



## Building ACLA-PH: A Multiplatform Corpus of Philippine Gayspeak

John Paulo Q. Herrera<sup>1</sup>

<sup>1</sup>De La Salle University, Manila, Philippines

Corresponding Author Email: [john.paulo.herrera@dlsu.edu.ph](mailto:john.paulo.herrera@dlsu.edu.ph)

DOI: <https://doi.org/10.36892/ijlls.v8i6.2827>

**APA Citation:** Herrera, J. P. Q. (2026). Building ACLA-PH: A Multiplatform Corpus of Philippine Gayspeak. *International Journal of Language and Literary Studies*, 8(6).86-98. <https://doi.org/10.36892/ijlls.v8i6.2827>

**Received:**

24/08/2026

**Accepted:**

07/10/2026

**Keywords:**

Corpus  
Linguistics,  
Digital  
Discourse,  
Multimodal  
Corpus,  
Philippine  
Gayspeak,  
Queer  
Linguistics

### Abstract

The rapid expansion of digitally mediated communication has transformed the nature of linguistic data and challenged traditional approaches to corpus construction. These challenges are particularly evident in documenting Philippine Gayspeak, a dynamic and multilingual linguistic repertoire characterized by continual lexical innovation, multimodal communication, and platform-dependent interaction. This study introduces ACLA-PH (A Corpus Linguistic Approach to Philippine Gayspeak), a multiplatform corpus designed to document contemporary Philippine Gayspeak across publicly accessible digital environments. Employing a methodological research design, the study examines the processes involved in corpus design, data selection, sampling, annotation, metadata architecture, multimodal representation, and ethical data collection. The findings demonstrate that constructing corpora for rapidly evolving digital discourse requires methodological principles that extend beyond traditional corpus linguistic practices. Building on these findings, the study proposes the Rolling Corpus Model, which conceptualizes corpus development as a process of planned and continual expansion, and the Adaptive Digital Corpus Framework (ADCF), which integrates authenticity, platform sensitivity, multimodal integration, layered annotation, ethical reflexivity, and adaptive sustainability. Although developed through Philippine Gayspeak, these methodological innovations offer a framework that may be adapted for documenting multilingual, multimodal, and rapidly evolving forms of digital discourse, thereby contributing to ongoing discussions on the future of corpus construction in the digital age.

### 1. Introduction

The rapid expansion of digitally mediated communication has fundamentally transformed the ways in which language is produced, circulated, and interpreted. Social media platforms, livestreaming services, podcasts, online gaming, and other digital environments have created new communicative spaces characterized by multimodality, multilingualism, and continuous interaction among geographically dispersed communities. Unlike traditional written or spoken discourse, contemporary digital communication extends beyond text to include images, videos, emojis, gestures, hashtags, audiovisual editing, and platform-specific affordances that collectively contribute to meaning-making. These developments have expanded the scope of linguistic inquiry while simultaneously challenging methodological assumptions that have long guided corpus-based research.

Corpus linguistics has established itself as one of the most influential methodological approaches in contemporary linguistics by enabling the systematic investigation of authentic language through electronically searchable corpora (Biber et al., 1998; McEnery & Hardie, 2012; Sinclair, 1991). Through carefully designed corpora, researchers have examined lexical patterns, grammatical variation, discourse organization, language change, and register

differences across diverse communicative contexts. Central to this tradition are principles such as authenticity, representativeness, systematic sampling, and transparent annotation, which ensure that corpus findings are grounded in naturally occurring language. While these principles remain fundamental to corpus linguistic inquiry, many were developed using relatively stable written and spoken language varieties whose communicative boundaries, genres, and production contexts are comparatively well defined.

The emergence of digital discourse complicates many of these methodological assumptions. Language now unfolds across interconnected digital platforms where communication is shaped by multimodal resources, algorithmic visibility, participatory cultures, and rapidly evolving interactional practices (Androutsopoulos, 2015; Zappavigna, 2018). Linguistic innovations frequently emerge through online communities, circulate across multiple platforms, and acquire new meanings within remarkably short periods of time. Consequently, conventional notions of corpus representativeness and stability become increasingly difficult to sustain when documenting language that is dynamic, platform-dependent, and continuously evolving. These developments suggest that corpus construction can no longer be understood solely as the compilation of textual data but must also account for the technological and social environments in which communication takes place.

These methodological challenges are particularly evident in the study of Philippine Gayspeak, often referred to as *swardspeak*. Emerging within Filipino LGBTQIA+ communities, Philippine Gayspeak is a dynamic linguistic repertoire through which speakers negotiate identity, solidarity, humor, resistance, and community belonging (Casabal, 2008; Manalansan, 2003; Pascual, 2016). Rather than constituting a fixed vocabulary, it is continually reshaped through lexical innovation, multilingualism, semantic play, and performative language practices as it circulates across changing social and digital contexts. From the perspective of queer linguistics, such linguistic practices are understood not merely as communicative resources but as sites where gendered and sexual identities are constructed, negotiated, and recognized through interaction (Bucholtz & Hall, 2005; Leap, 1996; Motschenbacher, 2010). The increasing visibility of Philippine Gayspeak across social media, livestreams, podcasts, entertainment programs, and other online communities therefore provides a compelling context for examining how corpus methodologies may be adapted to document rapidly evolving digital discourse.

Although previous scholarship has substantially advanced understanding of Philippine Gayspeak by documenting its lexical creativity, sociocultural functions, and role in identity construction (Casabal, 2008; Manalansan, 2003; Pascual, 2016; Ulla et al., 2024), comparatively little attention has been devoted to the methodological question of how this linguistic repertoire should be systematically represented through corpus linguistic approaches. Existing studies have primarily relied on interviews, lexical inventories, discourse examples, or relatively small datasets. While these approaches have generated valuable insights into the linguistic and cultural dimensions of Philippine Gayspeak, they are less capable of capturing its scale, multimodality, platform diversity, and continual transformation within digitally mediated environments. This methodological gap underscores the need for corpus construction strategies that recognize language as an evolving social and technological phenomenon rather than as a static collection of textual artifacts.

Against this backdrop, the present study introduces ACLA-PH (A Corpus Linguistic Approach to Philippine Gayspeak) as a multiplatform corpus designed to document contemporary Philippine Gayspeak across digitally mediated environments. Rather than focusing exclusively on the resulting corpus, the study examines the methodological decisions underlying its construction, including corpus design, platform selection, sampling, annotation, metadata architecture, multimodal representation, and ethical considerations. Building upon these experiences, the study proposes the Rolling Corpus Model and the Adaptive Digital Corpus Framework (ADCF) as methodological contributions to corpus linguistics. While developed through the context of Philippine Gayspeak, these frameworks are intended to

inform future corpus construction involving multilingual, multimodal, and rapidly evolving forms of digital discourse. In doing so, the study contributes to ongoing discussions on how corpus linguistics can remain responsive to the methodological demands of contemporary digital communication.

Specifically, this study sought to answer the following questions:

1. What methodological considerations informed the construction of ACLA-PH?
2. How can corpus construction be adapted for multilingual, multimodal, and rapidly evolving digital discourse?
3. What methodological framework emerged from the development of ACLA-PH?

## **2. Literature Review**

### **Corpus Linguistics and Corpus Construction**

Corpus linguistics has become one of the most influential methodological approaches in contemporary linguistic research by enabling the systematic investigation of authentic language through electronically searchable corpora. Rather than relying solely on linguistic intuition or isolated examples, corpus-based research examines naturally occurring language across large collections of texts, allowing researchers to identify recurring lexical, grammatical, and discourse patterns that may not be evident through introspection alone (Biber et al., 1998; McEnery & Hardie, 2012; Sinclair, 1991). This empirical orientation has substantially advanced research in language variation, phraseology, register analysis, discourse studies, language change, and applied linguistics, establishing corpus methods as a cornerstone of modern linguistic inquiry.

The value of corpus linguistics, however, depends largely on the quality of corpus construction. Sinclair (1991) emphasized authenticity as a defining characteristic of corpus-based research, arguing that corpora should consist of naturally occurring language rather than researcher-generated examples. Likewise, Biber et al. (1998) stressed representativeness and balanced sampling, maintaining that corpus findings are meaningful only when the data adequately reflect the language variety under investigation. McEnery and Hardie (2012) further argued that corpus design encompasses numerous methodological decisions, including data selection, annotation, metadata organization, corpus architecture, and analytical procedures. Consequently, corpora are not neutral repositories of language but carefully designed research resources whose methodological foundations directly influence subsequent linguistic interpretation.

Although these principles remain central to corpus linguistics, they were largely developed using relatively stable written and spoken language varieties. Newspapers, literary texts, academic writing, and transcribed speech generally possess identifiable communicative boundaries, relatively stable genres, and predictable production contexts. Such characteristics facilitate systematic sampling and corpus compilation because the linguistic units being documented are comparatively stable over time. As communication increasingly shifts toward digitally mediated environments, however, many of these assumptions become less straightforward. Digital discourse is dynamic, multimodal, and continuously evolving, raising questions about how authenticity, representativeness, and corpus stability should be conceptualized in online settings. Despite these foundational principles, existing corpus methodologies provide limited guidance for constructing corpora that represent multilingual, multimodal, and rapidly evolving digital discourse.

### **Digital Discourse and Multimodal Communication**

The rapid growth of digital communication has transformed not only the ways people use language but also the methodological challenges associated with documenting linguistic practices. Digital discourse differs from traditional written and spoken communication because meaning is produced through the interaction of multiple semiotic resources, including written text, speech, gesture, emojis, typography, audiovisual editing, hashtags, hyperlinks, and

audience participation. Rather than functioning independently, these resources collectively shape interpretation within technologically mediated environments.

Androutsopoulos (2015) argues that digital communication has fundamentally reconfigured multilingual practices by enabling speakers to draw upon diverse linguistic resources across interconnected online communities. Communication is no longer confined to geographically bounded speech communities but instead develops through networked interaction among users who continually negotiate language across platforms. Similarly, Zappavigna (2018) demonstrates that social media platforms actively shape discourse through technological affordances such as searchability, hashtags, visibility algorithms, and participatory interaction. These affordances influence not only how language is produced but also how linguistic forms circulate, gain prominence, and acquire new social meanings.

These developments present important implications for corpus construction. Traditional corpora frequently privilege written or spoken language as the primary unit of analysis, whereas digital communication requires greater attention to multimodality and interactional context. Meaning often emerges not from lexical items alone but from their relationship with visual design, gesture, timing, audience engagement, and platform-specific conventions. Consequently, corpus methodologies designed for digital discourse must preserve contextual information alongside textual data if they are to represent communicative practices accurately. These considerations suggest that corpus construction should extend beyond textual compilation toward documenting the broader communicative ecology within which language is produced and interpreted. Consequently, methodological frameworks capable of preserving these multimodal and platform-specific characteristics remain underdeveloped within corpus linguistics.

### **Queer Linguistics and Philippine Gayspeak**

Queer linguistics has consistently challenged essentialist understandings of language, gender, and sexuality by demonstrating that linguistic practices are central to the construction and negotiation of identity. Rather than treating gender or sexuality as fixed social categories reflected in language, queer linguistic scholarship emphasizes that identities emerge through discourse, interaction, and social practice (Bucholtz & Hall, 2005; Leap, 1996; Motschenbacher, 2010). From this perspective, language functions not merely as a means of communication but also as a resource through which speakers perform identities, establish community affiliations, negotiate belonging, and resist dominant social norms.

This perspective is reinforced by Agha's (2007) concept of *enregisterment*, which explains how recurring linguistic features become socially recognized as markers of particular identities, communities, or styles of speaking. Enregisterment highlights the relationship between linguistic form and social meaning, demonstrating that repeated use within particular communities gradually establishes recognizable registers associated with specific identities. Likewise, Zimman (2014) argues that language actively participates in the construction of gendered and sexual identities rather than simply reflecting pre-existing categories. Together, these theoretical perspectives underscore the importance of examining language within its interactional and sociocultural contexts.

Within the Philippine context, these theoretical insights are particularly relevant to Philippine Gayspeak, often referred to as *swardspeak*. Manalansan (2003) describes Philippine Gayspeak as a creative and dynamic linguistic repertoire through which Filipino gay men negotiate community, identity, and belonging. Subsequent studies have documented its lexical innovation, semantic creativity, multilingual borrowing, and sociocultural functions (Casabal, 2008; Pascual, 2016). More recently, Ulla et al. (2024) observed that Philippine Gayspeak has expanded beyond LGBTQIA+ communities into mainstream digital communication, where expressions circulate widely through social media, entertainment, and influencer culture. As these expressions move across different communicative environments, they continually acquire new meanings, making Philippine Gayspeak an especially dynamic object of linguistic inquiry.

Despite these advances, existing scholarship has focused primarily on describing linguistic features, documenting lexical innovation, or examining identity construction. Comparatively less attention has been devoted to the methodological question of how Philippine Gayspeak should be systematically documented through corpus linguistic approaches capable of representing its digital, multilingual, multimodal, and rapidly evolving nature. These observations suggest the need for corpus methodologies that can systematically document queer linguistic practices without sacrificing their interactional and sociocultural complexity.

### **Research Gap**

Taken together, the preceding scholarship reveals an important methodological gap. Corpus linguistics provides well-established principles for corpus construction, while digital discourse studies highlight the multimodal and platform-dependent nature of online communication. Queer linguistics further demonstrates that linguistic forms derive meaning through interaction, identity performance, and social practice, and scholarship on Philippine Gayspeak has documented the repertoire's creativity, sociocultural significance, and expanding digital presence. Yet these strands of research have rarely been integrated into a coherent methodology for constructing corpora representing dynamic queer digital discourse.

Most studies of Philippine Gayspeak continue to rely on interviews, lexical inventories, discourse excerpts, or relatively small datasets. Although these approaches have generated valuable linguistic and sociocultural insights, they provide limited guidance for constructing large-scale corpora capable of documenting continual linguistic innovation across multiple digital platforms. Likewise, existing corpus methodologies offer relatively little direction for preserving multimodal interaction, platform metadata, and evolving discourse practices within a single corpus architecture.

The present study addresses this gap by introducing ACLA-PH (A Corpus Linguistic Approach to Philippine Gayspeak) as a methodological case study in corpus construction. Rather than proposing a new theory of Philippine Gayspeak itself, the study examines how corpus methodologies may be adapted to document multilingual, multimodal, and rapidly evolving digital discourse. Building upon the experience of constructing ACLA-PH, the study proposes the Rolling Corpus Model and the Adaptive Digital Corpus Framework (ADCF) as methodological contributions that extend beyond Philippine Gayspeak and offer practical guidance for future corpus development involving other dynamic online language communities.

## **3. Methodology**

### **Research Design**

This study employed a methodological research design to examine the construction of ACLA-PH (A Corpus Linguistic Approach to Philippine Gayspeak), a multiplatform corpus developed to document contemporary Philippine Gayspeak across digitally mediated environments. Rather than investigating a particular linguistic feature or testing a specific hypothesis, the study focused on the methodological processes involved in corpus construction and the principles that informed decisions throughout its development. Accordingly, the research examined corpus design, platform selection, sampling, annotation procedures, metadata architecture, multimodal representation, and ethical considerations associated with collecting and organizing publicly accessible digital discourse.

The article reports the methodological component of the larger ACLA-PH dissertation project and therefore focuses on corpus construction rather than on the linguistic findings generated from the corpus.

A methodological research design was considered appropriate because the primary objective of the study was not to analyze the linguistic features of Philippine Gayspeak themselves but to develop a systematic framework for documenting them. The resulting corpus therefore served both as a research resource and as a methodological case study illustrating

how corpus linguistics may be adapted for multilingual, multimodal, and rapidly evolving forms of digital communication. Throughout corpus development, established principles of corpus linguistics—including authenticity, representativeness, and systematic documentation—were integrated with emerging perspectives from digital discourse studies and queer linguistics to address the methodological challenges presented by contemporary online communication.

### **Corpus Design**

The design of ACLA-PH was guided by the recognition that Philippine Gayspeak circulates across multiple communicative environments rather than within a single speech community or textual genre. Consequently, the corpus was conceptualized as a multiplatform corpus capable of representing linguistic practices occurring across diverse forms of publicly accessible digital discourse. Instead of privileging one platform, genre, or communicative mode, the corpus was designed to capture variation across different digital environments while maintaining methodological consistency in data collection and annotation.

The architecture of ACLA-PH was informed by six interrelated design principles that emerged during corpus development: authenticity, platform sensitivity, multimodal integration, layered annotation, ethical reflexivity, and adaptive sustainability. These principles collectively guided decisions regarding corpus composition, metadata organization, annotation procedures, and long-term maintenance. Rather than functioning independently, they operated simultaneously throughout the research process, ensuring that corpus construction remained responsive to the dynamic nature of Philippine Gayspeak and contemporary digital discourse.

### **Data Sources**

Data were collected between August 2025 and August 2026 from publicly accessible digital platforms where Philippine Gayspeak is actively produced, circulated, and negotiated. This period was selected to capture contemporary linguistic practices across multiple digital environments while allowing sufficient variation in lexical innovation, interactional practices, and platform-specific language use.

Selection emphasized communicative environments characterized by sustained interaction, linguistic creativity, and active user participation. Primary sources included publicly accessible content from TikTok, YouTube, X (formerly Twitter), podcasts, livestreams, and digital entertainment programs. Particular attention was given to Drag Race Philippines, whose performances, backstage interactions, and audience engagement provide productive contexts for lexical innovation, identity performance, and metalinguistic commentary. Public comments associated with these materials were likewise incorporated to capture patterns of uptake, negotiation, and reinterpretation among audiences.

The inclusion of multiple platforms reflects the understanding that Philippine Gayspeak, as one linguistic repertoire used within and beyond Filipino LGBTQIA+ communities, no longer circulates exclusively within localized communities but is also increasingly visible across mainstream digital culture. Expressions frequently move between entertainment media, influencer content, social networking platforms, and everyday online interaction. A multiplatform corpus was therefore necessary to document the diversity of communicative contexts in which Philippine Gayspeak is produced, circulated, and interpreted.

Table 1 summarizes the principal platform-specific components of ACLA-PH. At the time of analysis, the Drag Race Philippines Season 1 and Dogshow Divas subcorpora had been fully processed and analyzed, whereas the TikTok and X subcorpora remained under continuing expansion. This modular architecture enables the corpus to accommodate ongoing data collection while preserving transparency regarding the developmental status of each component.

**Table 1: Platform-Specific Components of ACLA-PH**

| Subcorpus                      | Communicative Environment                  | Current Corpus Size               |
|--------------------------------|--------------------------------------------|-----------------------------------|
| Drag Race Philippines Season 1 | Reality television / performance discourse | 89,830 processed tokens           |
| Dogshow Divas                  | YouTube conversational podcast             | 46,329 processed tokens           |
| TikTok                         | Short-form public video discourse          | Corpus under continuing expansion |
| X (formerly Twitter)           | Public text-based digital discourse        | Corpus under continuing expansion |

### Sampling Strategy

Sampling followed a purposeful platform-based strategy rather than traditional genre-based sampling. Conventional corpus construction often seeks proportional representation across predefined genres, such as newspapers, academic writing, or conversation. Such an approach is less appropriate for Philippine Gayspeak because linguistic innovation frequently transcends conventional genre boundaries and instead emerges through platform-specific interactional practices.

Three criteria guided corpus selection. First, linguistic productivity prioritized communicative environments in which Philippine Gayspeak is actively created, modified, and disseminated. Second, interactional richness emphasized platforms that encourage meaningful audience participation, allowing the corpus to capture naturally occurring exchanges between speakers and audiences. Third, public accessibility ensured that corpus construction complied with ethical principles by restricting data collection to publicly available materials. Together, these criteria prioritized communicative relevance rather than statistical proportionality, enabling the corpus to represent contexts where Philippine Gayspeak performs significant linguistic and social functions.

### Corpus Construction and Annotation

The corpus was designed as a multiplatform digital resource consisting of platform-specific subcorpora. At the time of analysis, completed components included the Drag Race Philippines Season 1 lexical corpus (89,830 processed tokens) and the Dogshow Divas lexical corpus (46,329 processed tokens), while additional TikTok and X subcorpora remained under continuing expansion.

**Table 2: Overview of ACLA-PH**

| Component              | Description                                                                                                |
|------------------------|------------------------------------------------------------------------------------------------------------|
| Data collection period | August 2025–August 2026                                                                                    |
| Platforms              | TikTok, YouTube, X, Podcasts, Livestreams, Drag Race Philippines                                           |
| Corpus architecture    | Multiplatform, multimodal digital corpus                                                                   |
| Annotation software    | ELAN                                                                                                       |
| Corpus software        | AntConc                                                                                                    |
| Metadata fields        | Platform, speaker category, interactional context, discourse function, code-switching, multimodal features |

Table 2 summarizes the overall architecture of ACLA-PH. Rather than functioning as a single homogeneous dataset, the corpus was designed as a multiplatform, multimodal resource composed of platform-specific subcorpora representing different communicative environments. Audiovisual data were transcribed and temporally aligned using ELAN, while corpus analyses, including frequency, concordance, collocation, and keyword analyses, were conducted using AntConc. Each corpus record was accompanied by layered metadata to facilitate both quantitative and qualitative analyses across platforms.

In addition to lexical transcription, ACLA-PH incorporates layered metadata documenting platform, source, speaker category, interactional context, discourse function, code-switching, and multimodal features. Audiovisual materials were temporally aligned using ELAN, while lexical retrieval and corpus analyses were conducted using AntConc. This architecture enables both quantitative corpus analyses and qualitative investigations of identity performance, interaction, and digital discourse.

Corpus construction proceeded through several interconnected stages designed to preserve both analytical rigor and contextual richness. Audiovisual materials were first transcribed using standardized orthographic conventions to ensure consistency across corpus files. Where communication involved substantial multimodal interaction, transcripts were temporally aligned using ELAN, allowing relationships among speech, gesture, gaze, facial expression, prosody, and other semiotic resources to be preserved.

Following transcription, corpus files were processed in AntConc to support lexical frequency analysis, concordance generation, collocation analysis, and keyword identification. Structural and interpretive annotation layers were subsequently integrated into a unified corpus architecture. Beyond lexical and grammatical information, each corpus record included metadata describing the platform, speaker category, interactional context, discourse function, code-switching practices, and relevant multimodal features. This layered annotation strategy enabled the corpus to support both quantitative analyses of recurring linguistic patterns and qualitative investigations of discourse, identity performance, and interaction.

### **Ethical Considerations**

Because ACLA-PH documents language drawn from digitally mediated environments, ethical considerations informed every stage of corpus construction. Only materials that were publicly accessible at the time of collection were included. Private conversations, restricted communities, and password-protected content were excluded. Where appropriate, personally identifiable information was minimized to balance methodological transparency with respect for individual privacy.

Ethical reflexivity also extended beyond issues of data access. Philippine Gayspeak represents a linguistic repertoire associated with historically marginalized communities whose language practices carry important cultural and social meanings. Accordingly, corpus construction sought to document these practices respectfully, recognizing that methodological choices influence how communities and their linguistic resources are represented. The objective was not merely to compile linguistic data but to develop a corpus that supports rigorous linguistic inquiry while remaining attentive to the social contexts from which the data emerge.

### **Data Analysis**

Data analysis combined corpus linguistic procedures with qualitative thematic analysis to examine both linguistic patterns and the methodological implications of corpus construction. Following transcription, audiovisual materials were aligned in ELAN to preserve temporal relationships among speech, gesture, gaze, facial expression, prosody, and other multimodal resources. This alignment enabled linguistic forms to be interpreted within their broader communicative contexts rather than as isolated textual units. The resulting transcripts were subsequently processed in AntConc to generate lexical frequency lists, concordances, collocations, and keyword analyses, facilitating the systematic identification of recurring linguistic patterns across corpus files.

Beyond lexical analysis, annotation records, field notes, and corpus construction decisions were subjected to iterative qualitative coding to identify recurring methodological issues encountered during corpus development. Codes were continually refined through constant comparison until broader conceptual categories emerged. These categories were then synthesized thematically, leading to the identification of six interrelated methodological principles that informed the development of ACLA-PH: authenticity, platform sensitivity,

multimodal integration, layered annotation, ethical reflexivity, and adaptive sustainability. Together, these analyses formed the basis for the Rolling Corpus Model and the Adaptive Digital Corpus Framework (ADCF) presented in the Results and Discussion.

#### **4. Results and Discussion**

##### **Emerging Principles for Digital Corpus Construction**

Analysis of the corpus construction process revealed three major methodological insights that informed the development of ACLA-PH. These findings concern (1) emerging principles for digital corpus construction, (2) the Rolling Corpus Model, and (3) the Adaptive Digital Corpus Framework. While authenticity, representativeness, and systematic annotation remain fundamental, the corpus development process revealed that these principles alone are insufficient for representing multilingual, multimodal, and rapidly evolving forms of online communication. Instead, corpus construction required continuous methodological adaptation in response to platform diversity, interactional complexity, and ongoing linguistic innovation.

One of the most significant findings concerns the role of authenticity. Consistent with Sinclair's (1991) emphasis on naturally occurring language, ACLA-PH prioritized discourse produced independently of research activities. However, corpus construction demonstrated that authenticity in digital environments extends beyond simply collecting naturally occurring texts. Because Philippine Gayspeak frequently emerges through interaction among content creators and audiences, authentic representation also requires preserving the communicative environments in which language is produced. Audience comments, livestream exchanges, podcasts, and entertainment programs therefore function not merely as sources of lexical data but as interactional spaces where linguistic forms are negotiated, reinterpreted, and disseminated. This finding suggests that authenticity in digital corpus construction should be understood as preserving naturally occurring communicative practices rather than simply collecting authentic textual materials.

A second finding concerns the importance of platform sensitivity. During corpus construction, it became evident that identical lexical forms often performed different communicative functions depending on the digital environment in which they occurred. For example, expressions such as *mother*, *charot*, and *slay* frequently shifted in pragmatic meaning across TikTok, Drag Race Philippines, and podcast discourse, where their interpretation depended on platform-specific interactional norms, audience participation, and performance practices. Such variation reflects the technological affordances and communicative norms of individual platforms rather than lexical change alone. This observation supports Androutsopoulos's (2015) view that digital communication is fundamentally shaped by networked environments and reinforces Zappavigna's (2018) argument that platform affordances influence both language production and circulation.

The construction of ACLA-PH also highlighted the importance of multimodal integration. Many instances of Philippine Gayspeak derived their communicative force not from lexical items alone but from the interaction of language with gesture, facial expression, intonation, audiovisual editing, typography, and audience response. Expressions such as *mother*, *slay*, or *charot*, for example, often conveyed meanings that depended upon performance rather than vocabulary in isolation. This finding reinforces previous observations in digital discourse research that meaning is distributed across multiple semiotic resources rather than being contained exclusively within textual forms. Accordingly, multimodal annotation using ELAN became essential for preserving interactional context and ensuring that linguistic analyses remained sensitive to the performative dimensions of digitally mediated communication.

##### **The Rolling Corpus Model**

A second major outcome of the study is the development of the Rolling Corpus Model, which reconceptualizes corpus construction as a process of planned and transparent expansion rather than one-time compilation. Throughout the development of ACLA-PH, it became

increasingly evident that Philippine Gayspeak evolves at a pace that exceeds the assumptions underlying many traditional corpus projects. New lexical items continually emerge through viral trends, entertainment programs, influencer culture, and online communities, while existing expressions undergo semantic extension, pragmatic reinterpretation, or gradual decline. A corpus frozen at a single point in time therefore risks becoming progressively less representative of the linguistic repertoire it seeks to document.

The Rolling Corpus Model addresses this challenge by conceptualizing corpus development as an evolving research infrastructure. Rather than replacing earlier materials, successive corpus versions preserve previous stages of linguistic development while incorporating newly emerging forms and communicative practices. This approach allows researchers to investigate both synchronic patterns within particular periods and diachronic processes of lexical innovation and change. More importantly, it shifts the focus of corpus construction from producing a finished linguistic archive toward maintaining a sustainable methodological resource capable of adapting alongside the language it documents.

This reconceptualization extends existing corpus linguistic practice by recognizing that digital discourse cannot reasonably be represented through permanently fixed datasets. Instead, corpus sustainability depends upon transparent procedures for expansion, documentation, and version control. The Rolling Corpus Model therefore provides a practical framework for documenting language varieties characterized by continual innovation while preserving the historical record necessary for longitudinal investigation.

### **The Adaptive Digital Corpus Framework**

Building upon the methodological findings generated during corpus construction, this study proposes the Adaptive Digital Corpus Framework (ADCF) as a general framework for documenting dynamic forms of digital discourse. The framework synthesizes six interrelated methodological principles that consistently informed decisions throughout the development of ACLA-PH: authenticity, platform sensitivity, multimodal integration, layered annotation, ethical reflexivity, and adaptive sustainability. Rather than functioning as sequential procedures, these principles operate simultaneously throughout corpus construction and collectively guide decisions regarding data collection, annotation, corpus maintenance, and analytical interpretation.

Although developed through Philippine Gayspeak, the ADCF also contributes to gender and sexuality studies by demonstrating how corpus methodologies can document the linguistic practices of marginalized communities without reducing them to isolated lexical inventories. Drawing on queer linguistics, the framework recognizes that linguistic forms derive meaning through identity performance, interaction, and social practice rather than through vocabulary alone (Bucholtz & Hall, 2005; Leap, 1996; Motschenbacher, 2010). Likewise, Agha's (2007) concept of *enregisterment* explains how recurring linguistic features become socially recognized as markers of particular communities and identities, while Zimman (2014) illustrates how language participates in the construction and negotiation of gendered and sexual identities. By preserving interactional context, multimodal features, and platform-specific practices, the ADCF enables researchers to investigate how queer identities are performed, negotiated, and transformed across digitally mediated environments.

Beyond its contribution to Philippine Gayspeak research, the ADCF is proposed as a methodological framework that may inform future corpus construction involving dynamic forms of digital discourse. The same principles may inform corpus construction involving youth language, multilingual social media, internet vernaculars, Indigenous language revitalization, online fan communities, and other rapidly evolving forms of digital communication. As online discourse continues to diversify, corpus methodologies must likewise become more adaptive, transparent, and sustainable. The framework proposed here therefore extends beyond a single linguistic repertoire to contribute to broader discussions concerning the future development of corpus linguistics in digitally mediated contexts.

## **5. Conclusion**

The rapid expansion of digitally mediated communication has fundamentally reshaped the linguistic environments in which contemporary language is produced, circulated, and interpreted. These developments present important methodological challenges for corpus linguistics, particularly when documenting linguistic repertoires characterized by continual innovation, multimodality, and platform-dependent interaction. As one dynamic linguistic repertoire used within and beyond Filipino LGBTQIA+ communities, Philippine Gayspeak exemplifies these challenges because its meanings emerge not only through lexical forms but also through performance, interaction, and the technological affordances of contemporary digital platforms. Consequently, documenting Philippine Gayspeak requires corpus methodologies that move beyond traditional assumptions of stable textual data toward approaches capable of representing language as an evolving social and communicative practice.

This study introduced ACLA-PH (A Corpus Linguistic Approach to Philippine Gayspeak) as a methodological response to these challenges. By examining the processes involved in corpus design, platform selection, sampling, annotation, metadata architecture, multimodal representation, and ethical data collection, the study suggests how corpus construction may be adapted to document multilingual, multimodal, and rapidly evolving forms of digital discourse. The methodological insights generated during corpus development resulted in two principal contributions: the Rolling Corpus Model, which conceptualizes corpus construction as a process of planned and transparent expansion, and the Adaptive Digital Corpus Framework (ADCF), which integrates authenticity, platform sensitivity, multimodal integration, layered annotation, ethical reflexivity, and adaptive sustainability into a coherent methodological framework.

Beyond its contribution to Philippine Gayspeak research, the study offers methodological implications for future corpus construction and contributes to ongoing discussions in corpus linguistics, digital discourse studies, and queer linguistics by illustrating how methodological decisions shape the representation of marginalized linguistic communities. Rather than replacing qualitative approaches, the proposed framework complements them by enabling systematic large-scale analyses while preserving the interactional and sociocultural contexts through which meaning is produced. The findings suggest that adaptive corpus methodologies may provide a useful approach for documenting digitally mediated language practices characterized by multilingualism, multimodality, and continual linguistic change. However, the broader applicability of the Rolling Corpus Model and the ADCF should be regarded as a methodological proposition that requires further investigation across other linguistic repertoires and digital communities.

Several limitations should nevertheless be acknowledged. The corpus draws exclusively on publicly accessible digital discourse and therefore does not capture language practices occurring within private conversations, closed online communities, or face-to-face interaction. Moreover, the corpus reflects the platforms selected during the study and may therefore underrepresent linguistic practices occurring in less visible, emerging, or platform-specific communities. The purposeful sampling strategy, while appropriate for documenting communicatively productive environments, may also introduce sampling bias by representing some discourse communities more extensively than others. In addition, although annotation procedures followed transparent protocols, the interpretation of discourse functions and multimodal features necessarily involved researcher judgment. Finally, the Adaptive Digital Corpus Framework proposed in this study has been developed from a single corpus and has not yet been independently validated through application to other linguistic repertoires or research contexts. Furthermore, the continual evolution of Philippine Gayspeak means that no corpus can be regarded as permanently complete. Ongoing corpus expansion, methodological

refinement, and validation will therefore remain necessary as new linguistic forms emerge and communicative practices continue to evolve.

Future research may evaluate and refine the proposed framework through its application to other multilingual and marginalized language communities, compare corpus development across different digital platforms, or integrate computational techniques such as automated annotation and machine learning to support larger-scale corpus construction. Such investigations would further strengthen methodological approaches to documenting the increasingly dynamic forms of language that characterize contemporary digital communication. More broadly, the present study suggests that corpus construction should continue to evolve alongside digital communication if corpus linguistics is to remain responsive to increasingly dynamic and socially situated language practices.

## 6. References

- Agha, A. (2007). *Language and social relations*. Cambridge University Press.
- Androutsopoulos, J. (2015). Networked multilingualism: Some language practices on Facebook and their implications. *International Journal of Bilingualism*, 19(2), 185–205. <https://doi.org/10.1177/1367006913489198>
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.
- Bucholtz, M., & Hall, K. (2005). Identity and interaction: A sociocultural linguistic approach. *Discourse Studies*, 7(4–5), 585–614. <https://doi.org/10.1177/1461445605054407>
- Casabal, N. G. (2008). Gay language: Defying the structural limits of English language in the Philippines. *Kritika Kultura*, (11).
- Leap, W. L. (1996). *Word's out: Gay men's English*. University of Minnesota Press.
- Manalansan, M. F., IV. (2003). *Global divas: Filipino gay men in the diaspora*. Duke University Press.
- McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- Motschenbacher, H. (2010). *Language, gender and sexual identity: Poststructuralist perspectives*. John Benjamins.
- Pascual, G. R. (2016). Sward speak (Gay lingo) in the Philippine context: A morphological analysis. *International Journal of Advanced Research in Management and Social Sciences*, 5(12).
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.
- Ulla, M. B., Macaraeg, J. M., & Ferrera, R. E. (2024). “What’s the word? That’s the word!”: Linguistic features of Filipino queer language. *Cogent Arts & Humanities*, 11(1).
- Zappavigna, M. (2018). *Searchable talk: Hashtags and social media metadiscourse*. Bloomsbury Academic.
- Zimman, L. (2014). The discursive construction of sex: Recovering gender in sociolinguistics. In S. Ehrlich, M. Meyerhoff, & J. Holmes (Eds.), *The handbook of language, gender, and sexuality* (2nd ed., pp. 13–34). Wiley-Blackwell.

## Author Bio

John Paulo Q. Herrera is a faculty member in the Department of English, Faculty of Arts and Letters, University of Santo Tomas. He previously served as Pedagogical Lead of the Faculty of Arts and Letters and as Thesis Coordinator of the English Language Studies Program. He earned his MA in English Language Education from De La Salle University,

**Building ACLA-PH: A Multiplatform Corpus of Philippine Gayspeak**

Manila, Philippines and is currently pursuing a PhD in Applied Linguistics in the same institution. His research interests include corpus linguistics, queer linguistics, discourse analysis, and digital language practices.