World Journal of Corpus Linguistics
Vol. 4, No. 2-2024, Page 44-52
Available online at https://pusatpublikasi.com/index.php/corpus/

A CORPUS-BASED ANALYSIS OF LEXICAL PATTERNS AND CULTURAL VALUES IN TRADITIONAL MALAY PANTUN

Yosefina Elsiana Suhartini
STIE Karya Ruteng
E-mail: tiniyosefin@gmail.com

Abstract
Pantun, a traditional Malay quatrain, stands as one of the most enduring cultural and linguistic expressions in the Malay-speaking world, encapsulating moral wisdom, ecological awareness, and aesthetic refinement transmitted across generations. This study employs a corpus-based analytical framework to examine the linguistic structures and cultural meanings embedded in traditional Malay pantun. The research aims to identify dominant lexical features, collocation patterns, and semantic domains that define pantun discourse, while interpreting how these linguistic patterns encode and sustain Malay cultural values and worldview. A specialized corpus of fifty authenticated classical pantun was compiled from verified public domain sources and analyzed using AntConc 4.0. Quantitative analysis revealed recurrent lexical clusters associated with natural imagery, interpersonal harmony, and ethical reflection. Concordance and collocation analyses uncovered systematic relationships between lexical choice and structural organization, particularly between the sampiran (introductory couplet) and maksud (message couplet), reflecting metaphorical correspondence central to pantun aesthetics. These findings demonstrate that pantun operates as a linguistic medium for articulating ecological sensitivity, moral reasoning, and social cohesion. Methodologically, the study highlights the value of corpus linguistic approaches in exploring oral poetic traditions, while contributing to digital humanities initiatives for documenting and preserving intangible cultural heritage through computational means.

Keywords: *corpus linguistics, pantun, Malay poetry, traditional literature, cultural linguistics, digital humanities.*

INTRODUCTION

Corpus linguistics has emerged as a powerful methodological framework for investigating language use through systematic analysis of authentic textual data. As McEnery and Hardie (2012) articulate, corpus linguistics involves the study of language data at scale through computer-aided analysis of extensive collections of transcribed utterances or written texts. This empirical approach enables researchers to identify linguistic patterns that may not be readily apparent through traditional close reading or introspective analysis, thereby providing quantitative evidence to support or challenge theoretical claims about language structure and use.

The application of corpus linguistic methods to literary texts, particularly to traditional poetic forms, represents a relatively recent development in digital humanities scholarship. While corpus stylistics has made significant advances in analyzing novels and dramatic texts, the systematic corpus-based investigation of oral poetry and traditional verse forms remains underexplored. This gap is particularly evident in the study of non-Western poetic traditions, where cultural-linguistic complexity and structural features distinct from European verse forms

present unique methodological challenges.

Pantun, the traditional Malay quatrain, occupies a unique position in world literature as one of the oldest and most sophisticated forms of oral poetry in Southeast Asia. Historically documented for at least 400 years in written records, pantun predates this written evidence by centuries in oral tradition. Wilkinson (1907) defined pantun as a quatrain in which the first line rhymes with the third and the second with the fourth, though this definition captures only its formal structure. More comprehensively, pantun represents a complex literary form characterized by its distinctive two-part structure: the sampiran or pembayang (opening couplet that sets imagery) and the maksud (closing couplet that conveys meaning), linked through intricate patterns of semantic and phonological correspondence.

The cultural significance of pantun in Malay society extends far beyond its literary-aesthetic function. As documented by Harun Mat Piah (2001) and other scholars of Malay literature, pantun serves as a primary vehicle for transmitting moral values, social norms, and traditional knowledge across generations. Pantun permeates virtually every aspect of traditional Malay social life, from formal ceremonies to everyday conversation, from courtship rituals to dispute resolution. The ability to compose and deliver pantun skillfully remains a marker of cultural competence and eloquence in Malay communities.

Despite its cultural importance and literary sophistication, pantun has received limited attention from corpus linguists and computational literary scholars. Existing studies of pantun have predominantly employed traditional literary analysis, anthropological approaches, or structural linguistic methods without leveraging the quantitative capabilities of corpus analysis. This represents a significant research gap, particularly given the potential of corpus methods to reveal patterns in lexical selection, semantic organization, and cultural conceptualization that may not be immediately visible through conventional analytical approaches.

This study addresses this gap by conducting a systematic corpus-based investigation of traditional Malay pantun. The research objectives are threefold: (1) to construct a specialized corpus of classical Malay pantun from authenticated public domain sources; (2) to identify and quantify dominant lexical patterns, collocations, and semantic fields through corpus analytical methods; and (3) to interpret these linguistic patterns in relation to Malay cultural values, worldview, and aesthetic principles. By combining quantitative corpus analysis with qualitative cultural interpretation, this study contributes to both corpus linguistics methodology and the preservation of intangible cultural heritage through digital means.

LITERATURE REVIEW

Corpus Linguistics: Theoretical Foundations

Corpus linguistics fundamentally represents a methodological rather than theoretical orientation to language study. As Biber and Conrad (2019) emphasize, corpus linguistics provides a set of procedures and tools for investigating language empirically, drawing on authentic usage data rather than constructed examples or introspective judgments. The field emerged from the confluence of structural linguistics, computational methods, and empiricist philosophy of science, establishing a paradigm that privileges observational data and quantitative evidence in linguistic description.

The foundational principle of corpus linguistics, as articulated by Sinclair (1991), holds that language patterns revealed through large-scale corpus analysis often diverge from patterns predicted by introspective grammatical theory. Sinclair's distinction between the 'open-choice principle' and the 'idiom principle' illuminates how corpus evidence demonstrates that language use follows formulaic patterns more extensively than traditional grammar might suggest. This insight has profound implications for understanding not only how language operates but also how meaning is constructed through conventional phrasing and collocation.

Tognini-Bonelli (2001) distinguishes between corpus-based and corpus-driven approaches, a differentiation crucial for understanding methodological variation within the field. Corpus-based research employs corpora to test pre-existing hypotheses derived from linguistic theory, while corpus-driven research derives theoretical categories and patterns inductively from corpus data itself. This study adopts a primarily corpus-driven stance, allowing patterns to emerge from the data while remaining informed by existing knowledge of pantun structure and Malay poetics.

Corpus Methods in Literary Analysis

The application of corpus methods to literary texts has developed into a robust subdiscipline known as corpus stylistics. Semino and Short (2004) demonstrate how corpus techniques can reveal systematic patterns in narrative discourse, point of view, and speech presentation across literary texts. Their work illustrates how quantitative analysis of linguistic features can complement and deepen qualitative literary interpretation, rather than replacing traditional close reading approaches.

Research on poetic language using corpus methods remains less developed than prose analysis, though important contributions have emerged. Hoover, Culpeper, and O'Halloran (2014) edited a comprehensive volume exploring digital literary studies, including corpus approaches to poetry. Their work demonstrates how frequency analysis, keyword extraction, and collocation studies can illuminate stylistic choices, thematic preoccupations, and intertextual relationships in poetic texts. However, applications to non-Western poetic traditions remain scarce, creating opportunities for methodological innovation and cross-cultural comparative study.

Pantun as a Literary and Cultural Form

Scholarly investigation of pantun spans multiple disciplines including linguistics, literary studies, anthropology, and performance studies. Daillie (1988) provides a comprehensive examination of the pantun universe, documenting its structural variations, thematic categories, and social functions across the Malay world. His work emphasizes pantun's role as both artistic expression and social practice, noting how compositional skill in pantun signals cultural sophistication and rhetorical competence.

The structural architecture of pantun has received particular attention from scholars. The form typically consists of four lines with an a-b-a-b rhyme scheme, though variations exist. The first two lines establish imagery, often drawn from nature, daily life, or proverbial wisdom, while the final two lines deliver the actual message or meaning. This structure creates what scholars have termed 'semantic parallelism' between sampiran and maksud, where the imagery of the opening couplet metaphorically resonates with or illustrates the message conveyed in the closing couplet.

Noriah Mohamed (2006) analyzes the cognitive and emotional dimensions of traditional Malay poetry, including pantun, arguing that these forms encode distinctive ways of experiencing and expressing feeling. Her work suggests that understanding pantun requires not merely linguistic competence but cultural knowledge of the imagery, metaphors, and associative patterns conventional in Malay tradition. This cultural-cognitive dimension makes pantun an ideal subject for corpus analysis that attends to both linguistic pattern and cultural meaning.

Recent scholarship has begun addressing pantun preservation and pedagogical applications. However, systematic quantitative analysis of pantun language remains limited. This study responds to calls by scholars like Harun Mat Piah (2001) for more rigorous linguistic

documentation of pantun, employing computational methods to reveal patterns that can inform both linguistic theory and cultural preservation efforts.

METHOD

Research Design

This study employs a mixed-methods research design integrating quantitative corpus analysis with qualitative cultural-linguistic interpretation. The approach aligns with principles of corpus-driven linguistics as outlined by Tognini-Bonelli (2001), allowing patterns to emerge from data while remaining theoretically informed by existing scholarship on Malay poetics and cultural linguistics. The research follows an iterative analytical cycle of quantitative pattern identification followed by qualitative interpretation and validation through consultation of cultural-linguistic sources.

Corpus Construction

A specialized corpus of 50 traditional Malay pantun was compiled from authenticated public domain sources, including classical collections documented by Wilkinson and Winstedt (1961), pantun repositories maintained by national libraries, and scholarly anthologies of traditional Malay literature. Selection criteria prioritized: (1) authenticated traditional pantun predating modern compositions; (2) structural completeness with clear sampiran-maksud division; (3) thematic diversity representing major pantun categories (advice, love, nature, wisdom); and (4) public domain status ensuring ethical research practices.

The corpus totals approximately 800 words (200 lines of four-line pantun) representing a specialized mini-corpus suitable for detailed linguistic analysis. While modest in size compared to general-purpose corpora, this scale aligns with established practices in specialized corpus research. As McEnery and Wilson (2001) note, recent trends in corpus linguistics demonstrate increasing interest in small, highly specialized corpora that enable detailed investigation of particular genres or text types. Each pantun was manually transcribed and verified against multiple sources to ensure textual accuracy.

Analytical Tools and Procedures

Corpus analysis utilized AntConc 4.0, a freely available corpus analysis toolkit developed by Laurence Anthony. As Anthony (2005) documents, AntConc provides comprehensive functionality including concordancing, word frequency analysis, keyword extraction, collocation analysis, and distributional visualization. The software's accessibility and robust feature set make it particularly suitable for humanities research and specialized corpus investigation.

The analytical procedure followed five stages:

1. Frequency Analysis: Generation of word frequency lists to identify lexical items occurring with statistically significant frequency, enabling identification of dominant vocabulary and conceptual domains.
2. Concordance Analysis: Systematic examination of high-frequency content words in context to determine usage patterns, semantic ranges, and structural positions within pantun.
3. Collocation Analysis: Statistical identification of words that co-occur with significantly higher frequency than chance would predict, revealing conventional word pairs and phraseological units characteristic of pantun discourse.
4. Semantic Field Analysis: Categorization of high-frequency lexical items into semantic domains to map conceptual organization and thematic preoccupations in pantun.

5. Cultural Interpretation: Integration of quantitative findings with qualitative analysis drawing on scholarship in Malay cultural studies, anthropology, and traditional poetics to interpret identified patterns' cultural significance.

Validity and Reliability

Research validity was ensured through multiple mechanisms: (1) use of authenticated sources from established scholarly collections; (2) manual verification of textual accuracy through comparison of multiple versions; (3) triangulation between quantitative corpus findings and qualitative cultural-literary scholarship; and (4) peer debriefing with specialists in Malay literature. Reliability was enhanced through systematic documentation of analytical procedures, enabling replication, and through use of established corpus tools with validated statistical measures. The iterative analytical process, cycling between quantitative analysis and qualitative interpretation, strengthens both the rigor and cultural validity of findings

RESULTS AND DISCUSSION

Lexical Frequency Patterns

Frequency analysis of the pantun corpus, excluding grammatical function words, reveals distinct lexical priorities reflecting the cultural-conceptual world of traditional Malay society. The highest-frequency content words include natural phenomena and objects: 'padi' (rice/paddy, 15 occurrences), 'hati' (heart/liver, 14 occurrences), 'bunga' (flower, 12 occurrences), 'air' (water, 11 occurrences), 'burung' (bird, 10 occurrences), and 'batu' (stone/rock, 9 occurrences). Human relationships and social roles also feature prominently: 'orang' (person/people, 13 occurrences), 'anak' (child, 8 occurrences), and 'adik' (younger sibling, 7 occurrences).

This frequency distribution demonstrates pantun's grounding in agrarian life and natural world observation. The prominence of 'padi' reflects rice cultivation's centrality to traditional Malay economy and cultural identity. As Winstedt (1969) documents, rice agriculture structures not only economic life but also ritual calendar, social organization, and cosmological understanding in Malay communities. Similarly, frequent reference to natural elements like water, flowers, and birds reflects intimate environmental knowledge and the use of nature imagery as a conventional poetic resource.

The high frequency of 'hati' (heart/liver) warrants particular attention. In Malay cultural-linguistic conceptualization, 'hati' functions as the seat of emotion, cognition, and moral character. The term appears in numerous compounds and idioms denoting emotional and ethical states. Corpus analysis reveals 'hati' occurring in diverse contexts across both sampiran and maksud sections, functioning sometimes as concrete imagery and sometimes as the locus of the pantun's message. This polysemy and contextual flexibility make 'hati' a key lexical item bridging the sampiran-maksud structural division.

Concordance Patterns and Contextual Usage

Concordance analysis illuminates how high-frequency lexical items function within pantun's distinctive two-part structure. Examination of 'bunga' (flower) demonstrates characteristic patterns. In sampiran contexts, 'bunga' typically appears with modifying terms specifying flower types or attributes: 'bunga melur' (jasmine flower), 'bunga mawar' (rose), 'bunga kembang' (blooming flower). These concrete specifications create vivid natural imagery establishing the sampiran's scene-setting function.

Conversely, in maksud contexts, flower imagery often serves metaphorical or symbolic functions. Phrases like 'bagai bunga' (like a flower) employ floral imagery to characterize

human attributes, particularly feminine beauty or transient youth. This pattern exemplifies the semantic relationship between sampiran and maksud: concrete natural observation in the opening couplet transforms into metaphorical commentary on human experience in the closing couplet.

Concordance analysis of 'hati' reveals its central role in pantun's ethical and emotional discourse. The term appears in diverse phrasal combinations: 'hati kasih' (loving heart), 'hati sakit' (hurt feelings), 'hati tersangkut' (attached heart/caught affections), 'hati budi' (character/virtue). These collocations demonstrate 'hati' functioning as the grammatical and semantic anchor for expressing emotional and moral states. The frequency and diversity of 'hati' constructions suggests that pantun serves importantly as a vehicle for emotional expression and moral reflection in Malay culture.

Collocation Analysis and Conventional Phrasing

Collocation analysis identifies word pairs and phrasal units that occur together with statistically significant frequency, revealing conventional expressions characteristic of pantun discourse. Significant collocations include 'pisang emas' (golden banana, appearing in the famous pantun about reciprocity), 'hutang budi' (debt of gratitude), 'dari mata' (from the eyes, in the phrase 'dari mata turun ke hati' meaning love at first sight), and 'turun ke hati' (descend to the heart).

These collocations demonstrate pantun's reliance on formulaic language and conventional phrasing. As Sinclair's (1991) idiom principle suggests, language use involves not only open lexical choice but also selection from conventionalized multi-word units. Pantun exemplifies this principle dramatically, with formulaic expressions serving both mnemonic and aesthetic functions. The conventional nature of certain phrasings enables improvisational composition and aids memorization, crucial for an oral poetic tradition.

Particularly significant is the collocation 'hutang budi' (debt of gratitude), which encapsulates a core value in Malay ethics. The phrase appears in one of the most famous pantun: 'Pisang emas dibawa belayar / Masak sebiji di atas peti / Hutang emas boleh dibayar / Hutang budi dibawa mati' (Golden bananas brought sailing / One ripens on the chest / Gold debt can be repaid / Gratitude debt carried to death). This pantun articulates the principle that moral obligations transcend material debts, a fundamental tenet of Malay social ethics. The collocation's frequency and cultural resonance demonstrate how pantun encodes and transmits cultural values through memorable phrasing.

Semantic Field Analysis

Categorization of high-frequency lexical items into semantic domains reveals the conceptual organization of pantun discourse. Five major semantic fields emerge:

1. Natural World. Including flora (padi, bunga, pisang, kayu), fauna (burung, ikan, ayam), and landscape features (air, batu, gunung, laut). This domain dominates sampiran sections, providing concrete imagery drawn from environmental observation.
2. Kinship and Social Relations. Terms for family relationships (anak, adik, kakak) and general social categories (orang, tuan). These terms frequently appear in maksud sections, grounding ethical teachings in specific relational contexts.
3. Emotion and Affect. Lexemes denoting feelings and emotional states (kasih, sayang, rindu, sakit), with 'hati' as the central organizing term. This semantic field pervades both sampiran and maksud, though emotional vocabulary concentrates more heavily in maksud.

4. Ethics and Values. Terms encoding moral concepts (budi, hutang, janji, adat). This domain appears almost exclusively in maksud sections, reflecting pantun's function in transmitting ethical principles.
5. Actions and States. Verbs and verbal expressions describing activities (pergi, pulang, menangis, tertawa) and states of being (jauh, dekat, tinggi, rendah).

The distribution of semantic fields across pantun's structural components illuminates its aesthetic and communicative strategy. Natural world vocabulary concentrates in sampiran sections, establishing concrete, sensory imagery that grounds the verse in shared environmental experience. Conversely, ethical and emotional vocabulary dominates maksud sections, where the pantun's message or teaching emerges. This pattern reflects the metaphorical relationship between sampiran and maksud: natural observation serves as vehicle for understanding human experience and moral truth.

Cultural Values and Worldview

Integration of corpus findings with cultural-anthropological scholarship reveals how linguistic patterns in pantun encode distinctive Malay cultural values and worldview. The prominence of 'budi' and related ethical vocabulary reflects the centrality of budi in Malay cultural philosophy. As scholars of Malay culture explain, budi encompasses refined character, moral excellence, cultivation of social grace, and reciprocal ethical obligation. The frequent appearance of budi-related lexemes in pantun demonstrates the form's role in transmitting and reinforcing this core cultural value.

Similarly, the prevalence of reciprocity vocabulary (balas, bayar, hutang) reflects the importance of reciprocal obligation in Malay social ethics. Traditional Malay society operates through complex networks of mutual obligation, where actions create debts that must be acknowledged and reciprocated. Pantun articulates these principles memorably, making ethical abstractions concrete through vivid imagery and rhythmic language.

The dominance of nature imagery reveals an intimate relationship between Malay communities and their natural environment. This relationship transcends mere economic dependence on natural resources, reflecting a worldview that sees human experience as fundamentally continuous with natural processes. The pantun convention of using natural imagery to illuminate human affairs suggests a cosmology that does not sharply separate human from natural domains, but rather perceives analogical relationships and moral lessons in natural phenomena.

The emotional vocabulary centered on 'hati' demonstrates a distinctive cultural conceptualization of personhood and affect. In Malay cultural psychology, emotions and thoughts are not clearly separated, both being seated in the 'hati'. This holistic conception of cognition-emotion contrasts with Western psychological models that distinguish thinking and feeling more sharply. The linguistic evidence from pantun, with its rich 'hati' phraseology, provides empirical support for anthropological accounts of Malay cultural psychology and demonstrates how corpus methods can illuminate cultural conceptualizations of self and emotion.

CONCLUSION

This corpus-based investigation of traditional Malay pantun demonstrates the efficacy of computational linguistic methods for analyzing traditional oral poetry and reveals systematic linguistic patterns that encode cultural values and worldview. Frequency analysis identified

lexical priorities reflecting agrarian life, natural world observation, and ethical-emotional concerns. Concordance analysis illuminated how high-frequency lexemes function differently in sampiran versus maksud contexts, supporting the aesthetic principle of metaphorical relationship between these structural components. Collocation analysis revealed conventional phraseology that facilitates oral composition and memorization while expressing core cultural values. Semantic field analysis demonstrated systematic relationships among natural imagery, social relationships, emotional expression, and ethical teaching.

These linguistic patterns reflect and reinforce distinctive aspects of Malay cultural worldview, particularly the centrality of budi (refined character and reciprocal obligation), the intimate relationship between human communities and natural environment, and a holistic conceptualization of cognition and emotion centered on 'hati'. The findings demonstrate that pantun functions not merely as aesthetic artifact but as a crucial vehicle for transmitting cultural values, social norms, and ethical principles across generations.

This research makes several contributions to scholarship. Methodologically, it extends corpus linguistic approaches to non-Western poetic traditions, demonstrating adaptations necessary for analyzing oral poetry with distinctive structural features. Theoretically, it provides quantitative evidence supporting anthropological accounts of Malay cultural values and worldview, showing how corpus methods can complement ethnographic research. Practically, it contributes to digital humanities efforts to preserve intangible cultural heritage through computational documentation and analysis.

Several limitations warrant acknowledgment. The corpus size, while appropriate for specialized analysis, limits generalizability of findings. The study focuses exclusively on traditional pantun, not examining modern compositions or regional variations. Analysis was conducted on written transcriptions, potentially missing prosodic and performance features crucial to oral poetry. Additionally, the study addresses only Malay-language pantun, not examining pantun traditions in other languages of the Malay world.

Future research should: (1) expand the corpus to include larger samples and regional variations; (2) conduct comparative analysis across traditional and modern pantun to identify diachronic changes; (3) investigate relationships between linguistic patterns and performance contexts; (4) extend analysis to pantun in other languages (Javanese, Sundanese, Minangkabau); (5) develop specialized corpus annotation schemes capturing pantun's distinctive structural features; and (6) explore applications of natural language processing and machine learning to automated pantun analysis and generation. Such research would further demonstrate corpus methods' potential for preserving and understanding traditional oral literature while contributing to theoretical linguistics and digital humanities.

REFERENCES

Anthony, L. (2005). AntConc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom. *Proceedings of the 2005 IEEE International Professional Communication Conference*, 15-16.

Biber, D., & Conrad, S. (2019). *Register, genre, and style*. Cambridge: Cambridge University Press.

Daillie, F.-R. (1988). *Alam pantun Melayu: Studies on the Malay pantun*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Harun Mat Piah. (2001). *Pantun Melayu: Bingkisan permata*. Kuala Lumpur: Yayasan Karyawan.

Hoover, D. L., Culpeper, J., & O'Halloran, K. (2014). *Digital literary studies: Corpus approaches to poetry, prose, and drama*. New York: Routledge.

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

McEnery, T., & Wilson, A. (2001). *Corpus linguistics: An introduction* (2nd ed.). Edinburgh: Edinburgh University Press.

Noriah Mohamed. (2006). *Sentuhan rasa dan fikir dalam puisi Melayu tradisional*. Bangi: Penerbit Universiti Kebangsaan Malaysia.

Semino, E., & Short, M. (2004). *Corpus stylistics: Speech, writing and thought presentation in a corpus of English writing*. London: Routledge.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work.* Amsterdam: John Benjamins.

Wilkinson, R. J. (1907). *A Malay-English dictionary*. Singapore: Kelly & Walsh.

Wilkinson, R. J., & Winstedt, R. O. (1961). *Pantun Melayu* (4th ed.). Singapore: Malaya Publishing House Limited.

Winstedt, R. O. (1969). *A history of classical Malay literature*. Singapore: Oxford University Press.