IndoDialect: A Multi-Dialect Conversational Dataset for Indonesian Local Languages

Anonymous ACL submission

Abstract

Dialectal variation is a defining characteristic of many low-resource languages, yet most NLP approaches overlook this complexity—often assuming uniformity or arbitrarily selecting a single variant without reporting which dialect is used. This assumption severely limits the applicability of language technologies in real-world communication, especially for languages without unified orthography where dialects could differ in vocabulary and morphology. In this paper, we present the first comprehensive study of dialect processing for Indonesian local languages, a linguistically rich and underrepresented setting. We introduce **IndoDialect**, a high-quality conversational dataset comprising **14.2K conversations** across **four major local languages** and **eleven dialects**, including five Javanese, two Minangkabau, two Sundanese, and two Dayak dialects. Building on this resource, we define three core tasks: (i) dialect-level language identification, (ii) machine translation across dialects, and (iii) dialect-controlled generation. Our experiments reveal that state-of-the-art multilingual models perform poorly in few-shot scenarios, excel at dialect identification after fine-tuning, but struggle with machine translation and dialect-controlled generation[1].

1 Introduction

Low-resource NLP has attracted growing attention for its potential to democratize language technologies (Cahyawijaya et al., 2023; Nigatu et al., 2024). However, existing approaches often simplify the complexity introduced by dialectal variation in these languages (Faisal et al., 2024). In practice, many low-resource languages lack standardized forms, resulting in substantial dialectal diversity that dominates real-world communication (Nerbonne, 2003; Koto and Koto, 2020; Aji et al., 2022). This diversity poses significant challenges for NLP systems, as orthographic conventions are often inconsistent and dialects can differ markedly in vocabulary, morphology, and syntax (Ziems et al., 2022).

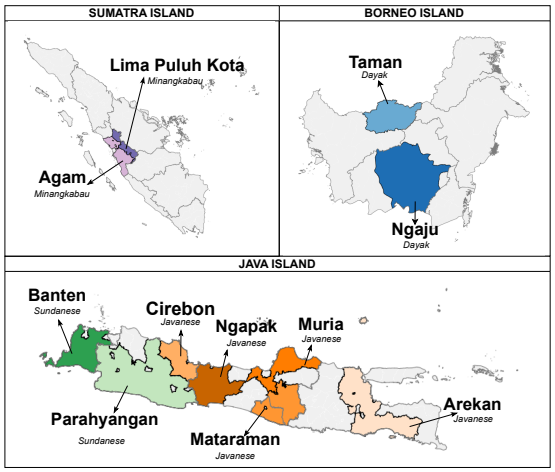

Figure 1: Dialect coverage in IndoDialect, showing 11 dialects from 4 Indonesian local languages.

Despite its importance, dialect processing remains underexplored in the context of low-resource languages. Prior work has extensively focused on dialects of medium– to high–resource languages such as English (Held et al., 2023; Xiao et al., 2023) and Arabic (Salameh et al., 2018; Bouamor et al., 2018), where resources and standardized benchmarks are more readily available. Shared tasks such as VarDial (Zampieri et al., 2017; Gaman et al., 2020) and AmericasNLP (Mager et al., 2021; Ebrahimi et al., 2023) have further advanced research on dialectal variation across languages. Existing studies have also explored dialect-to-standard translation and normalization including Japanese (Abe et al., 2018), Finnish (Partanen et al., 2019), and multilingual scenarios (Kuparinen et al., 2023; Bawden et al., 2022). However, Indonesian local languages remain under-resourced

[1]The code and data will be released via an anonymized repository.

at the dialect level, where linguistic fragmentation is widespread and dedicated NLP resources are scarce. This gap limits the inclusiveness and applicability of current NLP systems, particularly in multilingual settings where dialects play a central role in everyday communication.

Indonesia vividly illustrates this challenge. With over 700 languages spoken across the archipelago (Eberhard et al., 2021), dialectal diversity is extensive yet remains poorly documented. Linguistic studies have revealed substantial variation within major Indonesian languages, including Jambi Malay (Anderbeck, 2008), Javanese (Kartikasari et al., 2018), and Balinese (Maharani and Candra, 2018). These differences are strongly influenced by geographic and cultural factors (Van-der Klok, 2015), creating significant obstacles for language technologies that are typically trained on a single standardized form (Aji et al., 2022). Despite their widespread use in daily communication, dialects of Indonesian local languages have an extremely limited digital footprint. Addressing this variation is therefore essential for building robust and inclusive NLP systems that capture real-world linguistic practices.

This paper presents the first comprehensive study on dialect processing for Indonesian local languages, focusing on three tasks: dialect identification, machine translation, and dialect-controlled generation. Since dialectal variation is most evident in conversation, we build high-quality conversational datasets **written from scratch by native speakers** for authenticity and diversity. The study covers four major local languages and eleven dialects across three of Indonesia's largest islands—Java, Sumatra, and Borneo (Figure 1). These languages include Sundanese ($\approx$ 42M speakers), Javanese ($\approx$ 80M speakers), Minangkabau ($\approx$ 8M speakers), and Dayak ($\approx$ 4.2M speakers). To our knowledge, this is the first large-scale dialect-level benchmark for Indonesian local languages, addressing a critical need for inclusive and dialect-aware NLP systems.

Our contributions can be summarized as follows: **(i)** We introduce **`IndoDialect`**, the first dataset of Indonesian local languages at the dialect level, comprising 14.2K conversational sentences spanning five Javanese, two Minangkabau, two Sundanese, and two Dayak dialects, all manually written by native speakers. **(ii)** We define **three core tasks**—dialect-level language identification, machine translation across dialects-specific variation, and dialect-controlled generation—and evaluate models under both few-shot and fine-tuning settings. For dialect identification, multilingual models perform poorly in few-shot scenarios but achieve strong performance after fine-tuning. For machine translation, we find that phrase-based statistical approaches outperform neural models, whereas dialect-controlled generation remains challenging, as evidenced by low dialect identification scores in controlled settings. **(iii)** We conduct a thorough analysis across all three tasks, including model performance across dialects, sensitivity to text length, and detailed error analysis to uncover key challenges and opportunities for future research.

2 Related Work

Efforts in Indonesian NLP Several studies in Indonesian NLP have explored many tasks, including POS Tagging (Dinakaramani et al., 2014), NER (Gunawan et al., 2018), sentiment analysis (Wicaksono et al., 2014), summarization (Koto et al., 2020a), question answering (Fikri and Purwarianti, 2012), machine translation (Hermanto et al., 2015), and speech recognition (Zahra et al., 2009). Earlier reviews from Aji et al. (2022) note that many of these works did not release their datasets, which made the findings hard to reproduce. More recent projects have introduced shared benchmarks, such as IndoNLU (Wilie et al., 2020) and IndoLEM (Koto et al., 2020b), to provide common evaluation data and support better reproducibility and comparison across models. However, these resources remain focused on language-level modeling, and support for dialect-level processing is still limited.

Dialect Processing Shared tasks have played a central role in advancing dialect NLP research. VarDial (Zampieri et al., 2017; Gaman et al., 2020) focuses primarily on dialect identification for closely related language varieties, while AmericasNLP (Mager et al., 2021; Ebrahimi et al., 2023) highlights the challenges of low-resource machine translation for indigenous languages. More recently, DIALECTBENCH (Faisal et al., 2024) introduces a large-scale benchmark spanning 281 varieties across 10 NLP tasks by aggregating existing dialect resources. However, none of these efforts focus on Indonesian local languages. In contrast, `IndoDialect` is, to the best of our knowledge, the first dataset specifically designed for Indonesian local dialects, featuring a conversational dataset

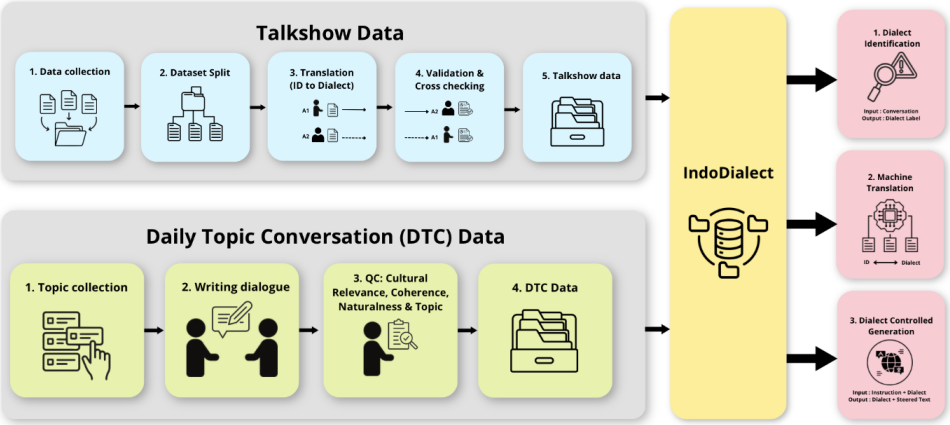

Figure 2: Overview of the **IndoDialect** data construction pipeline. The top panel illustrates the creation of the *Talkshow Data*, while the bottom panel shows the dialogue creation options based on selected topics for *DTC Data*.

written by native speakers.

Dialect identification and dialect-level machine translation have been studied across multiple languages, but dialect-controlled generation remains limited. Prior work shows that dialect identification is largely driven by surface lexical and orthographic cues (Zampieri et al., 2017), while studies on dialect translation and normalization demonstrate that character-level and phrase-based approaches remain competitive for closely related varieties (Kuparinen et al., 2023; Bawden et al., 2022). However, most existing work relies on short or isolated texts, with limited support for extended conversational settings. In the Indonesian context, these capabilities are largely absent: existing resources do not focus on Indonesian local languages or jointly support dialect identification and dialect-level machine translation. Through IndoDialect, we address this gap by introducing one of the first datasets for Indonesian dialect identification and dialect-level machine translation, and by presenting one of the earliest studies on dialect-controlled generation in this setting.

3 IndoDialect

3.1 Scope of Languages and Dialects

To address the scarcity of dialectal resources, we introduce **IndoDialect**, a high-quality conversational dataset designed to capture dialectal diversity in Indonesian local languages. As shown in Figure 1, **IndoDialect** focuses on **four major local languages**–Javanese, Sundanese, Minangkabau, and Dayak–representing different large islands in Indonesia: Javanese and Sundanese from Java, Minangkabau from Sumatra, and Dayak from Borneo.

Dialect	Sentence
English	I was serious when I said I would not leave you again.
Indonesian	Aku serius waktu bilang takkan meninggalkanmu lagi.
Arekan	Aku serius pas ngomong gak bakal ninggalno awakmu maneh.
Mataraman	aku tenanan wektu ngomong ora bakal ninggalke kowe maneh.
Muria	Aku tenan pas kondo ora bakal ninggalno kowe maneh.
Ngapak	Nyong serius pas ngomong ora bakal ninggalna ko maning.
Cirebon	Isun serius waktu ngomong beli arep ninggalnang sira maning.

Table 1: Example of variation across Javanese dialects for a sentence in IndoDialect.

The dataset spans **eleven dialects**, categorized according to *Ethnologue*,[2] including **five** Javanese dialects (Arekan, Mataraman, Muria, Ngapak, and Cirebon), **two** Sundanese dialects (Parahyangan and Banten), **two** Minangkabau dialects (Agam and Lima Puluh Kota), and **two** Dayak dialects (Ngaju and Taman). All dialects are represented in Latin script and reflect contemporary spoken usage, covering both closely related varieties with substantial lexical overlap and more distinct dialects. Table 1 illustrates an example of variation across Javanese dialects in IndoDialect. We also provide a detailed **data statement** (Bender and Friedman, 2018) for comprehensive information about our dataset in Appendix A.

3.2 Dataset Construction

Figure 2 illustrates the construction process of IndoDialect, which consists of two fully human-written pipelines. The first pipeline involves manually translating Indonesian talk show subtitles from YouTube OpenSubtitles[3] into multiple dialects. The second pipeline simulates natural conversations through a chat platform, where two native

[2] https://www.ethnologue.com/
[3] https://www.opensubtitles.org

		Talkshow Data				DTC Data						
Lang.	**Dialect**	**#Diag.**	**Qual. (%)**	**#Wrd**	**#Chr**	**#Diag.**	**Cult. (%)**	**Coh. (%)**	**Nat. (%)**	**Top. (%)**	**#Wrd**	**#Chr**
Indonesian	id	1000	100	52,931	279,606							
Javanese	are	1000	100	52,528	269,572	200	100	95	95	100	14,431	69,234
	mat	1000	100	52,080	284,530	200	100	100	100	100	13,824	75,446
	mur	1000	100	51,554	267,006	200	100	95	95	100	9,951	45,067
	nga	1000	100	54,778	281,690	200	100	95	95	100	16,327	77,907
	cir	1000	100	53,783	283,225	200	100	95	95	100	13,316	62,509
Sundanese	par	1000	99	55,470	295,130	200	100	95	95	100	22,541	108,456
	btn	1000	99	55,894	291,657	200	85	95	95	100	10,925	50,685
Minangkabau	agm	1000	100	55,429	273,256	200	100	100	100	100	12,684	54,882
	lkt	1000	98	54,492	267,694	200	100	100	100	100	12,806	56,612
Dayak	nij	1000	100	58,951	272,708	200	100	95	95	95	16,122	67,890
	tmn	1000	100	49,538	283,467	200	85	100	100	100	11,396	63,479
AGG.		**12,000**	**99.6**	**647,428**	**3,349,541**	**2,200**	**97.3**	**96.8**	**96.38**	**99.5**	**154,323**	**732,167**

Table 2: Dataset statistics and quality evaluation by language and dialect. For the *Talkshow Data*, we report the number of dialogues, overall quality scores (Qual), and length statistics. For the *DTC Data*, we report the number of dialogues (#Diag), human evaluation scores for cultural relevance (Cult), coherence (Coh), naturalness (Nat), and topic coverage (Top), along with length statistics; the DTC dataset does not include an Indonesian (`id`) split.

speaker annotators engage in dialogues on daily topics using their native dialects. Both pipelines incorporate rigorous quality control procedures, described in detail in a later section.

To ensure high-quality data, we conduct human evaluation across all eleven dialect classes in `IndoDialect`: `jv_mur` (Javanese Muria), `jv_mat` (Javanese Mataraman), `jv_are` (Javanese Arekan), `jv_nga` (Javanese Ngapak), `jv_cir` (Javanese Cirebon), `su_par` (Sundanese Parahyangan), `su_btn` (Sundanese Banten), `min_agm` (Minangkabau Agam), `min_lkt` (Minangkabau Lima Puluh Kota), `day_nij` (Dayak Ngaju), and `day_tmn` (Dayak Taman). This evaluation focuses on cultural relevance, coherence, and naturalness to guarantee the authenticity of the dataset.

Annotator Recruitment We employ at least two expert annotators who are native speakers of each dialect. To ensure quality, annotators are well-educated (at least senior high school graduate) and have lived in the dialect region for at least ten years, with at least one parent is native from that region. Candidates are screened through translation tasks, region-specific questions, and short dialogue and writing samples. Selected annotators participate in a one-hour training session before annotation begins where we explain annotation procedure and expectations. All annotators were recruited via Upwork, provided informed consent, and were compensated accordingly. Further details and guidelines are provided in Appendix A and Appendix B.

Data-1: Talkshow We follow the annotation strategy of Winata et al. (2023) by manually translating Indonesian text into 11 dialects, employing two native speakers for each dialect. Our source data comes from OpenSubtitles of Indonesian talk shows on YouTube. These talk shows are licensed under Creative Commons, allowing us to use them for our dataset. The original dataset consists of 1,000 dialogues, each with an average of 5-turns per dialogue. We split the data into two sets and assign each set to a different annotator, resulting in two independent translations for every dialect.

For quality control, annotators review a portion of each other's work to validate translation accuracy. To ensure careful reading during validation, we randomly modify 10% of the dialogues prior to review. Perturbations include four types: word deletions (e.g., removing negation markers), incorrect word substitutions, full utterance removals, and redundant word insertions that affect fluency. Annotators are expected to detect and correct all modifications; dialogues that fail to meet quality standards are discarded. This procedure helps identify errors and maintain consistency. As shown in Table 2, the average quality score for the Talkshow Data is 99.6%, meaning that annotators successfully corrected 99.6% of the intentionally altered samples—demonstrating rigorous validation and high-quality translations across all dialects. The final Talkshow dataset is a parallel corpus comprising approximately 11,000 dialectal dialogues aligned with 1,000 Indonesian source dialogues.

Data-2: Daily Topic Conversations (DTC) We simulate natural conversations between two annotators using a chat platform, based on daily topics drawn from `IndoCulture` (Koto et al., 2024). Each annotator pair writes a five-turn dialogue for each assigned topic. For every dialect, we collect 200 dialogue pairs, resulting in 200 natural conversations per dialect. Annotators are instructed to preserve cultural relevance in accordance with local norms and practices. Examples include Javanese greetings such as *sungkeman*, Sundanese wedding customs like *nyawer*, Dayak marriage requirements such as *palaku*, and Minangkabau engagement traditions like *batimbang tando*.

Although the dialogues are natural and high-quality, we conduct an additional quality check through manual evaluation by a third native speaker. This evaluation assesses cultural relevance, coherence, naturalness, and topic coverage on a scale of 0–100. As shown in Table 2, DTC achieves high scores above 95 across all aspects, indicating excellent quality. The final DTC dataset consists of 2.2K dialogues, totaling approximately 154K words and 732K characters, and is not structured as a parallel corpus.

3.3 Statistical Analysis

Table 2 shows that the Talkshow parallel dataset displays varied dialogue lengths across dialects, despite being a parallel corpus. The average number of words per dialogue in Indonesian is 52.9. Within Javanese, the Ngapak (`jv_nga`) dialect tends to have longer dialogues compared to other Javanese dialects. In contrast, among Sundanese and Minangkabau dialects, no clear differences are observed. For Dayak, the Ngaju (`day_nij`) dialect generally contains more words than Taman (`day_tmn`). This pattern is also consistent in the DTC dataset, with Sundanese being an exception.

We further analyze dialect similarity within each language using word-level Levenshtein distance, where values closer to 0 indicate higher similarity. Figure 3 reveals distinct patterns across dialect groups. Javanese dialects exhibit moderate and relatively uniform similarity, while Sundanese and Minangkabau dialects tend to be more similar, as indicated by lower distance scores. In contrast, Dayak dialects show higher distance values, reflecting clearer lexical differentiation between the two varieties.

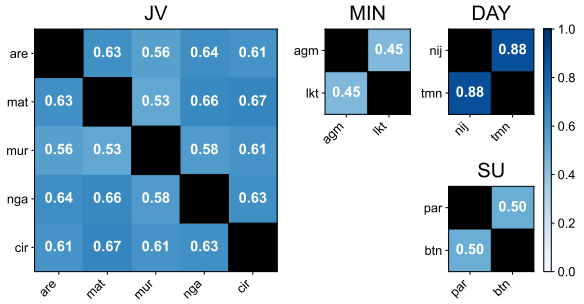

Figure 3: Word-level Levenshtein distance between dialect pairs.

4 Experimental Setup

We design three evaluation settings to investigate dialect processing in `IndoDialect`: (i) **Dialect Identification**, (ii) **Machine Translation**, and (iii) **Dialect-Controlled Generation**. These tasks capture key challenges in dialect-aware NLP. Dialect identification measures a model's ability to distinguish closely related language varieties, while machine translation across dialects evaluates whether models can generate dialect-specific forms rather than defaulting to standardized or dominant variants. Dialect-controlled generation further assesses whether models can produce responses in the target dialect under explicit control, reflecting their ability to generate contextually appropriate dialectal usage.

All experiments use the `IndoDialect` dataset, which is split into training, validation, and test sets with a 70:15:15 ratio, ensuring that the test set remains fully isolated. Prompt templates for few-shot settings are provided in Appendix C, and hyperparameter details are in Appendix D.

Dialect Identification Dialect identification is formulated as a multi-class classification task, where the model predicts one of eleven dialect labels, along with Indonesian (`id`) and an additional `others` class given a multi-turn dialogue. We evaluate on both Talkshow and DTC data to capture performance across controlled and conversational settings, using macro F1 to assign equal weight to all dialect classes. We evaluate traditional classifiers—Naïve Bayes (NB), SVM, and Logistic Regression (LR)—trained on TF–IDF features in a fully supervised setting, and fine-tune encoder-based models including IndoBERT (Koto et al., 2020b) and XLM-R (Conneau et al., 2020) (`BASE` and `LARGE` variants) on the Talkshow data before applying them to DTC. Large language models—Qwen3 (Team, 2025), Llama 3.1 (Grattafiori et al.,

Model	ID	Javanese					Sundanese		Minangkabau		Dayak		AVG
	id	are	mat	mur	nga	cir	par	btn	agm	lkt	nij	tmn	
Talkshow Data													
GPT-5	0.998	0.192	0.250	0.451	0.530	0.696	0.569	0.624	0.568	0.525	0.960	0.844	0.601
Llama 3.1	0.622	0.230	0.000	0.026	0.007	0.000	0.332	0.078	0.353	0.173	0.396	0.000	0.185
NB	**1.000**	**1.000**	0.818	0.792	**0.990**	**0.988**	0.500	**0.714**	**0.978**	**0.978**	**1.000**	**1.000**	**0.896**
IndoBERT$_{\text{LARGE}}$	0.993	0.982	**0.830**	**0.809**	0.978	0.966	**0.520**	0.675	0.965	0.966	0.995	**1.000**	0.890
DTC Data													
GPT-5		0.288	0.249	0.502	0.345	0.583	0.584	0.755	0.529	0.408	0.981	0.795	0.547
Llama 3.1		0.249	0.000	0.000	0.010	0.005	0.342	0.022	0.427	0.204	0.604	0.005	0.170
NB		**0.998**	**0.966**	**0.968**	**0.998**	**0.996**	0.923	0.931	**0.976**	**0.976**	**1.000**	**1.000**	**0.976**
IndoBERT$_{\text{LARGE}}$		0.964	0.903	0.930	0.989	0.986	**0.933**	**0.939**	0.974	0.974	**1.000**	**1.000**	0.963

Table 3: Dialect identification performance of best models across dialects on F1 score for Talkshow and DTC datasets. **bold text** = best performing model, underlined = second-best performing model.

2024), and closed-weight systems from OpenAI and Cohere (Cohere et al., 2025)—are evaluated via few-shot in-context learning on both datasets.

Machine Translation The machine translation task involves translating between Indonesian and each dialect in both directions (Indonesian→Dialect and Dialect→Indonesian). We use utterance-level pairs from the Talkshow parallel corpus, excluding utterances shorter than three words. Translation quality is evaluated using SacreBLEU (Post, 2018) and chrF. We include Copy and phrase-based statistical machine translation (PBSMT) (Lample et al., 2018) as baselines. IndoGPT, IndoBARTv2 (Cahyawijaya et al., 2021), and mT5$_{\text{BASE}}$ (Xue et al., 2021) are fine-tuned, while Qwen3, Llama 3.1, and closed-weight models are evaluated via supervised fine-tuning with LoRA or few-shot prompting.

Dialect-controlled Generation This task evaluates a model's ability to generate responses in a target dialect given a dialogue context and an Indonesian source utterance. We use the Talkshow data, treating each Indonesian utterance as input and its dialectal variant as reference. Models are evaluated under few-shot settings using SacreBLEU, chrF, and dialect identification accuracy, where the latter measures whether outputs are correctly classified into the target dialect by a fine-tuned IndoBERT$_{\text{LARGE}}$ model.

5 Result and Analysis

5.1 Dialect Identification

Best-Performing Model Table 3 reveals clear performance gaps across models. Generative approaches perform poorly, with 3-shot GPT-5 averaging 0.601 on Talkshow and 0.547 on DTC, while fine-tuned Llama-3.1 is even worse (0.185 and 0.173). In contrast, classical models like Naïve Bayes achieve strong results (0.896 and 0.976), and IndoBERT$_{\text{LARGE}}$ dominates with near-perfect scores (0.890 and 0.963). This pattern is consistent with prior work on dialect identification, where traditional models based on character- and word-level features remain competitive benchmarks for closely related language varieties (Zampieri et al., 2017; Malmasi and Zampieri, 2017). Generative models struggle most with closely related Javanese dialects (`are`, `mat`, `mur`), where confusion is common, while Dayak varieties are consistently identified with high accuracy. Performance drops slightly on DTC due to informal spelling and cultural expressions, but overall, encoder-based and classical models far outperform generative ones.

Effect of Shared Vocabulary To examine the impact of lexical similarity on dialect identification, we compute the Pearson correlation (ρ) between shared-vocabulary overlap (Fig. 4) and instance-level correctness for each dialect using IndoBERT$_{\text{LARGE}}$ on the Talkshow data (Table 3). Negative correlations indicate that higher lexical overlap is associated with increased confusion. This effect is most pronounced for Javanese `mat` ($\rho = -0.284$, F1: 0.830) and `cir` ($\rho = -0.266$, F1: 0.966), where dialects such as `mat` and `mur` share approximately 53% vocabulary yet must be distinguished as separate classes. In contrast, Dayak `tmn` shows perfect prediction, reflecting clearer lexical separability, while Sundanese `btn` ($\rho = 0.157$) exhibits a positive correlation, suggesting that shared vocabulary can aid prediction in some cases. These results show that high lexical overlap reduces discriminative cues, increasing the risk that models collapse closely related dialects into a single dominant variety. More detailed analysis can be found in Appendix E.

ID → Dialect$_i$								
Model	JAV		SUN		MIN		DAY	
	B	C	B	C	B	C	B	C
Copy	4.75	35.02	2.65	32.49	5.27	41.05	3.06	32.81
PBSMT	**31.97**	**62.83**	**21.82**	**54.18**	**32.93**	**65.65**	**18.78**	**56.06**
GPT-5	18.56	53.37	24.78	57.91	13.94	50.50	2.50	33.31
GPT-5-mini	14.06	48.36	20.50	54.49	7.76	43.68	2.01	31.24
CommandA	13.91	46.56	11.69	44.26	7.80	44.47	2.28	31.80
Qwen3-8B	23.84	56.80	15.72	49.38	25.42	59.81	8.96	43.13
Llama-3.1-8B	24.21	57.09	17.53	50.70	25.81	60.90	10.51	45.70
IndoGPT	12.11	46.11	10.58	41.38	14.67	42.20	5.46	30.07
IndoBARTv2	31.60	62.18	21.04	52.91	24.82	55.47	17.24	50.33
mT5$_B$	30.57	62.14	21.46	54.43	19.01	49.10	13.05	37.19

(a) Indonesian → Dialect

Dialect$_i$ → ID								
Model	JAV		SUN		MIN		DAY	
	B	C	B	C	B	C	B	C
Copy	4.75	35.42	2.63	33.43	5.23	40.34	3.07	33.25
PBSMT	**36.66**	**66.62**	23.37	57.03	37.01	**68.00**	21.60	55.48
GPT-5	34.09	**67.98**	**25.73**	**61.46**	25.77	61.02	9.02	40.86
GPT-5-mini	25.62	61.71	20.15	58.06	18.37	53.07	5.62	36.45
CommandA	30.41	62.69	19.83	54.97	22.33	54.44	8.49	38.92
Qwen3-8B	26.01	57.59	17.05	50.60	26.20	58.15	11.15	42.22
Llama-3.1-8B	22.20	53.66	14.92	47.33	23.77	55.10	10.67	40.83
IndoGPT	13.21	43.17	15.06	33.37	11.39	32.32	5.76	22.11
IndoBARTv2	36.54	64.69	23.79	53.93	35.52	61.01	21.46	47.20
mT5$_B$	33.70	63.81	23.19	55.25	**37.67**	63.36	**26.10**	46.91

(b) Dialect → Indonesian

Table 4: Language-group-level SacreBLEU (B) and chrF (C) scores for bidirectional translation. (a) shows Indonesian-to-dialect performance, while (b) shows dialect-to-Indonesian performance.

Model	**JV**			**MIN**			**SU**			**DAY**		
	BLEU	**chrF**	**DiA**	**BLEU**	**chrF**	**DiA**	**BLEU**	**chrF**	**DiA**	**BLEU**	**chrF**	**DiA**
OpenAI/GPT-5	8.03	**27.17**	23.73	6.31	**25.89**	23.73	7.16	**27.19**	54.92	6.44	**25.02**	17.00
OpenAI/GPT-5-mini	4.21	23.43	19.52	3.58	22.93	19.52	4.00	24.74	49.42	3.19	22.05	18.75
Cohere/CommandA	7.37	20.20	19.52	5.61	19.14	19.52	6.30	20.16	44.39	5.28	17.97	17.69
Llama-3.1-8B-Instruct	19.27	23.60	16.93	19.31	25.17	16.93	18.46	25.40	48.94	17.56	22.20	32.00
Qwen3/Qwen3-8B	**20.41**	25.10	36.13	**19.99**	25.64	36.13	**19.38**	26.07	46.60	**20.02**	24.51	**97.32**
IndoGPT	0.52	14.72	**84.69**	0.37	13.82	**84.69**	0.54	14.58	**75.51**	0.41	13.05	92.11
IndoBARTv2	1.05	20.49	61.87	0.84	17.79	61.87	1.08	20.56	64.58	1.15	19.54	90.45
mT5_base	15.46	21.29	60.90	19.48	25.23	60.90	17.28	22.43	38.69	19.16	23.51	94.56
Average	9.54	22.00	40.41	9.44	21.95	40.41	9.28	22.64	52.88	9.15	20.98	57.48

Table 5: Dialect-controlled generation results aggregated by language group. Scores report sacreBLEU (BLEU), chrF, and dialect identification accuracy (DiA). **bold text** = best scores, underlined = second-best scores.

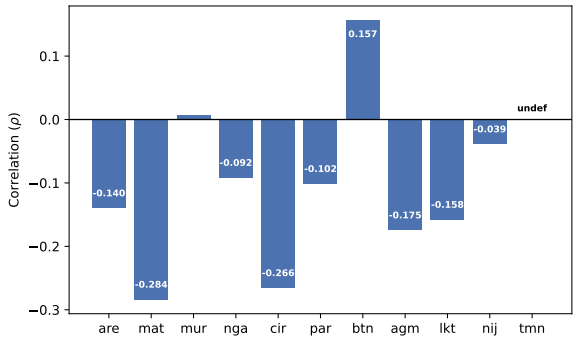

Figure 4: Correlation between shared-vocabulary overlap and instance-level dialect identification correctness.

5.2 Machine Translation

Indonesian-to-dialect Table 4a shows that translation performance is strongly influenced by the similarity between Indonesian and the target dialect. The Copy baseline achieves non-trivial scores for Minangkabau and some Javanese varieties, reflecting lexical overlap with Indonesian. PBSMT delivers the best overall results, with particularly strong performance on Minangkabau (BLEU: 32.93, chrF: 65.65) and Javanese (BLEU: 31.97, chrF: 62.83). Neural models show varied performance: IndoBARTv2 performs competitively across Javanese and Minangkabau, while mT5 achieves solid results on Sundanese. Large generative models (GPT-5, Qwen3, Llama-3.1) consistently underperform relative to supervised baselines, with the largest gaps on lexically distant Dayak dialects (GPT-5: BLEU 2.50, chrF 33.31). These results are consistent with prior findings that phrase-based statistical methods remain effective for translation between closely related language varieties under limited data conditions (Bawden et al., 2022; Lample et al., 2018).

Dialect-to-Indonesian As shown in Table 4b, translating from dialects into Indonesian is generally easier and more stable, reflecting Indonesian's status as a high-resource standardized language. PBSMT achieves the strongest results for Javanese (BLEU: 36.66, chrF: 66.62) and Minangkabau (BLEU: 37.01, chrF: 68.00). Neural models IndoBARTv2 and mT5 are closely competitive, with mT5 achieving the best results on Minangkabau (BLEU: 37.67) and Dayak (BLEU: 26.10). Large generative models perform notably better in this direction than in ID-to-dialect translation—

GPT-5 achieves strong chrF scores on Javanese (67.98) and Sundanese (61.46)—suggesting that models can understand dialectal input more reliably than they can generate dialect-specific forms. However, reference-based metrics like BLEU and chrF rely on single-reference surface overlap and may underestimate the quality of outputs that use valid alternative dialectal expressions, a limitation particularly relevant for generative models in this low-resource setting.

Human Evaluation To better understand translation errors beyond automatic metrics, we conduct a human evaluation of PBSMT outputs. Annotators label errors into five categories, with detailed definitions provided in the Appendix F. The error shown in Table 6 distribution varies across dialects. `jv_mat` shows the highest rates of Addition (62%) and Omission (79%), indicating frequent content modification. `day_nij` exhibits the most Dialect Errors (61%) and substantial Word Order Errors (44%), reflecting difficulty in preserving dialect-specific structure. `su_par` shows notable Morphology Errors (37%), while `min_agm` has lower overall errors but still significant dialect mismatches (50%). These results suggest that dialect distance and structural differences influence translation errors.

Dialect	Add	Omm	WO	DO	MO
day_nij	29.17	33.33	**44.44**	61.11	**40.28**
jv_mat	**62.28**	**78.89**	24.17	**73.33**	23.61
min_agm	27.03	20.27	36.49	50	14.87
su_par	32.86	50	25.71	35.71	37.14

Table 6: Human evaluation of PBSMT outputs by dialect. Values are in percentages. Errors include Addition (Add), Omission (Omm), Word Order (WO), Dialect (DO), and Morphology (MO); categories may overlap.

5.3 Dialect-controlled Generation

Table 5 shows that most models struggle to produce utterances faithful to the intended dialect. Qwen3 achieves the highest BLEU scores across all four language groups, and GPT-5 obtains the highest chrF scores, yet their dialect identification accuracy (DiA) remains modest, ranging from 17.00 to 54.92 for GPT-5 and from 36.13 to 97.32 for Qwen3. Qwen3 is a notable outlier on Dayak, reaching 97.32 DiA, but its dialect control on Javanese, Minangkabau, and Sundanese stays between 36 and 47, indicating that strong surface overlap with the reference does not imply strong dialect fidelity. In contrast, IndoGPT achieves the strongest dialect control overall, with DiA between 75.51 and 92.11 across all four languages, likely due to fine-tuning, but its generations collapse into repetitive and degenerate outputs, yielding BLEU below 1.1 and chrF below 15. Encoder–decoder models such as IndoBARTv2 and mT5_base offer a more balanced profile, producing coherent outputs with moderate BLEU and chrF while maintaining competitive DiA, with IndoBARTv2 in particular achieving the second-best DiA on JV (61.87), MIN (61.87), SU (64.58), and DAY (90.45).

Surface fidelity versus dialect realization. BLEU, chrF, and DiA capture complementary aspects of dialect-controlled generation. GPT-5 attains the highest chrF across all four language groups yet records the lowest DiA on Dayak (17.00) and modest DiA on Javanese and Minangkabau (23.73), indicating fluent outputs that normalize toward standard Indonesian. IndoGPT shows the opposite pattern, with the strongest DiA overall (75.51–92.11) but collapsed BLEU ($<$1.1) and chrF ($<$15), suggesting the classifier latches onto dialect-specific lexical markers even in repetitive or malformed outputs. Qwen3 sits between these extremes, with the highest BLEU but DiA of 36–47 on Javanese, Minangkabau, and Sundanese. These patterns show that overlap-based metrics alone are insufficient for evaluating dialect-controlled generation.

6 Conclusion and Future Work

We presented **`IndoDialect`**, the first dialect-level benchmark for Indonesian local languages, with 14.2K conversational sentences across four languages and eleven dialects, manually written by native speakers. Across three tasks, we find that encoder-based and classical models excel at dialect identification while large generative models lag behind; phrase-based statistical methods remain competitive for machine translation; and dialect-controlled generation exposes a persistent trade-off, where fluent models normalize toward standard Indonesian and dialect-faithful models collapse into degenerate outputs. Future work includes expanding dialect coverage to additional Indonesian languages, scaling to spoken and code-mixed varieties, and developing approaches that better handle orthographic variation and low-resource conditions. `IndoDialect` also opens avenues for analyzing dialectal representation in large-scale corpora such as `Common Crawl`, supporting more equitable NLP for linguistically diverse communities.

Limitations

IndoDialect covers eleven dialects across four Indonesian local languages, but does not capture the full diversity of Indonesian dialects, and findings may not generalize to less-documented varieties. The dataset is relatively small and based on controlled dialogue writing, which may differ from natural language use, and annotator demographics are relatively homogeneous. Our analysis focuses on lexical-level variation and text-based modeling, without accounting for phonological or syntactic factors, and does not establish causal relationships. In addition, word-level similarity measures and surface-based metrics such as BLEU may obscure true similarity or underestimate semantically correct outputs under orthographic variation.

Ethical Consideration

IndoDialect supports research on dialect processing for Indonesian local languages in low-resource settings. However, dialect boundaries are fluid and the dataset does not capture the full diversity of Indonesian dialects. As the data are constructed through controlled annotation and partially derived from subtitle sources, they may not fully reflect natural language use. These limitations should be considered when interpreting results or deploying models, and we discourage applications that may lead to harmful, misleading, or culturally inappropriate outcomes.

All human-authored data were manually reviewed to avoid harmful or inappropriate content. Annotators were fairly compensated[4], provided informed consent, and no personally identifiable information is collected or disclosed. AI assistants were used only for grammar checking, while all experimental design and analysis were conducted by the authors.

References

Kaori Abe, Yuichiroh Matsubayashi, Naoaki Okazaki, and Kentaro Inui. 2018. Multi-dialect neural machine translation and dialectometry. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

K. Alexander Adelaar. 1994. The classification of tamanic languages. In Tom Dutton and Darrell T. Tryon, editors, *Language Contact and Change in the Austronesian World*, pages 1–42. Walter de Gruyter, Berlin.

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Karl Ronald Anderbeck. 2008. *Malay dialects of the Batanghari river basin (Jambi, Sumatra)*. SIL International.

Edmund A Anderson. 1997. *The use of speech levels in Sundanese*. Pacific Linguistics.

Rachel Bawden, Jonathan Poinhos, Eleni Kogkitsidou, Philippe Gambette, Benoît Sagot, and Simon Gabay. 2022. Automatic normalisation of early Modern French. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3354–3366, Marseille, France. European Language Resources Association.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Dea Adhista, Emmanuel Dave, Sarah Oktavianti, Salsabil Akbar, Jhonson Lee, Nuur Shadieq, Tjeng Wawan Cenggoro, Hanung Linuwih, Bryan Wilie, Galih Muridan, Genta Winata, David Moeljadi, Alham Fikri Aji, Ayu Purwarianti, and Pascale Fung. 2023. NusaWrites: Constructing high-quality corpora for underrepresented and extremely low-resource languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 921–945, Nusa Dua, Bali. Association for Computational Linguistics.

[4] Annotators were recruited via Upwork and compensated USD 76 (approximately 1,270,000 IDR) for completing 200 translation and dialogue-writing tasks. Compensation was determined proportionally with respect to Indonesia's average monthly minimum wage (approximately 3,300,000 IDR).

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, Ayu Purwarianti, and Pascale Fung. 2021. IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Team Cohere, :, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Milad Alizadeh, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, and 211 others. 2025. Command a: An enterprise-ready large language model. *Preprint*, arXiv:2504.00698.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Arawinda Dinakaramani, Fam Rashel, Andry Luthfi, and Ruli Manurung. 2014. Designing an indonesian part of speech tagset and manually tagged indonesian corpus. In *2014 International Conference on Asian Language Processing (IALP)*, pages 66–69. IEEE.

Vici Dwisetyo, Ikhlas Abdi Perkasa, Beny Endriyanto, Raviona Annida, and Risfa Nur Aisyah. 2023. Linguistic divergence of dayak ngaju and kadorih: A dialectometric and cultural perspective. *Journal Compound: Improving the Quality of English Education*, 11(2):62–66.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2021. *Ethnologue: Languages of the World*, 24 edition. SIL International, Dallas, Texas.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219, Toronto, Canada. Association for Computational Linguistics.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECTBENCH: An NLP benchmark for dialects, varieties, and closely-related languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14412–14454, Bangkok, Thailand. Association for Computational Linguistics.

Abdurrisyad Fikri and Ayu Purwarianti. 2012. Case based indonesian closed domain question answering system with real world questions. In *2012 7th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, pages 181–186. IEEE.

Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. A report on the VarDial evaluation campaign 2020. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

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.

William Gunawan, Derwin Suhartono, Fredy Purnomo, and Andrew Ongko. 2018. Named-entity recognition for indonesian language using bidirectional lstm-cnns. *Procedia Computer Science*, 135:425–432.

William Held, Caleb Ziems, and Diyi Yang. 2023. TADA : Task agnostic dialect adapters for English. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 813–824, Toronto, Canada. Association for Computational Linguistics.

Andi Hermanto, Teguh Bharata Adji, and Noor Akhmad Setiawan. 2015. Recurrent neural network language model for english-indonesian machine translation: Experimental study. In *2015 International conference on science in information technology (ICSITech)*, pages 132–136. IEEE.

Erlin Kartikasari and Kisyani Laksono. 2022. Lexical and phonological differences in javanese in east java, indonesia. *Dialectologia*, 28:143–155.

Erlin Kartikasari, Kisyani Laksono, Agusniar Dian Savitri, and Diah Yovita Suryarini. 2018. A study of dialectology on javanese" ngoko" in banyuwangi, surabaya, magetan, and solo. *Humaniora*, 30(2):128–139.

Fajri Koto, Timothy Baldwin, and Jey Han Lau. 2022. Cloze evaluation for deeper understanding of commonsense stories in Indonesian. In *Proceedings of the First Workshop on Commonsense Representation and Reasoning (CSRR 2022)*, pages 8–16, Dublin, Ireland. Association for Computational Linguistics.

Fajri Koto and Ikhwan Koto. 2020. Towards computational linguistics in Minangkabau language: Studies on sentiment analysis and machine translation. In *Proceedings of the 34th Pacific Asia Conference on*

Language, Information and Computation, pages 138–148, Hanoi, Vietnam. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2020a. Liputan6: A large-scale Indonesian dataset for text summarization. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 598–608, Suzhou, China. Association for Computational Linguistics.

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024. IndoCulture: Exploring geographically influenced cultural commonsense reasoning across eleven Indonesian provinces. *Transactions of the Association for Computational Linguistics*, 12:1703–1719.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020b. IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 757–770, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Olli Kuparinen, Aleksandra Miletić, and Yves Scherrer. 2023. Dialect-to-standard normalization: A large-scale multilingual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13814–13828, Singapore. Association for Computational Linguistics.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 5039–5049.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics.

Putu Devi Maharani and Komang Dian Puspita Candra. 2018. Variasi leksikal bahasa bali dialek kuta selatan. *Mudra Jurnal Seni Budaya*, 33(1):76–84.

Shervin Malmasi and Marcos Zampieri. 2017. German dialect identification in interview transcriptions. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 164–169, Valencia, Spain. Association for Computational Linguistics.

John Nerbonne. 2003. Linguistic variation and computation (invited talk). In *10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary. Association for Computational Linguistics.

Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. 2024. The zeno's paradox of 'low-resource' languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17753–17774, Miami, Florida, USA. Association for Computational Linguistics.

J Noorduyn. 1963. Categories of courtesy in sundanese. *The Bible Translator*, 14(4):186–191.

Bernd Nothofer. 1977. *Dialektgeographische Untersuchung des Sundanesischen und des entlang der sundanesischen Sprachgrenze gesprochenen Javanischen und Jakarta-Malaiischen. Erster Teil*. Ph.D. thesis, Philosophische Fakultät, Universität zu Köln, Köln.

Bernd Nothofer. 1982. *The Languages of the Javo-Sumatra Hesion*, volume 95 of *Verhandelingen van het Koninklijk Instituut voor Taal-, Land- en Volkenkunde*. Nijhoff Publishers, The Hague.

Niko Partanen, Mika Hämäläinen, and Khalid Alnajjar. 2019. Dialect text normalization to normative standard Finnish. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 141–146, Hong Kong, China. Association for Computational Linguistics.

Patriantoro Patriantoro. 2021. The geography of dayak dialect in landak regency, west kalimantan. *Indonesian Journal of EFL and Linguistics*, pages 1–16.

Soepomo Poedjosoedarmo. 1968. Javanese speech levels. *Indonesia*, 6:54–81.

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.

Marah Rusmali, Amir Hakim Usman, Syahwin Nikelas, Nurzuir Husin, Busri Busri, Agusli Lana, M. Yamin, Isna Sulastri, and Irfani Basri. 1985. *Kamus Minangkabau–Indonesia*. Pusat Pembinaan dan Pengembangan Bahasa, Departemen Pendidikan dan Kebudayaan, Jakarta.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained Arabic dialect identification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1332–1344, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Chong Shin. 2024. A discussion of two taman languages in west kalimantan, indonesia. *Dialectologia*, 32:193–208.

Antonia Soriente and Kazuya Inagaki. 2012. Kalimantan languages: An overview of current research and documentation. *Current Trends of Linguistic Research of Indigenous Languages in Indonesia*, pages 1–17.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Husni Thamrin and Retty Isnendes. 2020. Sundanese dialect in sinar resmi traditional village in cisolok district, sukabumi regency (phonological perspective). In *Twelfth Conference on Applied Linguistics (CONAPLIN 2019)*, pages 89–96. Atlantis Press.

Jozina Vander Klok. 2015. The dichotomy of auxiliaries in javanese: Evidence from two dialects. *Australian Journal of Linguistics*, 35(2):142–167.

Alfan Farizki Wicaksono, Clara Vania, Bayu Distiawan, and Mirna Adriani. 2014. Automatically building a corpus for sentiment analysis on indonesian tweets. In *Proceedings of the 28th Pacific Asia conference on language, information and computing*, pages 185–194.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China. Association for Computational Linguistics.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Zedian Xiao, William Held, Yanchen Liu, and Diyi Yang. 2023. Task-agnostic low-rank adapters for unseen English dialects. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7857–7870, Singapore. Association for Computational Linguistics.

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

Amalia Zahra, Sadar Baskoro, and Mirna Adriani. 2009. Building a pronunciation dictionary for indonesian speech recognition system. In *The TCAST workshop, Singapore*.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain. Association for Computational Linguistics.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. VALUE: Understanding dialect disparity in NLU. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3701–3720, Dublin, Ireland. Association for Computational Linguistics.

A Data Statement for IndoDialect

A.1 General Information

Dataset title: IndoDialect
Dataset curators: Not listed due to double-blind review.
Dataset version: 1.0 (September 2025)
Data statement authors: Not listed due to double-blind review.
Data statement version: 1.0 (December 2025)

A.2 Executive Summary

IndoDialect is a curated dataset of multi-turn dialogues covering four Indonesian local languages and eleven dialects across Javanese, Sundanese, Minangkabau, and Dayak. It consists of two components: a translation-parallel corpus derived from Indonesian source dialogues and a WhatsApp-style conversational corpus designed to capture natural and culturally grounded dialect usage. All data are produced by native speakers of the target dialects and validated through human quality checking. IndoDialect supports research on dialect-level language understanding and generation, focusing on dialect identification and machine translation, and enables systematic analysis of dialect variation under both controlled and informal settings.

A.3 Curation Rationale

IndoDialect is created to address the limited coverage of dialect-level variation in existing multilingual and language-level datasets. The translation-parallel component ensures consistent semantic content across dialects for controlled comparison, while the WhatsApp-style component captures informal language, cultural references, and everyday usage patterns. All translations and dialogues are produced by native speakers and undergo human validation to ensure linguistic correctness and natural usage, balancing data consistency with authentic dialect expression for reliable evaluation across dialect-focused tasks.

A.4 Documentation for Source Datasets

The translation-parallel component of IndoDialect is derived from Indonesian source dialogues collected from publicly available conversational resources, including subtitle data and online video transcripts such as YouTube and OpenSubtitles. The WhatsApp-style conversational data are created directly by native speakers based on predefined topics from IndoCloze (Koto et al., 2022), rather than sourced from existing datasets. This design ensures natural dialogue, cultural grounding, and informal usage, while all data are reviewed during curation to remove sensitive or harmful content and maintain consistency across dialects.

A.5 Language Variety

Dayak is a collective term for a group of Austronesian languages spoken across Central and West Kalimantan. Many Dayak varieties have relatively small speaker populations, often ranging from ten thousand to one hundred thousand (Eberhard et al., 2021). These languages show considerable lexical and morphological diversity due to geographic separation and long-standing local traditions (Soriente and Inagaki, 2012; Patriantoro, 2021). Among the many Dayak varieties, two that are particularly relevant for this study are Dayak Ngaju and Dayak Taman. Dayak Ngaju is a West Barito language spoken along major rivers in Central Kalimantan and has historically functioned as a regional lingua franca (Dwisetyo et al., 2023). Dayak Taman covers two related but distinct varieties: Tamanic Taman and Sekadau Taman (Adelaar, 1994; Shin, 2024).

Javanese is an Austronesian language spoken mainly on Java, with over 80 million speakers (Eberhard et al., 2021). It is characterized by a well-developed system of speech levels, such as *ngoko* (informal) and *krama* (polite), which mark social hierarchy and interactional norms (Poedjosoedarmo, 1968). The language shows substantial regional variation in phonology, vocabulary, and pragmatic expression (Kartikasari and Laksono, 2022). A widely used classification divides Javanese into three major dialect areas: *Western Javanese*, *Central Javanese*, and *Eastern Javanese*, each containing several subdialects with distinct linguistic traits (Nothofer, 1982).

Minangkabau is an Austronesian language spoken mainly in West Sumatra and parts of Riau and Negeri Sembilan in Malaysia, with an estimated 5–6 million speakers (Eberhard et al., 2021). Rusmali et al. (1985) identifies 6 major dialect areas based on regencies in West Sumatra, namely *Agam*, *Lima Puluh Kota*, *Pariaman*, *Tanah Datar*, *Pesisir Selatan*, and *Solok*. These dialects differ in phonology, morphology, and traditional registers, reflecting regional identity and historical migration patterns.

Sundanese is the second most widely spoken language in Java, with about 30–40 million speakers (Eberhard et al., 2021). Like Javanese, it is an Austronesian language and has a politeness system that influences lexical and grammatical choices (Noorduyn, 1963). Anderson (1997) describes three Sundanese speech levels, namely *lemes* (polite), *sedeng* (middle), and *kasar* (rough). Sundanese also shows considerable regional variation, with dialects differing in phonology, prosody, and core vocabulary (Thamrin and Isnendes, 2020). A major dialect study by Nothofer (1977) identifies 4 principal dialect areas: *Banten*, *Bogor*, *Priangan*, and *Cirebon*, which includes the western part of Brebes.

A.6 Speaker and Annotator Demographic

In IndoDialect, the speakers and annotators are the same individuals involved in data creation. All participants are native speakers of their respective target dialects, fluent in Indonesian, and recruited from regions where the dialects are commonly spoken based on self-reported regional background. These regions include Pati and Demak (Muria), Cirebon (Cirebon), Banjarnegara (Ngapak), Sidoarjo and Surabaya (Arekan), Yogyakarta (Mataraman), Agam (Agam), Lima Puluh Kota (Lima Puluh Kota), Bandung and Garut (Parahyangan), Pandeglang and Serang (Banten), Kapuas Hulu (Taman), and Kapuas and Palangkaraya (Ngaju). All participants are between 18 and 29 years old, and gender information is recorded in aggregate form, with a mix of male and female representation across dialect groups.

Annotators are selected based on self-reported native proficiency in the target dialect, long-term residence in the target region for more than ten years with at least one parent being a native of that region, and demonstrated ability to perform translation and dialectal writing tasks. All annotators have completed at least high school education (SMA or equivalent), and most have completed or are pursuing an undergraduate degree. Three annotators report prior experience in data annotation, while the remaining annotators do not. For the Daily Topic Conversation (DTC) component, dialogues are written collaboratively by two annotators to simulate natural two-party, multi-turn conversations using everyday informal language. For the Talkshow Data component, annotators follow the register and content of the Indonesian source dialogues to maintain semantic consistency across dialects.

A.7 Compensations

Annotators were recruited through Upwork and compensated USD 76 (approximately 1,270,000 IDR) for completing 1,000 translation and 200 dialogue-writing tasks. The total workload corresponds to part-time effort and compensation was determined proportionally based on Indonesia's average monthly minimum wage (approximately 3,300,000 IDR). All participants provided informed consent, were aware that their contributions would be used for research purposes, and that the resulting dataset would be publicly released. The dataset was manually reviewed to ensure that it does not contain harmful, offensive, or sensitive content, and no personal or identifying information about annotators is included.

A.8 Intended Use and Responsible Use

IndoDialect is intended for academic research on dialect processing, including tasks such as dialect identification and machine translation, with a focus on low-resource and closely related language varieties. The dataset is constructed through controlled annotation and is designed to support systematic analysis; however, it may not fully capture the diversity of real-world dialect usage. As such, it should not be used to make assumptions about linguistic correctness, social hierarchy, or broader cultural practices beyond the specific contexts represented in the data.

IndoDialect is released for research and educational purposes. Although the dialogues are newly created and translated by native speakers, part of the dataset is derived from publicly available subtitle data, and users are responsible for ensuring compliance with the licensing terms of the original sources. The dataset is not intended for commercial use without appropriate review of licensing conditions, and we discourage applications that may lead to harmful, misleading, or culturally inappropriate outcomes. Users are encouraged to cite this work when using IndoDialect in academic or applied research.

B Annotator Guidelines

Annotators are native speakers of the target dialects and were trained to produce natural, contextually appropriate conversational data. The annotation process consists of two main tasks: dialogue trans-

lation and dialogue creation, along with an additional quality checking step to ensure consistency and accuracy.

B.1 Task A: Dialogue Translation

Annotators translate Indonesian dialogues into a target dialect while preserving meaning, tone, and conversational structure, as showing in Table 7. Translations are required to be natural and fluent rather than literal, reflecting realistic usage of the dialect. Annotators are instructed to consistently use the assigned dialect, maintain dialect-specific grammatical and lexical patterns, and apply appropriate politeness levels based on speaker roles and social context. Expressions may be adapted to reflect relationships such as peer-to-peer or family interactions, while unnecessary mixing with standard Indonesian should be avoided unless contextually appropriate.

Indonesian (Source)	**Dialect (Example)**
Aku mau pergi ke pasar sekarang. *(I want to go to the market now.)*	Aku arep lunga menyang pasar saiki. *(I want to go to the market now.)*

Table 7: Example of dialogue translation from Indonesian to a Javanese dialect.

B.2 Task A.2: Translation Quality Checking

After translation (Table 8), annotators review a subset of dialogues produced by others within the same dialect speaker to ensure quality and consistency. This process involves correcting typographical and grammatical errors, verifying that translations preserve the original meaning, and ensuring appropriate use of politeness levels and dialect-specific forms. The goal is to maintain consistency across annotators while preserving natural variation in language use.

Source	**Before (Incorrect)**	**After (Corrected)**
Saya mau makan dulu. *(I want to eat first.)*	Aku mau makan dulu. *(I want to eat first.)*	Kulo badhe dhahar rumiyin. *(I would like to eat first.)*

Table 8: Example of translation quality checking. The corrected version reflects appropriate politeness and dialect usage.

B.3 Task B: Dialogue Creation

Annotators collaboratively create multi-turn dialogues based on predefined topics and contexts (Table 9). Each dialogue consists of five turns (A–B–A–B–A) and is expected to develop naturally without rigid scripting. Annotators use dialect-specific vocabulary, expressions, and discourse markers, while maintaining coherence and relevance to the assigned topic. Informal or colloquial forms are encouraged where appropriate, and switching to other dialects or standard Indonesian should be avoided unless required by context.

Dialogues must also reflect culturally appropriate interactions, including correct use of politeness and role-based language. Annotators are instructed to avoid offensive or inappropriate content and to ensure that conversations flow naturally, maintain consistent tone, and end with a clear conversational closure.

Example Scenario *Role A = Pak Surya (older), Role B = Mbak Rini (younger). Context: Pak Surya invites Mbak Rini to check farmland ready for harvest and discuss selling strategies for the upcoming market.*

Speaker	**Utterance (Javanese)**	**English Translation**
Pak Surya	Mbak Rini, monggo nderek kulo mriksani sawah ingkang sampun siyap dipun panen.	Mbak Rini, please come with me to check the field that is ready to be harvested.
Mbak Rini	Inggih Pak, kulo siap. Menawi sampun dipun panen, rencanane badhe dipun sade wonten pundi?	Yes, sir, I am ready. After harvesting, where do you plan to sell it?
Pak Surya	Kulo remen badhe dipun sade wonten pasar minggu ngajeng, amargi regine langkung sae.	I would prefer to sell it at next week's market because the price is better.
Mbak Rini	Inggih Pak, menawi mekaten kita saged nyiapaken strategi sade saking sakmeniko.	Yes, sir, in that case we can start preparing a selling strategy from now.
Pak Surya	Leres, kita kedah ngatur distribusi supados hasilipun saged maksimal.	That's right, we need to manage the distribution so that the results can be maximized.

Table 9: Example of dialogue creation reflecting role-based politeness and hierarchical interaction, with English translations for clarity.

All dialogues are expected to maintain natural conversational flow, consistent tone, and appropriate closure. Annotators are instructed to avoid offensive or inappropriate content and to ensure that language use reflects realistic communication in the target dialect.

C Prompt Template

Dialect identification. We use an in-context learning prompt that asks the model to return the *top-3* most likely dialect labels for a given conversation window in a strictly JSON-formatted output, illustrated in Figure 5.

Machine translation. We prompt models to translate a source sentence from a specified dialect

Dialect Identification Prompt Template

[Few-shot block]
Example 1:
Conversation: <conversation_text_1>
Answer: <label_1>, <label_2>, <label_3>
...
Example k:
Conversation: <conversation_text_k>
Answer: <label_1>, <label_2>, <label_3>

Human: Given the conversation: <conversation_text>
Here are the options of dialects:
id, jv_are, jv_mat, jv_mur, jv_nga, jv_cir, su_par, su_btn, day_nij, day_tmn, min_agm, min_lkt, others
Assistant: Return ONLY JSON in exactly this format:
{
"input": "<conversation_text>",
"lid": ["<label1>", "<label2>", "<label3>"]
}
Make sure to include exactly three distinct labels from the given list, in order of likelihood.

Figure 5: Prompt template used for dialect identification.

to a target dialect, optionally preceded by k in-domain examples to guide the mapping, illustrated in Figure 6.

Machine Translation Prompt Template

You are a translation system. Translate the following text from <SOURCE_DIALECT> to <TARGET_DIALECT>.

[Few-shot block]
Example 1:
<SOURCE_DIALECT>: <source_text_1>
<TARGET_DIALECT>: <target_text_1>
...
Example k:
<SOURCE_DIALECT>: <source_text_k>
<TARGET_DIALECT>: <target_text_k>

<SOURCE_DIALECT>: <source_text>
<TARGET_DIALECT>:

Figure 6: Prompt template used for machine translation.

Dialect-controlled generation. We prompt models to continue a dialogue in a specified target dialect, given a conversation context in a source dialect. The model is instructed to generate only the next turn while preserving the speaker label and turn index format, illustrated in Figure 7.

D Hyperparameters

D.1 Dialect Identification

For dialect identification, we evaluate both statistical and encoder based models using controlled hyperparameter search. Statistical baselines including NB, SVM, and LR are trained with bag of words and TF IDF features, with model specific hyperparameters selected from predefined ranges, as summarized in Table 10.

Dialect-Controlled Generation Prompt Template

You are a dialogue assistant.
The conversation so far is written in <SOURCE_DIALECT>.
Continue the dialogue by answering as the same speaker as the last turn, but write the response in <TARGET_DIALECT>.
Preserve the speaker label and dialogue format exactly as in the context. Generate only a single next turn.

[Few-shot block]
Example 1:
<SOURCE_DIALECT> (conversation so far):
<context_text_1>
<TARGET_DIALECT> (next turn):
<target_reply_1>
...
Example k:
<SOURCE_DIALECT> (conversation so far):
<context_text_k>
<TARGET_DIALECT> (next turn):
<target_reply_k>

<SOURCE_DIALECT> (conversation so far):
<conversation_text>
<TARGET_DIALECT> (next turn):

Figure 7: Prompt template used for dialect-controlled generation.

Hyperparams	NB	SVM	LR
Feature		{BoW, TF-IDF}	
α	0.001–1.0	–	–
C	–	0.01–100	0.001–100
Kernel	–	{linear, rbf}	–

Table 10: Hyperparameters and search ranges for statistical models used in dialect identification.

Encoder-based models are fine-tuned using IndoBERT and XLM R variants with a shared training setup, where learning rates and batch sizes are explored, and other settings are kept fixed for consistency, as shown in Table 11.

Setting	Value / Search Space
Models	IndoBERT-base, IndoBERT-large, XLM-R-base, XLM-R-large
Max length	256
Learning rate	{1e-4, 5e-5, 1e-5, 5e-6, 1e-6}
Batch size	{8, 16, 32}
Epochs	50

Table 11: Hyperparameters for fine-tuning encoder-based models on dialect identification.

D.2 Machine Translation

We fine-tune IndoGPT, IndoBARTv2, and $mT5_{BASE}$ using a shared set of hyperparameters to ensure a fair comparison across models, as summarized in Table 12. All models are trained with multiple random seeds, while learning rates follow common practice for each architecture.

Hyperparameters	IndoGPT	IndoBARTv2	$mT5_{BASE}$
Learning rate	1e-4	1e-4	5e-4
Batch size		16	
Gamma	0.98	0.98	0.95
Max epochs		20	
Early stop		10	
Seed		{1 . . . 3}	

Table 12: Hyperparameters used for fine-tuning pre-trained models on the machine translation task.

E Dialect Identification

E.1 Model Performance

Tables 13 and 14 summarize dialect identification performance across models on the Talkshow and DTC datasets. Across both settings, encoder-based and classical models consistently outperform large generative language models, with $IndoBERT_{LARGE}$ and Naive Bayes achieving the strongest and most stable performance across dialects. Errors are concentrated among closely related Javanese dialects, particularly `mat` and `mur`, while Minangkabau and Dayak dialects are identified with near-perfect accuracy. Performance generally decreases on the DTC dataset due to informal spelling and conversational variation, but the relative ranking of models remains consistent, indicating that dialect similarity—rather than dataset style—is the primary source of confusion.

E.2 Effect on Input Length

We examine how $IndoBERT_{LARGE}$ performance varies with input length across Talkshow and DTC datasets. As shown in Figure 8, F1 scores generally improve with longer inputs, with the largest gains from short to medium-length texts and smaller improvements beyond that. This trend is stronger on DTC, where informal and conversational writing benefits more from added context, while Talkshow remains relatively stable due to its controlled style. Gains are not uniform across dialects: varieties like `su_par` and `day_tmn` improve noticeably with longer inputs, whereas closely related Javanese dialects remain difficult to separate even with more context.

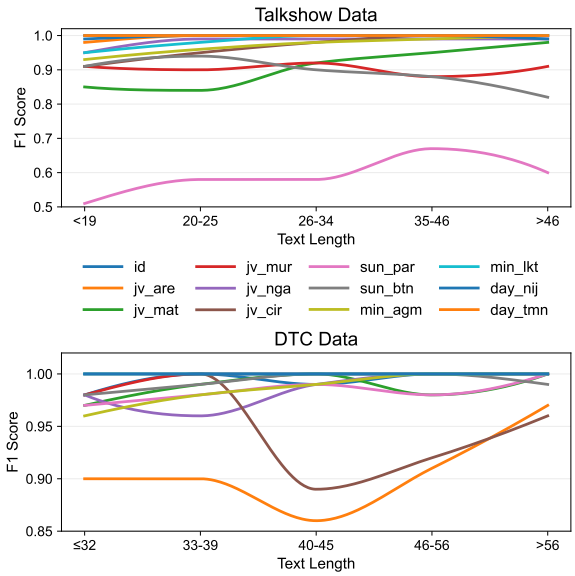

Figure 8: F1 score trends of $IndoBERT_{LARGE}$ across text length bins on the Talkshow Data and DTC Datasets.

E.3 Dialect-Level Confusion Analysis

Figure 9 shows that $IndoBERT_{LARGE}$ achieves near-perfect accuracy for most dialects, but confusion persists among closely related varieties. Javanese dialects (`are`, `mat`, `mur`, `nga`, `cir`) exhibit the highest overlap, with misclassifications concentrated between `mat` and `mur`. Sundanese dialects also show moderate confusion, particularly between `par` and `btn`, while Minangkabau and Dayak varieties are almost perfectly distinguished, reflecting stronger lexical separation.

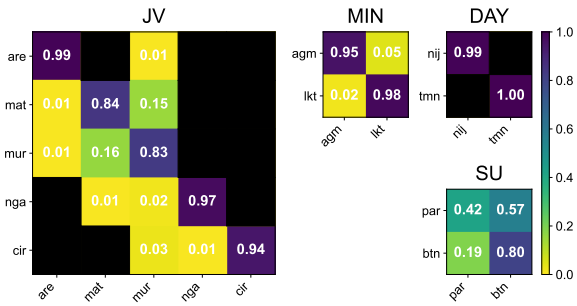

Figure 9: Confusion matrix of $IndoBERT_{LARGE}$ on the Talkshow Data dataset across dialect groups.

Figure 10 shows that $IndoBERT_{LARGE}$ performs well on the DTC dataset, but confusion remains among closely related dialects. This is most evident within the Javanese group, with frequent misclassifications between `mat` and `mur`, and to a lesser

extent between Sundanese `par` and `btn`. In contrast, Minangkabau and Dayak dialects are largely well separated despite conversational noise.

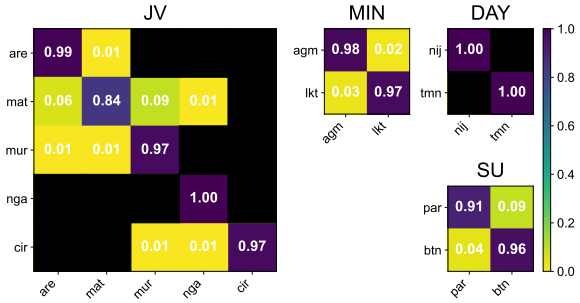

Figure 10: Confusion matrix of IndoBERT$_{\text{LARGE}}$ on the DTC Dataset across dialect groups.

F Machine Translation

Indonesian-to-dialect Table 15 presents detailed SacreBLEU scores for Indonesian-to-dialect translation across all dialects.

Dialect-to-Indonesian Table 16 reports detailed SacreBLEU scores for dialect-to-Indonesian translation across all dialects.

Human Evaluation While automatic metrics capture overall trends, they do not reveal specific error types. To better understand model behavior for Indonesian-to-dialect translation, we conduct a human evaluation of PBSMT outputs. Annotators classify errors into five categories: **Addition** (extra information not in the reference), **Omission** (missing words or phrases), **Word Order Error** (unnatural or confusing sequence), **Dialect Error** (use of incorrect dialect), and **Morphology Error** (incorrect or unnatural word forms).

Table 17 presents representative examples from the human evaluation of PBSMT outputs on Section 5.2, illustrating common error types observed across dialects. Annotators highlight additions, omissions, word order issues, dialect mismatches, and morphological errors to provide qualitative insight into the strengths and limitations of phrase-based translation in dialectal settings.

G Dialect-Controlled Generation

Table 18 reports dialect-level accuracy for dialect controlled generation across all dialects, providing a fine-grained view of model performance beyond the aggregated results in the main paper.

Model	ID	Javanese					Sundanese		Minangkabau		Dayak		AVG
	id	are	mat	mur	nga	cir	par	btn	agm	lkt	nij	tmn	
Multilingual LMs													
GPT-5	0.998	0.192	0.250	0.451	0.530	0.696	0.569	0.624	0.568	0.525	0.960	0.844	0.601
GPT-mini	0.979	0.006	0.392	0.151	0.041	0.113	0.601	0.038	0.377	0.041	0.339	0.147	0.269
Command A	0.985	0.447	0.370	0.024	0.372	0.411	0.051	0.663	0.518	0.240	0.684	0.333	0.425
Qwen3 8B	0.268	0.196	0.076	0.000	0.177	0.000	0.039	0.000	0.000	0.000	0.000	0.127	0.074
Llama 3.1 8B	0.622	0.230	0.000	0.026	0.007	0.000	0.332	0.078	0.353	0.173	0.396	0.000	0.185
Classical Machine Learning													
NB	1.000	0.997	0.818	0.792	0.990	0.988	0.500	0.714	0.978	0.978	1.000	1.000	0.896
SVM	0.995	0.977	0.832	0.775	0.988	0.981	0.514	0.664	0.947	0.950	1.000	1.000	0.885
LR	0.995	0.979	0.827	0.771	0.988	0.979	0.515	0.661	0.955	0.958	1.000	1.000	0.886
Pre-trained Models													
IndoBERT$_{\text{BASE}}$	0.995	0.975	0.815	0.787	0.970	0.975	0.528	0.671	0.961	0.962	0.995	0.998	0.886
IndoBERT$_{\text{LARGE}}$	0.993	0.982	0.830	0.809	0.978	0.966	0.520	0.675	0.965	0.966	0.995	1.000	0.890
XLM-R$_{\text{BASE}}$	0.995	0.987	0.818	0.795	0.975	0.978	0.540	0.640	0.964	0.966	0.997	0.995	0.887
XLM-R$_{\text{LARGE}}$	0.995	0.982	0.845	0.826	0.973	0.975	0.491	0.630	0.971	0.972	0.995	0.998	0.888

Table 13: Dialect identification performance on F1 of the Talkshow dataset. Results are grouped by model family and language–dialect columns.

Model	Javanese					Sundanese		Minangkabau		Dayak		AVG
	are	mat	mur	nga	cir	par	btn	agm	lkt	nij	tmn	
Multilingual LMs												
GPT-5	0.288	0.249	0.502	0.345	0.583	0.584	0.755	0.529	0.408	0.981	0.795	0.547
GPT-mini	0.036	0.418	0.208	0.023	0.318	0.633	0.056	0.386	0.000	0.274	0.053	0.219
Command A	0.309	0.296	0.005	0.154	0.061	0.029	0.669	0.478	0.170	0.669	0.230	0.279
Qwen3 8B	0.207	0.071	0.049	0.117	0.000	0.015	0.000	0.005	0.014	0.000	0.194	0.061
Llama 3.1 8B	0.249	0.000	0.000	0.010	0.005	0.342	0.022	0.427	0.204	0.604	0.005	0.170
Classical Machine Learning												
NB	0.998	0.966	0.968	0.998	0.996	0.923	0.931	0.976	0.976	1.000	1.000	0.976
SVM	0.968	0.926	0.959	0.983	0.992	0.911	0.921	0.964	0.963	1.000	1.000	0.963
LR	0.947	0.883	0.947	0.983	0.996	0.927	0.935	0.959	0.958	1.000	1.000	0.958
Pre-trained Models												
IndoBERT$_{\text{BASE}}$	0.951	0.914	0.934	0.986	0.992	0.901	0.911	0.959	0.959	1.000	0.999	0.955
IndoBERT$_{\text{LARGE}}$	0.964	0.903	0.930	0.989	0.986	0.933	0.939	0.974	0.974	1.000	1.000	0.963
XLM-R$_{\text{BASE}}$	0.950	0.854	0.911	0.973	0.987	0.928	0.932	0.974	0.975	1.000	0.999	0.953
XLM-R$_{\text{LARGE}}$	0.980	0.826	0.878	0.989	0.986	0.952	0.953	0.970	0.971	1.000	0.999	0.955

Table 14: Dialect identification performance on F1 of the DTC dataset. Results are grouped by model family and language–dialect columns.

ID → Dialect$_i$												
Model	Javanese					Sundanese		Minangkabau		Dayak		avg
	are	mat	mur	nga	cir	par	btn	agm	lkt	nij	tmn	
Copy	4.88	3.85	5.24	3.02	6.77	2.79	2.50	5.14	5.39	2.70	3.42	4.15
PBSMT	**40.73**	25.20	**35.16**	28.88	29.89	18.54	25.09	**40.22**	**25.64**	**24.71**	**12.85**	**27.90**
GPT-5	14.31	23.34	29.17	11.89	14.10	**26.37**	23.19	16.62	11.26	2.53	2.47	15.93
GPT-5-mini	10.07	22.51	23.52	6.86	7.32	21.45	19.55	9.08	6.43	1.68	2.34	11.89
Command A	10.48	16.61	21.46	9.83	11.15	12.30	11.07	7.61	7.99	2.20	2.35	10.28
Qwen3 8B	29.84	19.76	25.59	22.13	21.89	15.14	16.30	30.50	20.34	12.14	5.78	19.95
Llama-3.1 8B	30.18	19.04	25.53	22.48	23.84	16.73	18.32	31.48	20.14	13.99	7.02	20.80
IndoGPT	16.85	6.72	12.06	13.39	11.52	6.57	14.59	23.77	5.56	8.24	2.67	11.09
IndoBARTv2	39.27	**25.40**	32.67	**30.18**	**30.49**	17.45	24.64	27.42	22.22	22.47	12.01	25.84
mT5$_{\text{BASE}}$	37.68	23.44	33.07	29.10	29.57	17.48	**25.45**	16.89	21.12	16.60	9.49	23.63

Table 15: SacreBLEU scores for Indonesian-to-dialect translation across models (Indonesian → Dialect$_i$)

Dialect$_i$ → ID												
Model	**Javanese**					**Sundanese**		**Minangkabau**		**Dayak**		**avg**
	are	mat	mur	nga	cir	par	btn	agm	lkt	nij	tmn	
Copy	4.88	3.85	5.22	3.01	6.80	2.76	2.49	5.10	5.35	2.67	3.46	4.14
PBSMT	**43.11**	38.53	34.11	25.59	41.95	24.49	22.24	38.12	**35.90**	**23.66**	19.54	**31.57**
GPT-5	37.87	**39.79**	**35.40**	23.47	33.92	25.80	**25.65**	26.11	25.42	9.14	8.90	26.50
GPT-5-mini	26.30	33.00	28.91	15.13	24.77	21.12	19.17	18.89	17.84	4.09	7.14	19.67
CommandA	32.08	32.06	33.70	21.98	32.24	20.46	19.19	22.62	22.03	7.95	9.03	23.03
Qwen3 8B	29.10	26.48	24.62	18.06	31.77	18.87	15.23	26.82	25.58	10.51	11.79	21.71
Llama-3.1 8B	24.49	21.49	21.06	17.10	26.84	16.84	13.00	25.77	21.77	10.74	10.60	19.06
IndoGPT	11.88	10.87	8.07	13.24	21.99	4.33	13.20	12.08	10.71	8.07	3.45	10.72
IndoBARTv2	41.32	39.28	33.92	25.50	**42.70**	**25.81**	21.78	37.58	33.46	21.43	**21.49**	31.30
mT5$_{BASE}$	32.65	37.09	31.78	**26.27**	40.69	25.22	21.15	**39.87**	35.47	22.54	15.76	29.86

Table 16: SacreBLEU scores for dialect-to-Indonesian translation across models (Dialect$_i$ → Indonesia).

Dialect	Reference	Hypothesis
day_nij	tuh uluh uluh kantor regional tuh jadi pasti angat are waktu akan tes wawancara malah ikau ah je handak mundur *(Those people at the regional office are surely angry; during the interview you suddenly wanted to withdraw.)*	toh oloh kantor regional toh jadi pasti angat dipermainkan ikau ie jadi kalute nah ikau mampalepah katika akan assesment wawancara malah bara kamunya je mundur *(Those people at the regional office are surely angry; you were being toyed with, you became confused, you ended up letting go during the assessment interview, and in the end you yourself withdrew.)*
su_par	ti leutik nini galak ka bayu unggal bayu nyokot dahareun sok dicarekan kue lebaran disumputkeun padahal keur lebaran pa *(Since childhood, grandma often scolded Bayu whenever he took food and hid the Eid cookies.)*	ti leutik nini ni galak jeung bayu unggal bayu mawa kadaharan ambek kueh oge konsentrasi kana lebaran disumputkeun padahal deui oge konsentrasi kana lebaran pa *(Since childhood, grandma was often angry with Bayu; whenever Bayu brought food he got upset about the cakes, still focused on Eid, hiding them even though he was also focused on Eid.)*

Table 17: Example of human evaluation analysis for PBSMT outputs. Highlighted spans indicate different error types: Addition, Omission, Word Order Error, Dialect Error, and Morphology Error. English translations are provided only to help readers understand the context and are not used for evaluation.

Model	**Javanese**					**Sundanese**		**Minangkabau**		**Dayak**	
	are	**mat**	**mur**	**nga**	**cir**	**par**	**btn**	**agm**	**lkt**	**nij**	**tmn**
GPT-5	13.33	48.67	54.00	1.33	1.33	33.33	76.51	50.00	44.30	20.67	13.33
GPT-5-mini	4.00	43.62	48.00	2.00	0.00	27.70	71.14	58.39	33.11	11.33	26.17
CommandA	12.08	28.67	40.54	14.29	2.03	27.70	61.07	34.00	31.54	4.08	31.29
Llama-3.1-8B	8.00	40.00	23.33	10.67	2.67	15.33	82.55	22.67	63.76	23.33	40.67
Qwen3-8B	28.00	28.67	44.67	25.33	54.00	3.33	89.86	48.00	54.36	94.63	100.00
IndoGPT	**88.22**	**91.73**	**70.67**	**83.33**	**89.49**	**57.33**	**93.69**	**76.45**	**84.34**	89.11	95.11
IndoBARTv2	34.89	73.15	61.56	75.78	63.98	42.67	86.49	66.89	53.02	**97.56**	83.33
mT5$_{base}$	53.78	88.59	37.33	75.33	49.44	40.89	36.49	67.11	76.96	95.78	**93.33**

Table 18: Dialect-level accuracy (%) for dialect steering across all 11 dialects.