruction to "translate the following sentence to Cantonese"and simply having the English sentence as a prompt failed to provide Cantonese output.

To determine whether the output was in Cantonese, two strategies were employed. The first was to ensure the output was in Traditional Chinese instead of Simplified Chinese, as Hong Kong Cantonese is written in Traditional characters. The second is to observe whether the output contains Cantonese-specific characters, while avoiding Mandarin-specific characters. If a model's output was not in Cantonese, attempts to re-prompt the model to receive Cantonese output would be taken, such as asking to "translate the following sentence to Yue Traditional Chinese" or to "translate the following sentence to Cantonese with Cantonese features".

The performance of each model was then evaluated with the assistance of an annotator from the annotation team: the best output of each model was then corrected to produce a corrected sentence. Once corrections for all outputs were done for all models, the corrected versions of each sentence were aggregated along with the reference translation to create a set of reference sentences, and each of the models' outputs were evaluated using SacreBLEU (Post, 2018) and ChrF++ (Popović, 2017).

The following models in the 1 billion parameter range were explored: 600M NLLB-200 (Team et al., 2022), 1.5B Qwen 2.5 Instruct (Team, 2024), 1B Llama 3.2 Instruct (Grattafiori et al., 2024), 1.1B Bloomz (Muennighoff et al., 2023), and 1.2B mT0 Large (Muennighoff et al., 2023). However, only NLLB was able to produce Cantonese, even after repeated re-prompting of the other models.

In the 8 billion parameter range, the following models were explored: 7B Qwen Chat (Bai et al., 2023), 8B Llama 3.1 (Grattafiori et al., 2024), and 8B Aya Expanse (Dang et al., 2024). However, Qwen was not able to produce Cantonese, even after repeated re-prompting. So, in the 8 billion parameter range, only Llama 3.1 and Aya were considered.

Upon further investigation, while all models above supported "Chinese", only NLLB-200 and Aya Expanse's training data included data that was explicitly "Cantonese". Llama and Qwen's training data could not be found, and the training data used for other models did not include Cantonese data.

The SacreBLEU and ChrF++ scores of the three models evaluated were as follows: While

Model # Params	NLLB-200 600M	Aya Expanse 8B	Llama-3.1 8B
SacreBLEU	74.5	79.9	82.1
ChrF++	75.0	72.7	82.0

Table 3: Comparison of Translation Models Using SacreBLEU and ChrF++.

NLLB-200 had the lowest average SacreBLEU and ChrF++ scores, the 8B models only did around 10% better in evaluation, while having almost an order of magnitude more parameters.

NLLB-200 was chosen as the model for machine translations for two reasons. The first reason is its smaller size: by virtue of having less parameters, it can be run on a larger variety of machines, making it more accessible to those without dedicated hardware for ML models. The second reason is that its lower average evaluation scores could lead to the creation of a more useful dataset: more errors in the machine translations mean more errors to mark down, meaning there are more examples of what machine translations are doing wrong. However, NLLB-200 is less than 10% worse than Aya 8B and Llama 3.1 8B, meaning the errors being identified are a lot more fine-grained than a model that just produces irrelevant outputs.

B.2 Mandarin

For consistency across languages, we used the same model for Mandarin as we did for Cantonese. As discussed in B.1, 600M NLLB (Team et al., 2022) was selected based on a thorough analysis of model quality and parameter size. Using a single model allows us to maintain consistency in prompts and output formatting across languages. Moreover, NLLB shows decent performances in Mandarin, thus making it an ideal choice for producing the `mt` outputs.

B.3 Wu Chinese

For Wu Chinese, we chose Qwen 2.5 Max (Team, 2024) for producing MT outputs since it remains the only publicly accessible large language

model with stable Wu Chinese ability, apart from DeepSeek, which could not be used due to institutional restrictions. We tried other models, including Llama and NLLB-200, which were unable to produce usable Wu Chinese output even after prompt engineering. We ensured Qwen 2.5 Max was regularly able to produce Wu Chinese translations using basic prompts. Following several rounds of experimentation, we chose the following prompt for producing consistent Wu Chinese translations with Qwen:

> Please translate the following English texts into idiomatic Shanghainese.
> 1. Ensure the translation is accurate.
> 2. Each sentence in English should be matched by a separate sentence in Shanghainese, with the complete meaning preserved.
> For example, for a complex sentence containing multiple clauses (and thus multiple periods), all parts should be translated. For example:
> *"Downhill snowsports, which include skiing and snowboarding, are popular sports involving sliding down snow-covered terrain with skis or a snowboard attached to your feet."*
> This example contains two sentences, and both should be fully translated.

C Annotation Interface

The TranslationCorrect tool (Wasti et al., 2025), a web-based annotation platform designed for machine translation post-editing and error tagging, is used for our annotations. As shown in Figure 3, for each example, annotators can see the `src`, `mt`, and `ref` sentences, and whether it has been annotated.

The tool allows annotators to highlight spans in the machine-translated output and assign error severity levels and error categories from predefined label sets (See Figure 4). All annotation data were stored in our database and later exported for analysis.

D Annotation Labels

We adapted our error severity and category definitions based on MQM guidelines (Burchardt, 2013) and the AfriCOMET framework (Wang et al., 2024), with modifications informed by language-specific characteristics of Sinitic Languages. More detailed information on the guidelines and modifications can be found in E. After multiple rounds of pilot annotation and qualitative analysis as described in Section 3, we refined the labels to better capture common translation issues observed in our data. Table 4 and Table 5 show the definition of severity levels and error categories, respectively.

E Additional Details on Annotation Guidelines

While we base our annotation guidelines on prior work in MQM (Burchardt, 2013) and AfriCOMET (Wang et al., 2024), we make several adjustments during the pilot and main annotation stages described in Section 3.2. We describe such adjustments in detail below.

Inappropriate Proper Nouns. For proper nouns such as specific names of people and places, if their translations are not the same as in the reference sentences, then annotators should classify them as Spelling errors, instead of Mistranslation. This guideline is based on the assumption that the reference translations from Flores+ (Goyal et al., 2022; NLLB Team et al., 2024; Yu et al., 2024) are of high quality, as they were created by professional human translators. The purpose is to distinguish between general semantic mistranslation and inappropriate translations of the proper nouns, which may vary historically or regionally. Labeling such differences as Spelling errors helps the future work, as models could better learn about the boundaries between semantic-level errors and surface-level variations.

Full-width form punctuations. In Sinitic writing systems, full-width punctuation marks are commonly used. However, the 600M NLLB-200 model (Team et al., 2022) consistently outputs sentences with punctuations in half-width forms. The annotators were instructed to skip such annotation cases, as the goal is to distinguish between misuse of punctuations and incorrect forms of them. In the future QA stage, we plan to consult linguistic experts and may introduce a new error category, if this issue will significantly affect translation quality according to the experts.

Omission of quality score evaluations. In our dataset, we focus exclusively on the error span positions, error type, and severity, rather than assigning overall quality scores to the machine translations. Our design reflects our goal of helping the models better identifying and classifying errors, instead of performing quality assessments. Omitting quality scores also improves the translation efficiency and helps avoid subjectivity and disagree-

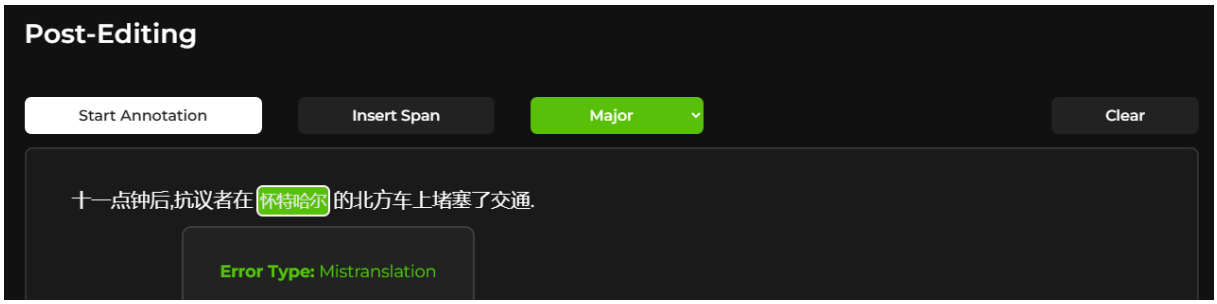

Figure 4: Annotators could select spans in the machine-translated output, assign an error type (e.g., "Mistranslation"), and specify the severity level (e.g., "Major"). In this example, the span has been tagged as a major mistranslation.

Severity Level	Definition
Major	The error introduced causes a significant change in the meaning of the translated sentence.
Minor	The error does not change the core meaning of the translated sentence, but introduces a slight issue affecting fluency or readability.

Table 4: Definitions of severity levels used for error annotation.

ment over score interpretation, thus improving consistency and efficiency in the annotation process.

Error Category	Definition
Addition	The highlighted span in the translation corresponds to information that does not exist in the source text.
Omission	The highlighted span corresponds to content manually inserted by the annotator into the translation, representing information present in the source text but missing from the original MT output.
Mistranslation	The highlighted span in the translation does not have the exact same meaning as the corresponding span in the source segment.
Untranslated	The highlighted span in the translation is a copy of the corresponding span in the source segment, but should have been translated into the target language.
Grammar	The highlighted span corresponds to issues related to grammar or syntax in the translated text, excluding spelling and orthography.
Spelling	The highlighted span corresponds to spelling issues. Mistranslations of names (e.g., locations, people) are also categorized as spelling errors.
Typography	The highlighted span corresponds to issues related to punctuation or diacritics, except omission of punctuations.
Unintelligible	The exact nature of the error cannot be determined, indicating a major breakdown in fluency.

Table 5: Definitions of error categories used for annotation.