JURNAL KOMUNIKASI

P-ISSN: 1907-848X, E-ISSN: 2548-7647
Homepage: https://journal.uii.ac.id/jurnal-komunikasi

Introducing CILLCO: A corpus model of vernacular Indonesian as a cultural capital

Devi Ambarwati, Intan Pradita, Bayu Permana Sukma & Yenny Karlina

To cite this article:
Puspitasari, D. A., Pradita, I., Sukma, B. P., & Karlina, Y. (2026). Introducing CILLCO: A corpus model of vernacular Indonesian as a cultural capital. *Jurnal Komunikasi*, 20(1), 161–174. https://doi.org/10.20885/komunikasi.vol20.iss1.art10

© 2026 The Author(s). Published by Program Studi Ilmu Komunikasi, Universitas Islam Indonesia

Published online: April 30, 2026

Submit your article to this journal

To link to this article: DOI: 10.20885/komunikasi.vol20.iss1.art10

View Crossmark data

JURNAL KOMUNIKASI

P-ISSN: 1907-848X, E-ISSN: 2548-7647
Volume 20, Nomor 1, April 2026, Hal 161-174
DOI: 10.20885/komunikasi.vol20.iss1.art10

Check for updates

Introducing CILLCO: A corpus model of vernacular Indonesian as a cultural capital

Devi Ambarwati[1], Intan Pradita[2*], Bayu Permana Sukma[3] & Yenny Karlina[4]

[1,4] Research Centre for Language, Literature and Community, The National Research and Innovation Agency, Jakarta, Indonesia.
[2] English Language Education Department, Universitas Islam Indonesia, Yogyakarta, Indonesia. Email: intan.pradita@uii.ac.id
[3] Linguistics and English Language, Faculty of Humanities, Arts and Social Sciences, Lancaster University, Lancaster, United Kingdom.
[*] Corresponding author

Article Info

Article History
Submit:
Oktober 27, 2025
Accepted:
April 29, 2026
Published:
April 30,2026

Keywords:
Corpus linguistics, sociolinguistics, vernacular Indonesian

Abstract: Despite Indonesia's extraordinary linguistic diversity, the communicative realities of vernacular Indonesian remain significantly underrepresented in existing corpus resources, most of which privilege formal and written registers. This study introduces CILLCO (Corpus of Indonesian Language, Linguistics, and Communities), a multimodal digital corpus jointly developed by the Research Center for Language, Literature, and Community at the National Research and Innovation Agency (BRIN) and the English Department of Universitas Islam Indonesia (UII), with the objective of documenting and analyzing vernacular language varieties as they occur in everyday spoken, digital, and community contexts. CILLCO was constructed using a multi-component design grounded in the Lancaster corpus methodology. CILLCO integrates three primary data streams such as spoken vernacular, digital vernacular, and community narrative texts that were collected through field recording, web scraping, and community collaboration across five macro-regions of Indonesia. As of 2025, the corpus comprises approximately 2 million tokens annotated through a layered framework that includes tokenization, part-of-speech tagging, morphological analysis, sociolinguistic metadata, and pragmatic discourse marking. Results reveal that vernacular Indonesian is characterized by systematic morphosyntactic simplification, a rich inventory of pragmatic particles showing regional variation, orthographic play in digital communication, and extensive code-mixing with English and regional languages. These findings demonstrate that colloquial Indonesian constitutes a coherent, rule-governed linguistic system rather than a deviation from standard norms. The implications of this study extend across corpus linguistics, communication studies, and digital humanities, offering an empirical foundation for investigating language use, identity, and sociocultural change in Indonesia and, more broadly, for decentered, corpus-driven communication research in Southeast Asia.

Copyright © 2026 Authors. This is an open-access article distributed under the terms of the Creative Commons Attribution-ShareAlike 4.0 International License (http://creativecommons.org/licences/by-sa/4.0/)

INTRODUCTION

Indonesia is one of the most linguistically and culturally diverse communication contexts in the world, with more than 700 regional languages in addition to Bahasa Indonesia as a national lingua franca (Ewing, 2014; Sakhiyya & Martin-Anatias, 2023). The country provides an unprecedented setting for investigation into communication and language across social, regional, and digital boundaries. This multilingual ecology affects Indonesians' communicative repertoires (Pradita et al., 2026) and digital communication (Puspitasari et al., 2024). In this rapidly changing context, language contact, vernacularization, and digital communication are central issues for both linguistics and communication studies.

Notwithstanding its richness, Indonesia lacks large-scale corpus resources that systematically represent the communicative realities of vernacular Indonesian. Modern corpora such as the Indonesian Leipzig Corpus, the PAN Localization corpus, and the SEAlang Project primarily focus on standard and written registers of the language (Johannessen et al., 2025). As a result, work based on corpora examining how Indonesians communicate in the most common domains of interpersonal interaction is largely absent. This lacuna, in turn, constrains our empirical knowledge of how digital communication practices affect linguistic behavior, identity formation, and cultural expression across Indonesia's diverse communities.

However, despite their communicative prominence, these varieties have been only minimally represented in formal linguistic resources. Many of Indonesia's existing corpora (e.g., IndonesianWaC, the Leipzig Corpora Collection) are highly institutionalized; composed largely of language found in news portals, Wikipedia, and government publications, they exhibit a bias towards formal registers and an under-representation of informal, interactive, or digital text. This gap has widened further with the rise of digital communication. Media such as Twitter, TikTok, and WhatsApp are new communicative spaces in which language creativity, code-mixing, and multimodality are manifested (Shin et al., 2021). In this context, language in digital spaces becomes a hybrid communicative code in which Indonesian, English, and local languages collide and overlap, constituting a digital vernacular mode of communication (Harun & Kassim, 2025). These mixed genres reflect how users negotiate identity, humor, and effect on Indonesia's networked public sphere. In terms of infrastructure, the corpus that facilitates vernaculars remains underdeveloped in Southeast Asia. The English, Thai, and Vietnamese resources have been built, in part, through international collaboration (Robin et al., 2017). Indonesian resources rely primarily on lexical databases and machine translation datasets (Khairunnisa et al., 2023; Cahyawijaya et al., 2023). Resources, such as the PAN Localization Indonesian Corpus and the digital lexicon of *Kamus Besar Bahasa Indonesia*, are built with a focus on computational linguistics rather than on sociolinguistic or cultural studies. There is a paucity of publicly available annotated corpora representing spoken and vernacular Indonesian, compromising computational empirical research on variation, contact, and discourse.

To address these gaps, the Corpus of Indonesian Language, Literature and Community (CILLCO) was inaugurated in 2024 based on a partnership between the

Research Center for Language, Literature and Community of the National Research and Innovation Agency (BRIN), together with the Department of English, Universitas Islam Indonesia (UII). The work seeks to create an open, multimodal digital corpus with broad community involvement, documenting and analyzing vernacular Indonesian as informal forms occurring in talk-in-interaction that violate the prescriptive norms of contemporary broadcast media but are nonetheless the social default mode of communication for millions of Indonesians. CILLCO is an example of a global corpus initiative like the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA), as well as regional efforts such as Korpus Bahasa Indonesia (KOIN). Global corpus initiatives aim to document and capture the use of certain languages by retrieving documentation of language use. To this extent, the BNC documented British English, the COCA documented the use of American English, and the KOIN documented the use of Indonesian. CILLCO, however, offers a comprehensive view of Indonesia's diversity, especially ethnographic participation, through digital data and computational annotation. The corpus aims to represent language as a social and communicative phenomenon. CILLCO functions as a medium through which identity is enacted and authority negotiated in face-to-face interactions and online indexicalities (Snell, 2018). This project will therefore benefit linguistic research as well as a broader scope of communication studies in Indonesia, and it is consistent with BRIN's national strategic agenda to develop inclusive research infrastructures. It is committed to applied linguistics, digital communications, and language technology.

The novelty of this study lies in the development of CILLCO as the first large-scale, multimodal, and ethically designed corpus of vernacular Indonesian that systematically captures spoken, digital, and community language data across all major sociolinguistic regions of Indonesia. Unlike previous Indonesian corpora, which have been largely restricted to formal or written registers, CILLCO integrates ethnographic fieldwork, digital data collection, and community participation within a single, TEI-compliant annotation infrastructure. This makes CILLCO uniquely positioned to advance empirical research on vernacular language use, digital communication practices, and sociolinguistic variation in a multilingual context, offering a methodological contribution that bridges corpus linguistics, communication studies, and digital humanities.

METHOD

The methodology used to collect data from the Corpus of Indonesian Language, Literature, and Community (CILLCO) adopted the linguistic corpus development methodology proposed by corpus scholars at Lancaster University (McEnery & Hardie, 2012). Unlike the currently available Indonesian corpora, which tend to feature a preponderance of standardized written forms, CILLCO is designed with an emphasis on vernacular variation as sub-standard, community-embedded instantiations across spoken and digital genres.

Data Collection

This corpus was initiated in 2024 as a national collaborative research project by the Research Center for Language, Literature, and Community under the National Research and Innovation Agency (BRIN), in collaboration with the English Department of Universitas Islam Indonesia (UII). The design was guided by a

collaborative team consisting of linguists, computational linguists, sociolinguists, and cultural researchers. Methodologically, CILLCO exclusively deals with vernacular Indonesian, thereby requiring separate design principles, data sources, and annotation protocols. As of October 2025, CILLCO consists of roughly 2 million tokens from spoken, written, and digital sources. The goal for the corpus in Phase II (2026–2028) is 50 million tokens, spanning Indonesia's sociolinguistic zones. The corpus is created using open-source tools such as Sketch Engine and AntConc and is provided via cloud infrastructure on BRIN. All texts have been saved as plain text in UTF-8 format (. txt) with accompanying XML-based metadata.

CILLCO is designed to be a comparable multimodal corpus comprising three primary components. First, the Spoken Vernacular Component consists of transcripts of spontaneous conversations, interviews, and oral narratives collected in naturalistic settings. Second, the Digital Vernacular Component comprises online communications in social media, community blogs, YouTube comments, chat groups, and digital storytelling sites. Third, the Community Narrative Component encompasses local texts from grassroots organizations, cultural associations, and informal publications that assert community voice and identity. All elements are stratified internally by region, register, and speaker profile for comparative and sociolinguistic purposes. The main subcomponents are summarized in Table 1.

Data Source

Table 1
Structure of CILLCO by component and subcorpus

Component Type	Tokens	Percentage	Primary Source
Digital vernacular (WhatsApp, social media)	1,200,000	60%	WhatsApp group chats, Twitter posts
Spoken vernacular (recorded speech)	500,000	25%	Interviews, natural conversation
Community & Literary Data	300,000	15%	Local narratives, online storytelling
Total	**2,000,000**	**100%**	—

Source: Authors' own elaboration based on CILLCO corpus data.

Each sub-corpus represents not only a distinct mode of communication but also a sociolinguistic register, enabling micro-level comparisons (e.g., youth speech vs. community ritual discourse) and macro-level trend analysis across Indonesia. Recorded speech was obtained from field recordings and digital ethnography. Researchers participated in community meetings in both urban and rural environments of Java, Kalimantan, Sulawesi, and Nusa Tenggara. Recruitment was conducted using a snowball sampling technique to capture variance in age, gender, and occupation. Our data collection followed the principle of naturalness (Carter & McCarthy, 2017). Our recordings captured natural speech rather than elicited responses.

Research Participants

Participants were provided with informed consent forms prior to each recording session, clearly indicating the anonymity protocols and the intended use of the data. Audio was recorded (44.1 kHz, 16-bit) with hand-held portable recorders. Recordings were manually transcribed using analysis tools such as Praat, then orthographically normalized to Indonesian phonotactics while maintaining vernacular markers (e.g., *gak, nih, dong*). Transcribers were also trained to mark pauses, slight overlaps, laughter, and discourse particles as has been done elsewhere using Santa Barbara Corpus conventions (Carter & McCarthy, 2017).

The digital component was gathered through web scraping and API data extraction within a strict ethical and legal framework. Publicly available data sources (Twitter/X, YouTube, Wattpad, and Reddit Indonesia) were crawled for Indonesian-language content using automatic language detection (LangID.py). User-generated identifying personal information was omitted to meet privacy standards. CILLCO accounts for the computer-mediated counterparts of online vernaculars, including spelling play (*akuh*, *gemoy*, *kepo*), emotive reduplication (*lohhh*, *bangettt*), and multimodal markers (emojis, hashtags). Orthographic variation was controlled by applying light normalization to the data, while maintaining expressive forms for stylistic analysis. Metadata fields capture platform, date, and user demographics (if public), facilitating diachronic and cross-platform comparison.

Data Analysis: Corpus Annotation

To reflect community expression, CILLCO comprises non-standard literary and informational genres such as local magazines, zines, and oral transcriptions from local storytellers. These materials were gathered in collaboration with NGOs, student groups, and cultural centers. These hybrid registers, which blend oral vernacular and written form, are aimed at illustrating how communities assert local identity and cultural resistance through Indonesian. CILLCO employs a layered annotation framework that combines automatic tagging with manual correction. Annotation proceeds in four layers. First, tokenization and part-of-speech (POS) tagging are performed automatically using the IndoNLP Tagger (Cahyawijaya et al., 2023), which has been retrained on vernacular corpora to account for informal morphosyntax. Reduplication, cliticization (*aku-nya, nggak-lah*), and expressive affixes (*-in or -nya*) are handled by custom rules; the modified version identified 49 POS tags, which the team named POS TAG Indo-Verna. Second, a morphological and lemmatization layer implements a two-level lemmatizer to extract both standard (e.g., *tidak* for *nggak*) and colloquial forms (e.g., *gue, lo, dong*), enabling contrastive analysis between the two. Third, sociolinguistic annotation tags each utterance with region, speaker demographics (age, sex, and education), interaction type, and medium; these metadata are embedded in TEI-compliant XML headers following the Pannonia Corpus format (Robin et al., 2017) as shown in Figure 1. Fourth, selected sub-corpora receive additional pragmatic and discourse annotation for discourse markers, stance expressions, and pragmatic particles, with tag sets adapted from (Carter & McCarthy, 2017) and each marker tagged with a discourse function such as emphatic, mitigative, or turn-taking.

Figure 1
Source code sample for annotation

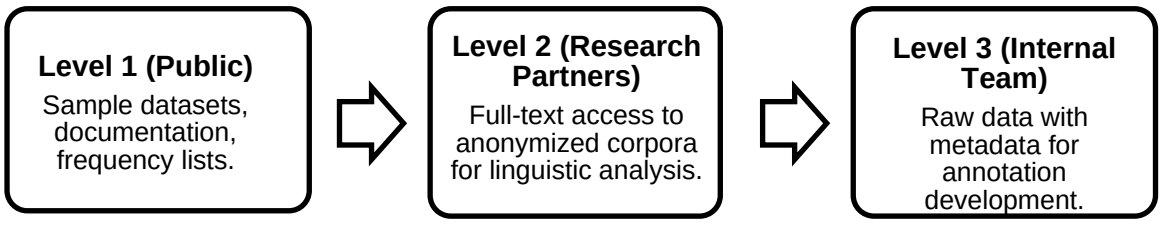
```
<teiHeader>
 <fileDesc>
  <titleStmt>
   <title>CILLCO Spoken Vernacular: Jakarta_Youth_01</title>
   <respStmt>
    <name>BRIN-UII Field Team</name>
    <resp>Transcription and annotation</resp>
   </respStmt>
  </titleStmt>
  <extent>12,438 words</extent>
  <publicationStmt>
   <availability status="restricted">Ethically cleared; anonymized</availability>
  </publicationStmt>
 </fileDesc>
 <profileDesc>
  <langUsage>
   <language ident="idv">Vernacular Indonesian</language>
  </langUsage>
  <particDesc>
   <person sex="female" age="21" edu="university"/>
  </particDesc>
  <settingDesc>
   <region>Jakarta</region>
   <domain>Urban informal speech</domain>
  </settingDesc>
 </profileDesc>
</teiHeader>
```

Source: BRIN Dataverse https://hdl.handle.net/20.500.12690/RIN/B9LJYY

Selected subcorpora receive additional annotation for discourse markers, stance expressions, and pragmatic particles. Tagsets are adapted from (Carter & McCarthy, 2017), covering items such as *kan, dong, nih, lah, deh*, etc. Each marker is tagged with a discourse function (e.g., emphatic, mitigative, turn-taking).

CILLCO is hosted on BRIN's Open Language Infrastructure (OLI) cloud platform. Researchers access the corpus through a secure web interface that integrates concordancing, keyword analysis, and collocation search functionalities. Access levels are tiered:

Figure 2
CILLCO's access level



Source: CILLCO Access Level Framework (the authors' research model) .

The CILLCO team employs stratified sampling to obtain representation across five macro-regions: Western Indonesia (Sumatra), Central (Java), Eastern (Sulawesi), Bali–Nusa Tenggara, and Kalimantan. In each region, data were categorized by age group (teenagers 13–19, adults 20–45, older adults 46+) and communication mode (oral, written paper-based, and digital). This model allows for the equal representation of demographic and geographic dimensions. The sampling was based on (Biber & Conrad, 2019). The distribution of texts reflects the occurrence of communicative activities in everyday situations. For example, urban digital data make up roughly 30% of the corpus as their usage is widespread in contemporary communication. The resulting corpus can be weighted or collapsed during the analysis to reflect different sociolinguistic orientations.

Data Trustworthiness

Quality assurance involved multiple stages. First, transcription validation was carried out by two annotators independently, and the inter-annotator agreement (Cohen's kappa) was over 0.92. Second, annotation verification was conducted through manual checking of POS and lemma on a random sample of annotated data. Third, metadata applicability was ensured by running scripts to check conformance to the XML schema and correct encoding. Fourth, a linguistic review was conducted at a senior level to ensure representativeness, including regional and gender balance.

CILLCO upholds a rigorous language research ethics referred to in the BRIN's Ethical Guidelines for Linguistic and Cultural Data which aligns with standard international practices, including those set out in the Linguist Data Consortium (LDC) Code of Ethics. All voice and digital data are captured with informed consent, rendered anonymous, and hosted on encrypted servers. All identifying information, such as names, contact information, and social media handles, is expunged. Private messages or limited posts are not included in digital data. The researchers retained only the minimal necessary data to preserve linguistic content while discarding nonlinguistic information about individuals. CILLCO's infrastructure integrates three technical layers. First, the data storage layer uses cloud-based storage with redundant backups maintained on BRIN's national server, with file formats including .txt for text and .wav for audio. Second, the processing layer employs a Python-based preprocessing pipeline incorporating tokenization, normalization, and tagging modules. Third, the user interface layer was built with Flask and Vue.js, enabling interactive querying and visualization; the features provided include concordances and collocation search, keyword-in-context (KWIC) display, frequency and dispersion plots, and metadata filtering by region and time period. Among CILLCO's original methodological aspects, the paradigm of interinstitutional collaboration is worth mentioning. BRIN contributes infrastructure, ethical oversight, and corpus management skills. UII provides academic development, linguistic analysis, and student research training. This collaboration promotes the sustainability of human resource development in corpus linguistics as a fledgling field of study in Indonesia.

RESULTS AND DISCUSSION

Overview of Corpus Composition

At the end of the first stage of its development in 2025, CILLCO contains around 2 million tokens, including digital, spoken, and community vernacular data. The corpus is a collection of 2,480 individual plain-text files, categorized by type and region. The data distribution also mirrors Indonesia's population density and linguistic diversity: most spoken-language data comes from Java, Kalimantan, Sulawesi, Bali-Nusa Tenggara, and Sumatra, whereas digital contributions are spread across the archipelago. A more detailed summary is given in Table 2.

Table 2

CILLCO corpus composition by region and mode (2025)

Region	Spoken (%)	Digital (%)	Community (%)	Total (%)
Java	32	20	15	67
Sumatra	18	8	3	29
Sulawesi	9	5	3	17
Bali–Nusa Tenggara	6	4	2	12
Kalimantan	4	2	1	7
Total	**69**	**39**	**24**	—

Source: Authors' own elaboration based on CILLCO corpus data.

Note. Percentages exceed 100% because several texts overlap in categories (e.g., spoken–digital hybrid materials).

This distribution suggests a clear bias towards spoken data (roughly 55% of total tokens), which confirms CILLCO's pledge for genuine oral vernacular representation. One third of the corpus is made up of digital texts, showing how prominently they feature in online (Indonesian) communication. Even if limited, community materials provide important contexts for cultural and narrative expression.

Linguistic Profiling of Vernacular Indonesian

The in-house quantitative evaluation provides evidence of stable linguistic features that differentiate vernacular Indonesian from the standard language. The next subsections present high-level results from a preliminary analysis using AntConc and Sketch Engine (2025 dataset).

Lexical Frequency and Informality

The most common words in the corpus are, as expected, from previous discussions on spoken Indonesian (Ewing, 2005). High-frequency items *aku, nggak, aja, nih, dong, loh,* and *kan*—particles that are rare or downplayed in formal writing. The ratio 5:1 of *nggak* to *tidak* in terms of relative frequency reflects the considerable preponderance that informal negation has in everyday language. First- and second-person pronouns gue and lo, respectively, appear prominently in urban sub-corpora (Jakarta, Bandung), while aku and kamu (me and you) dominate rural or mixed registers. Thus, it is evident from lexical dispersion analysis that vernacular particles have a high cross-register frequency, spreading across the whole text but not so much as to form a cluster, which implies comfortable usage. This is consistent with the distribution observed in studies of English discourse particles

(Beier et al., 2025). Moreover, confirms that Indonesian vernacular markers have a pragmatically fluctuating status.

Morphosyntactic Simplification

Simplification patterns consistent with spoken and digital language were found in morphological analysis. Suffixed reduction- including the elision of meN- and ber- prefixes - is present in 46% of verbal tokens, e.g., *pikir* (thinks) instead of *memikirkan* (to think), *main (play)* instead of *bermain (to play*). Also, object markers and prepositions tend to be dropped. This happened with *aku kasih kamu uang (I give you money)* instead of *aku memberi kamu uang* (I give you money). These observations are in line with (Harun & Kassim, 2025), who argues that Indonesian vernacular grammar operates on the principle of economy of expression, rather than deficit, and is fine-tuned to enhance processing speed at fast speaking rates and in digital writing. This generalization is supported by CILLCO, which provides datasets depicting how linguistic simplification indexes communicative situations requiring speed, intimacy, and reduced monitoring.

Pragmatic Particles and Discourse Organization

A distinctive feature of Indonesian vernacular is its extensive inventory of pragmatic particles. CILLCO's annotation identifies over 40 recurrent particles, including *kan, lah, deh, dong, loh, ya*, and *nih*. Their frequencies vary regionally, as shown in Table 3.

Table 3

Top vernacular particles by region (per 10,000 words)

Particle	Java	Sumatra	Sulawesi	Bali	Overall Rank
kan	84	71	59	44	1
dong	68	42	36	22	2
loh	54	33	40	27	3
nih	49	30	24	16	4
lah	41	27	18	35	5

Source: Authors' own elaboration based on CILLCO corpus data.

The findings reveal that *kan* is the most widespread among these particles, serving as a marker of common ground or alignment of expectations. Particles like *dong* and *nih* are used to mark emphasis and immediacy in high interpersonal engagement, as is characteristic of spoken interaction. These findings echo pragmatic systems in other Asian varieties, including Japanese *ne* and *yo*, or Mandarin *ba*, and point to typological similarities in discourse organization. CILLCO's pragmatic annotation also enables collocational investigation on particle clustering (e.g.*, kan ya, dong, deh*), which illuminates the layered nature of Indonesian interactional styles. Such constellations are important for the modeling of conversational politeness and stance-taking, which have not yet been systematically studied in corpus-based Indonesian research.

Orthographic Play and Digital Vernaculars

Through digital subcorpora, we see a world of orthographic play and paralinguistic code. Users bend spelling to

express affect, rhythm, and humor, such as *bangettt*, gemesss, or *please dong santuyy*. CILLCO's n-gram analysis also finds that stretched vowels and consonants (N=3, 7; e.g., *lohh, hahhaaa*) appear in 3.7% of all social media tokens. Those features function as digital prosody markers (Beier et al., 2025). Moreover, it compensates for the lack of intonation and gesture in computer-mediated writing. The corpus also shows extensive code-mixing between Indonesian and English, particularly in youth speech (*ngeliat outfit dia tuh so aesthetic banget*). This bilingual hybridity constitutes a typical case of translanguaging, where speakers seamlessly draw on an array of linguistic resources to mark cosmopolitan identity. There is quantitative evidence for this in CILLCO: 14% of the tokens in the Indonesian digital component are English lexical items, which is relatively significant compared with other previous Indonesian corpora (Cahyawijaya et al., 2023). A keyword and collocation analysis shows that the thematic fields of emotions, relations, humor, and social criticism dominate Indonesian vernacular discourse. For example, *cowok*, *cewek*, and chat are common collocates of *baper* ('overly emotional'), highlighting affective expression in romantic and online spaces. *Receh* ('trivial', 'cheap'), on the other hand, collocates with *meme, konten, and TikTok, displaying digital humor as one of the most dominant genres* in internet-based Indonesian culture. These collocational biases highlight how language reflects sociocultural values, emotional expressivity, relational negotiation, and shared play. CILLCO's information, therefore, supplements qualitative ethnographic accounts of Indonesian youth culture, grounding sociocultural interpretation in empirical evidence.

Regional and Sociodemographic Variation

CILLCO metadata enables nuanced regional and demographic analysis. Initial findings reveal strong geo-stratification in phonetic and lexical selection. For example:

a. Eastern Indonesian varieties (Makassar, Manado) show constant retention of *torang* (we) and kita (us), not found in Standard Indonesian.
b. Javanese-influenced varieties that have a mixed politeness system relying on *sampean* or *kamu* (you) according to the status of the interlocutor.
c. Youth vernaculars: These youth vernaculars do merge along pan-regional lines. Urban teens from the Sundanese and Betawi communities soften the profanity "*Anjing*" (dog) into "*Anjir*," "*Anjiir*," "*Anjay*," "*Njir*," and enrich the laughter "*lol*" into "*wkwkwk*," "*hahaha*," "*hehehe*." These phenomena are digitally mediated and documented through tokenization.

Sociodemographic observation of men shows stylistic diversity by gender. Female speakers employ expressive reduplication (*gemes bangettt, cutee lohh*) and male speakers prefer taboo or slang expressions (*anjay, bro, goblok*). These results are consistent with cross-linguistic variation in gender style (Johannessen et al., 2025; Coates, 2015), but within the particular moral-discursive ecology of contemporary Indonesia, on the one hand, local norms of modesty, and on the other, a preference for humor in online self-presentation.

Comparative Discussion: Vernacular vs. Standard Indonesian

CILLCO's data call into question time-honored assumptions about Indonesian diglossia. The Baku/Non-Baku continuum is a variant of the traditional distinction made between Standard and Non-standard Indonesian. Quantitative comparisons with formal corpora (e.g., the Leeds Indonesian Corpus) show that many so-called "non-standard" features occur regularly and systematically, which implies grammatical stability rather than randomness. For example, the absence of the prefix *meN* exhibits certain syntactic and pragmatic restrictions with high-frequency verbs frequently occurring in spontaneous speech. Sentence-final particles likewise play fixed discourse roles, such as indicating agreement, hesitancy, or a wish. Thus, CILLCO confirms the intuition that colloquial Indonesian is a systematic language variety rather than simply a degenerate version of standard Indonesian.

This conclusion supports the principle of vernacular regularity (Harun & Kassim, 2025), which states that the least self-conscious styles of speech contain the most organized patterns overtly structured at a linguistic level. Applying this to Indonesian, it implies that the vernaculars are acceptable depending on how the language really operates in society (Pradita et al., 2026).

New Media and Vernacular Communication in Indonesia

The corpus CILLCO shows that digital communication has emerged as a major space for linguistic and cultural practices among Indonesians. Scrutiny of discussions in WhatsApp group chats, social media posts, and online comments shows that digital communication not only accelerates language change but also changes how the speaker's express identity, emotion, and authority. Informal Indonesian, widely known as *Bahasa gaul digital*, serves as a resource for communication across genres, whereby users build social relationships and negotiate politeness in digitally mediated environments.

The first observation concerns the hybridity of language forms and codes across digital platforms. Code-mixing across Indonesian, English, and regional languages (Javanese, Sundanese, Banjar, Bugis) constitutes what Ewing, (2014) describes as polylingual repertoires. Code-mixing is also not haphazard but strategically patterned: English lexical insertions, for example, are used to indicate cosmopolitan orientation or humor, whereas regional items express solidarity and intimacy. This resonates with Blommaert's (2010) conceptualization of superdiversity, in which communicative competence entails the skillful exploitation of multiple linguistic and semiotic resources in liquid digital spaces.

Another common pattern is the use of pragmatic markers and stance particles, namely loh, sih, dong, nih, and kan, which serve as resources for interpersonal alignment and contextualization. Speakers deploy micro-features of talk — typically associated with face-to-face interaction, that are now recoverable in text-based exchanges, rather than physical co-presence, as part of the interactional work being carried out (Kirner-Ludwig, 2022). For example, the particle *kan* typically induces mutual understanding, and dong indicates insistence or playful emphasis. This linguistic practice is illustrative of the way Indonesians are transforming norms of spoken discourse within a multimodal digital environment and traversing into the

phenomenon of written orality (Carter & McCarthy, 2017). In the digital realm, CILLCO's data found that equally significant is the main theme of vernacular sovereignty. In online communities, WhatsApp groups in particular, everyday communicators, rather than institutional elites, control the circulation of discourse. This more democratic linguistic authority represents what is termed participatory culture, in which sense-making becomes dispersed and jointly controlled (Sakhiyya & Martin-Anatias, 2023). CILLCO's data also reveals how digital vernacular communication is a venue for informal knowledge sharing and micro-level civic engagement, concurring with research on the digital publics of Southeast Asia.

From a communication theory perspective, these findings emphasize the media-ecological nature of Indonesian discourse. The ubiquity of messaging apps and social media tools fosters hybrid communication ecologies in which written, spoken, and visual modes intertwine. Users exchange short, iterative messages that sustain social co-presence and collective affect (Ewing, 2014). In linguistic terms as well, these practices push the limits of Indonesian as a communicative system, from its prescriptive definition to an active, multi-coordinate vernacular network. Methodologically, CILLCO shows that corpus-based methods can contribute to communication sciences by providing empirical evidence of discourse behavior in digital environments. The integration of ethnographic participation with metadata tagging and multimodal annotation enables researchers to trace the flow of communicative practices in the cultural eco-system, ultimately by and between agents within social institutions across history, culture, and technology. This integration of corpus linguistics and communication studies addresses the demands of computational discourse ethnography (Ewing, 2014; Johannessen et al., 2025) as a lens for understanding meaning-making in digital life.

Therefore, the CILLCO research provides evidence of both local and global orientations in Indonesian digital communication. It foregrounds the mundane communicative agency of users who rearticulate national and cultural identities in online interaction through language play, code-switching, and multimodal self-presentation. For communication studies as well, this corpus offers not only language data, but a lens through which to trace Indonesia's changing digital discourse ecology.

CONCLUSION

CILLCO introduces a corpus design tailored to Indonesia's multilingual sociolinguistic landscape. Through its ethnographic fieldwork, digital scraping, and use of TEI-compliant metadata, it secures both representativeness and replicability. The hybrid annotation approach, which combines automated tagging with manual sociolinguistic coding, tends to exemplify how computational methods can be made compatible with humanistic inquiry. Thus, the corpus is a digital humanities effort in linguistics at the intersection of sampling rigor and cultural sensitivity. CILLCO's initial analyses also provide much to savor in the anatomy of the Indonesian vernaculars. It can be shown from lexical, morphosyntactic, and pragmatic properties that informal varieties are not arbitrary eruptions of language, but rather systematically organized linguistic systems.

Despite its contributions to digital humanities, CILLCO faces a set of methodological challenges. These include the imbalanced representation of

Indonesian regions. The regional under-representation of digital communication from East Indonesia, such as Papua and Maluku, is due to logistical constraints. Further studies are recommended to address these challenges by expanding regional representation and developing more advanced language identification systems.

Acknowledgement

This study would like to thank the National Research and Innovation Agency and the Faculty of Social and Cultural Sciences, Universitas Islam Indonesia, for funding the research.

Declaration of Interest

The authors declare that there are no conflicts of interest.

REFERENCES

Beier, E., Cohn, M., Trammel, T., Ferreira, F., & Zellou, G. (2025). Marking prosodic prominence for voice assistant and human addressees. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *51*(6), 986–1003. https://doi.org/10.1037/xlm0001396

Biber, D., & Conrad, S. (2019). *Register, genre, and style* (Second edition). Cambridge University Press. https://doi.org/10.1017/9781108686136

Blommaert, J. (2010). The sociolinguistics of globalization. Cambridge: Cambridge University Press.

Cahyawijaya, S., Lovenia, H., Aji, A. F., Winata, G., Wilie, B., Koto, F., Mahendra, R., Wibisono, C., Romadhony, A., Vincentio, K., Santoso, J., Moeljadi, D., Wirawan, C., Hudi, F., Wicaksono, M. S., Parmonangan, I., Alfina, I., Putra, I. F., Rahmadani, S., … Purwarianti, A. (2023). NusaCrowd: Open source initiative for indonesian NLP resources. *Findings of the Association for Computational Linguistics: ACL 2023*, 13745–13818. https://doi.org/10.18653/v1/2023.findings-acl.868

Carter, R., & McCarthy, M. (2017). Spoken grammar: Where are we and where are we going? *Applied Linguistics*, 38(1), 1–20. https://doi.org/10.1093/applin/amu080

Coates, J. (2015). Women, men and language: A sociolinguistic account of gender differences in language (3rd ed.). *Routledge*. https://doi.org/10.4324/9781315645612

Ewing, M. C. (2014). Language endangerment in indonesia. *International Journal of Education*, *8*(1).

Harun, N. Z., & Kassim, S. J. (2025). *The urban vernacular in southeast asia: Settlement as serendipity* (1st ed.). Routledge. https://doi.org/10.4324/9781003413868

Johannessen, J. B., Lundquist, B., Rodina, Y., Tengesdal, E., Kaldhol, N. H., Türker, E., & Fyndanis, V. (2025). Cross-linguistic effects in grammatical gender assignment and predictive processing in L1 Greek, L1 Russian, and L1 Turkish speakers of Norwegian as a second language. *Second Language Research*, *41*(1),

217–259. https://doi.org/10.1177/02676583241227709

Khairunnisa, S. O., Chen, Z., & Komachi, M. (2023). Dataset enhancement and multilingual transfer for named entity recognition in the indonesian language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(6), 1–21. https://doi.org/10.1145/3592854

Kirner-Ludwig, M. (2022). Data collection methods applied in studies in the journal *Intercultural Pragmatics* (2004–2020): A scientometric survey and mixed corpus study. *Intercultural Pragmatics*, *19*(4), 459–487. https://doi.org/10.1515/ip-2022-4002

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*.

Pradita, I., Sutrisno, A., Hardjanto, T. D., & Gündüz, C. (2026). *Triangulated corpus-informed assessment of multilingual academic writing*.

Puspitasari, D. A., Karlina, Y., Hernina, H., Kurniawan, K., Sutejo, S., & Danardana, A. S. (2024). Language choices and digital identity of high school student text messages in the new capital city of indonesia: Implication for language education. *International Journal of Language Education*, *8*(1). https://doi.org/10.26858/ijole.v8i1.63833

Robin, E., Götz, A., Pataky, É., & Szegh, H. (2017). Translation studies and corpus linguistics: introducing the pannonia corpus. *Acta Universitatis Sapientiae, Philologica*, *9*(3), 99–116. https://doi.org/10.1515/ausp-2017-0032

Sakhiyya, Z., & Martin-Anatias, N. (2023). Reviving the language at risk: A social semiotic analysis of the linguistic landscape of three cities in Indonesia. *International Journal of Multilingualism*, *20*(2), 290–307. https://doi.org/10.1080/14790718.2020.1850737

Shin, D.-S., Cimasko, T., & Yi, Y. (2021). Multimodal Composing in K-16 ESL and EFL education: Multilingual perspectives. In *Multimodal Composing in K-16 ESL and EFL Education: Multiling. Perspectives* (p. 207). Springer Singapore. Scopus. https://doi.org/10.1007/978-981-16-0530-7

Snell, J. (2018). Solidarity, stance, and class identities. *Language in Society*, 47(5), 665–691. https://doi.org/10.1017/S0047404518000970