Sundanese Discourse Particles in Dialogue: A Human-Annotated Corpus with Computational Baselines

Irwan Setiawan[1,2] **Kun-Ta Chuang**[1] **Hung-Yu Kao**[3]
[1] Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan;
[2] Department of Computer and Informatics Engineering, Politeknik Negeri Bandung, Indonesia
[3] Department of Computer Science, National Tsing Hua University, Taiwan;

irwan@jtk.polban.ac.id,ktchuang@mail.ncku.edu.tw, hykao@cs.nthu.edu.tw

Abstract

Sundanese discourse particles are frequent in dialogue but difficult to model because their interpretation depends on information structure, stance, sequential organization, and lexical identity. We present a human-annotated corpus of 1,243 occurrences of ten particles from three scripted and literary dialogue sources. Three native-speaker experts independently labeled every occurrence using a particle-neutral twelve-label codebook, followed by consensus adjudication of all non-unanimous cases. Primary-function agreement was strong (Fleiss' $\kappa = 0.828$) and increased to 0.938 after TOPIC_MANAGEMENT and FOCUS_EMPHASIS were merged. For téh, fine-grained κ was 0.251, but agreement was complete at the merged information-structure level. We provide occurrence-aware masking, dialogue-unit-grouped splits, and lexical, TF-IDF, and XLM-R baselines. On the fixed test set, particle-majority and particle-plus-context XLM-R obtained 0.670 and 0.669 macro-F1. Across five dialogue-unit-grouped folds, word TF-IDF with particle and context reached 0.722 ± 0.029, compared with 0.670 ± 0.001 for particle-majority; leave-one-source-out means were 0.749 ± 0.067 and 0.662 ± 0.014. Gains were smaller on multifunctional particles, and context-only models remained weak. The results establish a reproducible resource while showing both lexical shortcuts and limited, variable contextual gains in low-resource pragmatic classification.

Keywords: Sundanese; discourse particles; pragmatic annotation; corpus resource; lexical shortcuts; multilingual NLP

1 Introduction

Discourse particles are short, often grammatically optional forms that connect an utterance to its local discourse and interactional context. They can organize topic and focus, evoke contrast, express stance, shape directives, manage turns, seek confirmation, restrict alternatives, or direct attention. Their contribution is rarely captured by a fixed dictionary equivalent because interpretation depends on position, neighboring turns, speech-act structure, prosody, and shared assumptions (Schiffrin, 1987; Fraser, 1999; Aijmer, 2002; Waltereit, 2001; Fischer, 2006; Degand et al., 2013).

This problem is especially visible in Sundanese, an under-resourced Austronesian language spoken primarily in West Java. Forms such as *mah, téh, da, atuh, waé, gé, nya,* and *tah* are frequent in dialogue but do not map consistently onto Indonesian or English words. Earlier studies document the broader function-word system and analyze contrasts among *téh, mah,* and *téa* (Mulyono and Sutari, 1981; Müller-Gotama, 1994), while recent corpus research confirms differences in frequency and combinatorial behavior (Rahmawati and Roselani, 2024). Comparative annotation of several frequent particles under one context-sensitive framework remains limited.

The resource gap matters computationally. Existing Sundanese NLP has emphasized morphology, generation, and broader resource aggregation (Cahyawijaya et al., 2021; Cahyawijaya et al., 2023; Setiawan and Kao, 2024), whereas local pragmatic functions remain largely unencoded. A manually verified pragmatic corpus can support linguistic description, translation, and context-aware classification. It can also expose a recurrent evaluation problem: particle identity may predict the dominant label so

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 312–321
November 6-7, 2026 ©2026 Association for Computational Linguistics

well that high aggregate accuracy does not demonstrate contextual understanding.

This study analyzes all 1,243 validated occurrences identified in the three selected dialogue sources. Every case was independently labeled by three native speakers with relevant expertise, followed by consensus adjudication of all non-unanimous decisions. We address three research questions: (RQ1) How are the ten particles distributed across pragmatic functions and sources? (RQ2) How reliably can these functions be annotated, and where is disagreement concentrated? (RQ3) To what extent can particle-only, context-only, and particle-plus-context systems distinguish these functions under grouped and source-transfer evaluation?

The main contributions are as follows: (1) a complete three-source corpus of ten Sundanese discourse particles; (2) a particle-neutral twelve-label codebook with pre-adjudication reliability and consensus gold labels; (3) sensitivity analyses that localize the topic-focus boundary; (4) a human masked diagnostic; and (5) occurrence-aware fixed, grouped five-fold, and leave-one-source-out evaluations with lexical, TF-IDF, and multilingual-transformer baselines.

2 Related Work

2.1 Discourse particles and annotation

Discourse and pragmatic markers have been analyzed as sequentially dependent elements, as signals of relations between discourse segments, and as devices for stance, interaction, and speech-act modification (Schiffrin, 1987; Fraser, 1999; Waltereit, 2001; Fischer, 2006; Yen and Chung, 2025). Their multifunctionality makes annotation more interpretive than surface-form labeling. Reliable schemes therefore require operational definitions, contextual windows, independent coding, uncertainty handling, and transparent agreement statistics (Cohen, 1960; Lombard et al., 2002; Artstein and Poesio, 2008; Krippendorff, 2019). Multilingual discourse resources have likewise adopted shared communicative-function inventories and explicit mapping procedures (Silvano et al., 2022).

2.2 Sundanese and information structure

For Sundanese, Mulyono and Sutari (1981) provide a broad account of function words, Müller-Gotama (1994) analyzes *téh, mah,* and *téa*, and Gunawan and Fauziyah (2021) trace the historical particle ma. Information-structure research distinguishes discourse topic, focus, contrast, and questions under discussion, but also recognizes their overlap in actual discourse (Lambrecht, 1994; Krifka, 2008; Roberts, 2012). These distinctions are central to *mah* and *téh*. Other forms are more interactional: *atuh* can affect stance or directive force, *nya* can manage alignment or sequence, and *tah* can direct attention.

2.3 Low-resource pragmatic modeling

Regional-language NLP resources are expanding, but discourse-pragmatic annotation remains sparse. Recent ROCLING work on Taiwanese Hakka likewise shows how carefully constructed corpora and explicit evaluation tasks can support technology development for an under-resourced regional language (Liao et al., 2025). Recent studies suggest that multilingual models may struggle with pragmatic distinctions even when they perform well on lexical or syntactic tasks (Ma et al., 2025; Jung, 2025). For discourse particles, this creates a specific risk: a model may memorize a dominant form-function mapping rather than use surrounding discourse. Our evaluation therefore separates lexical, masked-context, and combined conditions and reports macro-F1 and performance on multifunctional particles.

3 Data and Annotation

3.1 Sources and corpus preparation

The corpus combines *Juragan Hajat* (JH), a scripted drama with explicit speakers; *Si Kabayan Ngingu Munding* (DKI2), a six-page OCR-derived drama; and manually processed dialogue from pages 21–46 of *Si Kabayan Jadi Dukun* (SKJD). The analysis is exhaustive for this selected material, not for the complete works or Sundanese dialogue in general. OCR candidates were checked against available page images, and original spellings were retained alongside normalized forms. The resulting corpus composition and the number of validated targets in each source are summarized in Table 1.

Source	Type	Units	Tokens	Targets
JH	Scripted drama	710	8,170	898
DKI2	OCR drama	89	1,005	108
SKJD	Literary dialogue	355	3,533	237
Total	—	1,154	12,708	1,243

Table 1. Corpus composition and validated targets

A dialogue unit was defined as one speaker turn in the scripts and one manually segmented unit of direct speech or narration in the literary text. Tokenization separated punctuation while preserving reduplication, clitic-like sequences, and surface orthography. A unit could contain multiple target particles; each occurrence received a unique identifier and was annotated separately while the full turn and immediately neighboring turns remained visible.

3.2 Particle-neutral function inventory

Every valid occurrence received exactly one primary function and, when supported, one distinct concurrent secondary function. A competing interpretation was recorded separately from a simultaneous secondary contribution. All twelve labels were formally available for every particle; the protocol imposed no particle-specific restriction or default mapping. Annotators had to cite contextual evidence such as a recoverable alternative, discourse-topic relation, reason relation, requested action, restricted set, alignment target, sequential operation, or attention target. The complete label inventory and its operational definitions are presented in Table 2.

The twelve labels formed six macro families: information structure; account and stance; illocution and interpersonal meaning; scope; alignment and sequence; and attention. MITIGATION remained in the inventory because it occurred as a secondary contribution, typically with restriction or directive force, although it was never selected as the final primary label in this corpus.

3.3 Independent annotation and adjudication

Three native speakers of Sundanese with relevant expertise independently labeled all 1,243 occurrences under a locked protocol. They saw the target turn, immediately neighboring turns, speaker and source metadata, particle position, and utterance type, but not one another’s labels. Reliability was calculated on these independent labels. The annotators then jointly adjudicated every non-unanimous primary or secondary decision by reviewing context and evidence notes. Consensus, rather than mechanical majority voting, produced the final gold label.

Fleiss' kappa was the main three-rater statistic, supplemented by unanimous and majority agreement and pairwise Cohen's kappa. Sensitivity analyses merged TOPIC_MANAGEMENT and FOCUS_EMPHASIS into INFORMATION_STRUCTURE and excluded universal NONE cases from secondary-function agreement. Confidence ratings were retained as adjudication metadata rather than used as prediction targets. The complete annotation, adjudication, and masked-diagnostic workflow is shown in Figure 1.

3.4 Masked diagnostic

A purposive stratified sample of 180 occurrences covered all ten particles, with additional representation of rare and multifunctional forms. The target occurrence was replaced by [PART], while speaker identity and the previous, target, and following turns remained visible. Two human annotators independently applied the same codebook without the original particle or gold label. Since masking removes part of the linguistic signal and can alter naturalness, this was treated as a perturbation diagnostic, not as an independent validation of the full-token gold labels.

3.5 Computational evaluation

The resource supports three settings: particle-only prediction from normalized identity; context-only prediction from a three-turn window in which only the annotated occurrence is replaced by [PART]; and particle-plus-context prediction from the overt particle and the same window. Exact character offsets identify the selected occurrence, including turns containing repeated forms. The target is the adjudicated primary function. Fine-grained labels are primary; an additional sensitivity analysis merges TOPIC_MANAGEMENT and FOCUS_EMPHASIS into INFORMATION_STRUCTURE. The fixed 869/187/187 train/development/test split keeps all target occurrences from each source-specific dialogue unit in one partition. Because inputs include adjacent turns, some context text is shared across partitions. We additionally use five outer

group-stratified folds, with grouped inner development folds, and leave-one-source-out evaluation over JH, DKI2, and SKJD.

Label	Operational definition
TOPIC_MANAGEMENT	Establishes, maintains, resumes, delimits, or shifts what the discourse is about.
FOCUS_EMPHASIS	Makes a locally salient constituent or proposition especially prominent.
CONTRAST	Evokes opposition among explicit or recoverable alternatives or expectations.
EXPLANATION_JUSTIFICATION	Presents a reason, account, evidence, or justification.
STANCE_AFFECT	Conveys evaluation, surprise, irritation, resignation, or speaker attitude.
DIRECTIVE_FORCE	Shapes the force or organization of a request, command, warning, or urging.
MITIGATION	Reduces interpersonal pressure or softens commitment.
RESTRICTION_MINIMIZATION	Narrows alternatives or presents one option as sufficient or minimal.
ADDITIVE_SCALAR	Adds an item to a set or invokes scalar inclusion.
ALIGNMENT_CONFIRMATION	Seeks or displays confirmation, agreement, or shared orientation.
SEQUENCE_TURN_MANAGEMENT	Organizes turn entry, continuation, transition, response progression, or closure.
ATTENTION_DEIXIS	Directs attention to a referent, event, evidence, warning, or upcoming segment.

Table 2. Fine-grained primary-function inventory.

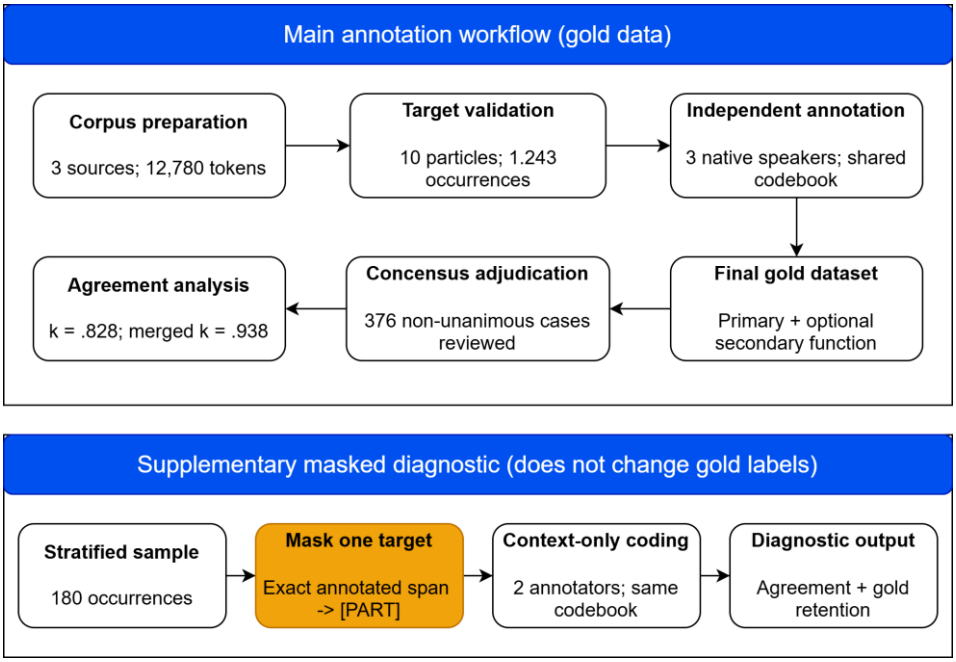

Figure 1. Main gold-data workflow and supplementary occurrence-masked diagnostic.

Each baseline answers a distinct question: global majority measures label imbalance; particle-majority measures lexical shortcuts; word and character TF-IDF test sparse contextual and orthographic cues; and xlm-roberta-base tests whether multilingual representations improve over lexical identity. TF-IDF uses class-balanced SGD logistic regression with L2 regularization (5,000 iterations; tolerance 1e-4; seed 2026). Development macro-F1 selects α from {1e-5, 3e-5, 1e-4, 3e-4, 1e-3, 3e-3}, word n-grams from {1, 1-2}, character n-grams from {3-5, 3-6, 4-6}, minimum document frequency from {1, 2}, and optional sublinear term frequency. XLM-R uses maximum length 256, eight epochs, batches of 8/16, gradient accumulation 2, learning rate 2e-5, weight decay 0.01, warmup 0.1, and seeds 13/29/47. Checkpoints are selected by development macro-F1 with early-stopping patience 2. Training uses unweighted cross-entropy. Macro-F1 is primary, supplemented by accuracy, weighted-F1, and performance on mah, téh, atuh, and nya.

4 Results

4.1 Validation and reliability

The final dataset contained 1,243 validated particle occurrences with no missing primary labels. Joint review covered 376 unique cases: 278

involved primary disagreement, 141 involved secondary disagreement, and 43 involved both. Primary agreement was strong (κ = 0.828; 77.6% unanimous), while the complete secondary field reached κ = 0.808. After excluding the 902 cases where all annotators selected NONE, secondary κ decreased to 0.675, providing a more informative estimate of agreement on cases in which a secondary function was potentially present. The overall, merged-label, secondary-function, and masked agreement results are summarized in Table 3.

The most frequent primary confusion was FOCUS_EMPHASIS versus TOPIC_MANAGEMENT (184 cases), followed by CONTRAST versus TOPIC_MANAGEMENT (51) and ALIGNMENT_CONFIRMATION versus STANCE_AFFECT (43). Particle-level reliability was lowest for *téh* (κ = 0.251). All 990 independent annotations for the 330 *téh* occurrences fell into either TOPIC_MANAGEMENT or FOCUS_EMPHASIS. After merging the two labels, unanimous and pairwise raw agreement reached 100%. Because all merged annotations belonged to one category, particle-level κ was undefined for this sensitivity condition. Thus the low fine-grained value localizes a systematic analytical boundary rather than disagreement across unrelated pragmatic families.

Per-source reliability remained strong: fine-grained κ was 0.817 for JH, 0.823 for DKI2, and 0.863 for SKJD. After merging topic and focus, κ exceeded 0.924 in every source. JH did not exhibit markedly higher agreement than the two smaller sources despite contributing most instances. In all 184 non-unanimous *téh* cases, the consensus label coincided with the label selected by two annotators. This pattern indicates that adjudication clarified a stable binary boundary rather than reversing majority judgments.

4.2 Frequency and functional profiles

The most frequent particles were *mah* (342) and *téh* (330), followed by *atuh* (163), *nya* (95), *da* (85), *waé* (72), *gé* (67), *tah* (57), *baé* (20), and *wéh* (12). *Mah* and *téh* together accounted for 54.1% of all targets. Secondary functions were retained in 271 cases (21.8%), with recurrent pairings such as CONTRAST plus TOPIC_MANAGEMENT and STANCE_AFFECT plus ALIGNMENT_CONFIRMATION. The particle-specific distributions are summarized in Table 4.

These concentrated profiles are corpus-attested tendencies rather than exhaustive lexical definitions. *Mah* is predominantly contrastive, whereas *téh* alternates between local focus and topic management. *Atuh* divides between affective stance and directive organization, while *nya* supports alignment or turn progression. These four particles require contextual disambiguation and are therefore reported separately in model evaluation.

4.3 Source-stratified variation

JH contributes 898 of 1,243 occurrences, so aggregate distributions can reflect one script's style. Source-stratified profiles nevertheless reveal notable descriptive variation. Mah is more strongly contrastive in JH (94.4%) than in SKJD (64.6%). Atuh is more directive in DKI2 (66.7%) than in JH (40.3%) or SKJD (14.3%). Téh remains divided between focus and topic across all sources, while nya is predominantly alignment-oriented in each source. The complete source-stratified profiles of these multifunctional particles are reported in Table 5.

These differences cannot be attributed uniquely to genre, author, period, or register because each factor is confounded with source. They motivate the group-aware and source-transfer evaluations reported below, while the small number of sources requires cautious interpretation.

4.4 Masked diagnostic

The two masked annotators agreed at the fine-function level in 76.7% of cases (κ = 0.717) and at the macro-family level in 82.2% (κ = 0.770). Fine-grained agreement with the gold labels was only 15.6% for Annotator A and 17.8% for Annotator B, and masked distributions shifted toward STANCE_AFFECT and DIRECTIVE_FORCE. Among 113 cases in which both masked annotators agreed but both differed from the gold labels, the dominant patterns were loss of information-structure cues, dominance of stance or directive interpretations, loss of lexical-scope information, and underdetermined fallback readings. We additionally examined a 25-item clean subset consisting of cases in which masking preserved grammaticality and did not remove an obligatory constructional cue. The masked inter-annotator

agreement and gold-retention results are summarized in Table 6.

The low gold-retention scores do not imply that discourse context is uninformative. Rather, the overt particle is itself part of the pragmatic evidence. After masking, annotators can often recover the broad action of the turn, such as affect, correction, or directive force, but lose lexical cues that distinguish contrast, information structure, restriction, and addition. The three-turn window may also omit earlier discourse topics and shared assumptions, while the absence of prosody and gesture further weakens topic-focus interpretation. The masked condition is therefore systematically underdetermined rather than a clean replacement for full-token annotation or an alternative procedure for validating the gold labels.

4.5 Computational baselines

On the fixed test set, particle-majority achieved 85.6% accuracy and 0.670 macro-F1. Its high accuracy reflects lexical concentration: all occurrences of six particles share one adjudicated primary label. On the four multifunctional particles, macro-F1 decreased to 0.481. On the fixed split, word and character TF-IDF with particle plus context reached 0.627 and 0.619 macro-F1; their context-only counterparts reached 0.104 and 0.167. Occurrence-aware context-only XLM-R obtained 0.203 ± 0.009 accuracy, 0.037 ± 0.001 macro-F1, and 0.077 ± 0.002 macro-F1 on multifunctional particles. It predicted only CONTRAST and FOCUS_EMPHASIS across the three runs. Particle-plus-context XLM-R obtained 0.854 ± 0.003 accuracy and 0.669 ± 0.001 macro-F1, nearly reproducing particle-majority.

Results were less dependent on one partition under five dialogue-unit-grouped folds. Particle-majority averaged 0.670 ± 0.001 macro-F1, whereas word and character TF-IDF with particle plus context reached 0.722 ± 0.029 and 0.702 ± 0.027. On multifunctional particles, the corresponding scores were 0.482 ± 0.002, 0.521 ± 0.046, and 0.482 ± 0.023. Context-only word and character models remained weak at 0.209 ± 0.019 and 0.243 ± 0.034. Thus contextual features provide a modest average gain, but their benefit is variable and much smaller on the ambiguity-centered subset.

Leave-one-source-out results further test transfer. Word TF-IDF with particle plus context scored 0.667, 0.750, and 0.831 macro-F1 when DKI2, JH, and SKJD were held out, compared with 0.650, 0.656, and 0.681 for particle-majority. The unweighted source means were 0.749 ± 0.067 and 0.662 ± 0.014. Merging topic and focus raised fixed-split particle-majority from 0.670 to 0.748 and particle-plus-context XLM-R from 0.669 to 0.748, without changing the conclusion that much aggregate performance is explained by particle identity. Table 7 reports the fixed-split fine-label results.

Metric	Overall	Merged topic/focus	Secondary	Masked
Inter-annotator κ	0.828	0.938	0.675	0.717
95% CI	0.794–0.859	0.915–0.959	0.611–0.736	0.629–0.799
Unanimous agreement	77.6%	92.4%	58.7%	76.7%
Mean pairwise κ	0.829	0.938	0.677	0.717
Key note	*téh* κ=0.251	*téh* agreement=100%; κ undefined	excludes universal NONE	A vs. B fine

Table 3. Agreement and sensitivity analysis.

Particle	N	Primary profile	Other primary profile / note
mah	342	CONTRAST 89.5%	TOPIC_MANAGEMENT 10.5%
téh	330	FOCUS_EMPHASIS 63.9%	TOPIC_MANAGEMENT 36.1%
atuh	163	STANCE_AFFECT 58.9%	DIRECTIVE_FORCE 41.1%
nya	95	ALIGNMENT_CONFIRMATION 83.2%	SEQUENCE_TURN_MANAGEMENT 16.8%
da	85	EXPLANATION_JUSTIFICATION 100%	—
gé	67	ADDITIVE_SCALAR 100%	—
tah	57	ATTENTION_DEIXIS 100%	—
wéh/waé/baé	104	RESTRICTION_MINIMIZATION 100%	12/72/20 cases

Table 4. Particle-specific primary-function profiles.

Particle	Source	Profile 1	Profile 2
mah	DKI2	81.5% CONTRAST	18.5% TOPIC_MANAGEMENT
mah	JH	94.4% CONTRAST	5.6% TOPIC_MANAGEMENT
mah	SKJD	64.6% CONTRAST	35.4% TOPIC_MANAGEMENT
téh	DKI2	63.3% FOCUS_EMPHASIS	36.7% TOPIC_MANAGEMENT
téh	JH	65.8% FOCUS_EMPHASIS	34.2% TOPIC_MANAGEMENT
téh	SKJD	58.3% FOCUS_EMPHASIS	41.7% TOPIC_MANAGEMENT
atuh	DKI2	33.3% STANCE_AFFECT	66.7% DIRECTIVE_FORCE
atuh	JH	59.7% STANCE_AFFECT	40.3% DIRECTIVE_FORCE
atuh	SKJD	85.7% STANCE_AFFECT	14.3% DIRECTIVE_FORCE
nya	All sources	79% – 90% ALIGNMENT_CONFIRMATION	10% – 21% SEQUENCE_TURN_MANAGEMENT

Table 5. Source-stratified profiles of multifunctional particles.

Comparison	N	Agreement	κ
A vs. B, fine	180	76.7%	0.717
A vs. B, macro	180	82.2%	0.770
A vs. gold, fine	180	15.6%	0.081
B vs. gold, fine	180	17.8%	0.105
A vs. B, clean subset	25	80.0%	0.765

Table 6. Masked reliability and gold-retention metrics.

Model	Input	Acc.	Macro-F1	Multi F1
Global majority	None	0.214	0.032	0.070
Particle majority	Particle	0.856	0.670	0.481
Word TF-IDF	Masked context	0.134	0.104	0.073
Word TF-IDF	Particle + context	0.722	0.627	0.459
Char TF-IDF	Masked context	0.214	0.167	0.106
Char TF-IDF	Particle + context	0.695	0.619	0.407
XLM-R	Masked context	0.203	0.037	0.077
XLM-R	Particle + context	0.854	0.669	0.480

Table 7. Fine-label results on the fixed grouped split; XLM-R values are three-seed means.

4.6 Contextual examples

The quantitative profiles are illustrated by the following examples. English renderings are functional rather than word-for-word translations because the particles often lack a stable lexical equivalent.

***mah*—CONTRAST:** *Tuh da, keuheul aing mah*. "Look, I am the one who is annoyed." The sequence *tuh da* introduces the speaker's response, while *mah* foregrounds the speaker against an implied alternative and makes the contrast recoverable from the local exchange.

***téh*—FOCUS_EMPHASIS:** *Sia téh sok masang togel.* "You are the one who often buys lottery numbers." *Téh* highlights the locally salient participant, resembling an answer or corrective focus.

***atuh*—DIRECTIVE_FORCE:** *Atuh salajengna dihaturanan ka sepuhna panganten pameget.* "Next, we respectfully invite the groom's father to speak." *Atuh* organizes the transition into a formal directive rather than primarily expressing emotion.

***nya*—ALIGNMENT_CONFIRMATION:** *Ibu mah sanés heureuy nya.* "I am not joking, you know." Turn-final *nya* seeks shared orientation and encourages acceptance of the proposition.

***tah*—ATTENTION_DEIXIS:** *Tah kadéngé, kadéngé ku ceuli sia.* "There, you heard it with your own ears." *Tah* directs attention to immediately available evidence.

The examples also clarify why particle-neutral definitions are necessary. The same surface form can participate in distinct discourse operations, while different forms may occur in superficially similar positions. Position and utterance type are useful cues, but neither is deterministic. The final annotation therefore combines local discourse status, recoverable alternatives, speech-act force, and sequential evidence rather than applying one-form–one-function rules.

5 Discussion

5.1 Functional differentiation and reliability

The complete selected inventory shows differentiated particle profiles under one shared

annotation framework. The main contribution lies not in the raw particle–function associations alone, but in the combination of independent annotation of the full dataset, explicit evidence requirements, pre-adjudication reliability analysis, and transparent localization of disagreement. The sensitivity result for *téh* preserves a linguistically meaningful topic-focus distinction in the gold data while also supplying a broader INFORMATION_STRUCTURE level for robust evaluation.

Fine and merged labels serve different purposes. Fine labels preserve distinctions relevant to description and translation, while macro families identify whether errors cross pragmatic boundaries. For *téh*, merged-label agreement was 100%, showing that disagreement was confined to topic versus focus. Particle-level κ is undefined after the merge because all annotations fall into one category. Computational results for *téh* should therefore be read together with the low pre-adjudication reliability of this fine-grained boundary.

5.2 Interactional particles and translation

Atuh and *nya* illustrate why fixed English equivalents are inadequate. *Atuh* contributes affective stance or directive organization depending on the action and sequence, while *nya* supports alignment or turn progression. Concentrated particles also require context for translation: *gé* may correspond to "also" or "even," and restrictive forms may be rendered as "only," "just," or omitted depending on discourse structure. Their contribution spans propositional and interactional meaning rather than simple lexical substitution.

5.3 Lexical shortcuts and benchmark design

The fixed-split XLM-R result shows that high aggregate performance can largely reflect particle identity rather than contextual understanding. The grouped TF-IDF evaluations add an important qualification: sparse context features improve average macro-F1 across alternative group partitions and held-out sources, but gains are smaller and more variable for mah, téh, atuh, and nya. Accuracy alone therefore overstates pragmatic competence; lexical baselines, macro-F1, per-particle scores, and source-transfer results should be reported together.

Weak context-only results should be interpreted cautiously because masking removes part of the linguistic evidence. The context-only XLM-R runs collapsed to two frequent labels under the unweighted training objective; this documents instability on a small, imbalanced target rather than showing that context is universally uninformative. The human diagnostic similarly showed that annotators often recovered broad turn-level actions but not the original fine-grained labels. Occurrence-specific masking prevents unrelated repetitions from being removed, yet the condition remains a perturbation test rather than a replacement for full-token annotation.

5.4 Source bias and resource use

The dataset is a foundational resource, not a representative sample of spontaneous Sundanese. The fixed and grouped-fold partitions keep target dialogue units disjoint, but adjacent-turn windows can share context text across partitions. Leave-one-source-out evaluation measures transfer among the three texts; none of these evaluations separates source from author, genre, period, register, characters, or production conditions. The strong transfer scores therefore demonstrate robustness within this small source set, not population-level generalization.

Future work should add balanced spontaneous interaction with prosody and evaluate imbalance-aware training and prompted multilingual LLMs. We prioritize fully reproducible local baselines here because prompt choice, model version, and proprietary access can otherwise obscure comparison. The release package records occurrence offsets, split checksums, hyperparameter grids, software versions, and model outputs.

6 Limitations

The corpus contains scripted and literary dialogue rather than spontaneous interaction and lacks prosody, gesture, and acoustic realization, although these can carry pragmatic information (Wan et al., 2025). The sources are unequal and confound author, genre, period, register, characters, and production conditions. DKI2 is OCR-derived, SKJD covers selected pages, and *wéh* and *baé* are rare.

The TOPIC_MANAGEMENT–FOCUS_EMPHASIS distinction for *téh* remains interpretive, although merged-label agreement is

complete. Some tables are sparse, observations are clustered within units, speakers, and sources, and masking changes the task by removing the classified particle.

Computational evidence remains limited by corpus size, label concentration, and only three source texts. Five-fold grouped and leave-one-source-out results reduce dependence on one split but cannot disentangle source from genre or author. Shared adjacent-turn context across fixed and grouped-fold partitions limits the independence of context-based evaluations. XLM-R uses one proof-of-concept configuration without imbalance-sensitive training, and no prompted LLM baseline is included.

7 Conclusion

We presented a human-annotated corpus of 1,243 occurrences of ten Sundanese discourse particles from three scripted and literary dialogue sources. Primary-label agreement among three independent annotators reached $\kappa = 0.828$ and increased to 0.938 after merging the topic–focus categories. For *téh*, fine-grained κ was 0.251, while merged-label agreement reached 100%, localizing disagreement to a single information-structural boundary. Source-stratified analysis showed descriptive variation, while the masked diagnostic revealed relatively strong inter-annotator agreement but limited recovery of the original gold labels.

On the fixed split, particle-majority and particle-plus-context XLM-R achieved nearly identical macro-F1, confirming strong lexical shortcuts. Across grouped folds and held-out sources, word TF-IDF obtained modest contextual gains, but performance on multifunctional particles remained lower and more variable. The corpus therefore supports pragmatic evaluation that combines fine and merged labels, exact occurrence masking, grouped partitions, per-particle reporting, and source-transfer analysis.

References

Karin Aijmer. (2002). English Discourse Particles (Vol. 10). John Benjamins Publishing Company. https://doi.org/10.1075/scl.10

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. Computational Linguistics, 34(4), 555–596. https://doi.org/10.1162/coli.07-034-R2

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Muhammad Satrio Wicaksono, Ivan Parmonangan, Ika Alfina, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Septiandri, James Jaya, Kaustubh Dhole, Arie Suryani, Rifki Afina Putri, Dan Su, Keith Stevens, Made Nindyatama Nityasya, Muhammad Adilazuarda, Ryan Hadiwijaya, Ryandito Diandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Damapuspita, Haryo Wibowo, Cuk Tho, Ichwanul Karo Karo, Tirana Fatyanosa, Ziwei Ji, Graham Neubig, Timothy Baldwin, Sebastian Ruder, Pascale Fung, Herry Sujaini, Sakriani Sakti, and Ayu Purwarianti. 2023. NusaCrowd: Open Source Initiative for Indonesian NLP Resources. In Findings of the Association for Computational Linguistics: ACL 2023, pages 13745–13818, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.868

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. https://doi.org/10.18653/v1/2021.emnlp-main.699

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20(1), 37–46. https://doi.org/10.1177/001316446002000104

Liesbeth Degand, Bert Cornillie, and Paola Pietrandrea (Eds.). 2013. Discourse Markers and Modal Particles (Vol. 234). John Benjamins Publishing Company. https://doi.org/10.1075/pbns.234

Kerstin Fischer. 2006. Approaches to Discourse Particles. https://doi.org/10.1163/9780080461588

Bruce Fraser. 1999. What are discourse markers? Journal of Pragmatics, 31(7), 931–952. https://doi.org/10.1016/S0378-2166(98)00101-5

Aditia Gunawan and Evi Fuji Fauziyah. 2021. The particle “ma” in Old Sundanese. Wacana, 22(1), 207. https://doi.org/10.17510/wacana.v22i1.1040

Chan Young Jung. 2025. Same Spelling, Different Functions Mah: Evaluating Language Models’ Understanding of Singlish Particles. In

Proceedings of the 39th Pacific Asia Conference on Language, Information and Computation, pages 15–25, Hanoi, Vietnam. Association for Computational Linguistics.
Manfred Krifka. 2008. Basic Notions of Information Structure. Acta Linguistica Hungarica, 55(3–4), 243–276.
Klaus Krippendorff. 2019. Content Analysis: An Introduction to Its Methodology (4th ed.). SAGE Publications, Inc. https://doi.org/10.4135/9781071878781
Knud Lambrecht. 1994. Information Structure and Sentence Form. In Cambridge Studies in Linguistics. Cambridge University Press. https://doi.org/10.1017/CBO9780511620607
Yuan-Fu Liao, Chih-Chung Kuo, Chao-Shih Huang, Yu-Siang Lan, Han-Chun Lai, and Wen-Han Hsu. 2025. Taiwanese Hakka Across Taiwan Corpus and Formosa Speech Recognition Challenge 2025 – Dapu & Zhao'an Accents. In Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 427–434, National Taiwan University, Taipei City, Taiwan. Association for Computational Linguistics.
Matthew Lombard, Jennifer Snyder-Duch, and Cheryl Campanella Bracken. 2002. Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability. Human Communication Research, 28, 587–604. https://doi.org/10.1111/j.1468-2958.2002.tb00826.x
Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the Era of Large Language Models: A Survey on Datasets, Evaluation, Opportunities and Challenges. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8679–8696, Vienna, Austria. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.425
Franz Müller-Gotama. 1994. The Sundanese particles *teh, mah*, and *tea*. In Papers from the Second Annual Meeting of the Southeast Asian Linguistics Society (pp. 235–250). Arizona State University, Program for Southeast Asian Studies.
Iyo Mulyono Sukandi, and Ice Sutari. 1981. Kata tugas bahasa Sunda. Pusat Pembinaan dan Pengembangan Bahasa, Departemen Pendidikan dan Kebudayaan.
Rahmawati Rahmawati and Ni Gusti Ayu Roselani. 2024. Pola penggunaan partikel penegas bahasa Sunda: Analisis berbasis korpus. Widyaparwa, 52(2), 301–317.
Craige Roberts. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. Semantics and Pragmatics, 5(6), 1–69. https://doi.org/10.3765/sp.5.6
Deborah Schiffrin. 1987. Discourse Markers. In Studies in Interactional Sociolinguistics. Cambridge University Press. https://doi.org/10.1017/CBO9780511611841
Irwan Setiawan and Hung-Yu Kao. 2024. SUSTEM: An Improved Rule-based Sundanese Stemmer. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(6), 1–29. https://doi.org/10.1145/3656342
Purificação Silvano, Mariana Damova, Giedrė Valūnaitė Oleškevičienė, Chaya Liebeskind, Christian Chiarcos, Dimitar Trajanov, Ciprian-Octavian Truică, Elena-Simona Apostol, and Anna Baczkowska. 2022. ISO-based Annotated Multilingual Parallel Corpus for Discourse Markers. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 2739–2749, Marseille, France. European Language Resources Association.
Richard Waltereit. 2001. Modal particles and their functional equivalents: a speech-act-theoretic approach. Journal of Pragmatics, 33(9), 1391–1417. https://doi.org/10.1016/S0378-2166(00)00057-6
I-Ping Wan, Yu-Ju Lai, and Pu Yu. 2025. Computational Approaches to Quantitative Analysis of Pause Duration in Taiwan Mandarin. In Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 116–123, National Taiwan University, Taipei City, Taiwan. Association for Computational Linguistics.
Yu-Che Yen and Siaw-Fong Chung. 2025. Stance and Cohesion: The Use of However and While in AI-Human Argumentative Discourse. In Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 348–357, National Taiwan University, Taipei City, Taiwan. Association for Computational Linguistics.