information

MDPI

Article

Unravelling Lexical and Narrative Patterns in the *Hikayat Lonthoir*: A Computational Linguistics Approach

Muhamad Iko Kersapati [1,*], Francesco Perono Cacciafoco [2,*], Bimasyah Sihite [1], Shiyue Wu [3], Khofiyana Putri Widyaningrum [4], Mohamad Atqa [1] and Elvis A. B. Toni [5]

[1] The Ministry of Culture, Jakarta 10270, Indonesia
[2] Department of Applied Linguistics, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China
[3] Department of Linguistics, University of Hawai'i at Mānoa, Honolulu, HI 96822, USA
[4] Department of Geography, Universitas Indonesia, Depok 16424, Indonesia
[5] Department of English Education, Widya Mandira Catholic University, Kupang 85225, Indonesia
* Correspondence: muhamad.kersapati.21@alumni.ucl.ac.uk (M.I.K.); francesco.perono@xjtlu.edu.cn (F.P.C.)

Abstract

Hikayat Lonthoir, a rare saga manuscript collection originating from the Banda Archipelago, Maluku, Indonesia, retains significant Indigenous oral history amidst the Western colonial narrative. This study seeks to leverage computational methods to analyze the historic manuscript that constitutes a combination of OCR-supervised transcription, corpus linguistic profiling, semantic clustering (Word2Vec + K-Means), and named entity network analysis. A validation of the dataset is performed on 2793 cleaned word tokens towards Indonesian and Malay dictionaries, showing that 50.3% overlapped with both dictionaries, with strong cross-dictionary agreement ($\kappa = 0.76$). The lexical analysis indicates that monarchy/governance, kinship, maritime vocabulary, and extensive morphological productivity (*me-*, *di-*, *ter-*, *pe-/per-*, *-nya*, *-an*), while semantic and network analyses identify two narrative cores, developed into Aarne–Thompson–Uther (ATU) and Stith Thompson's Motif Index of Folk Literature classification systems. These findings demonstrate how computational methods can extract structural, thematic, and relational patterns from historical manuscripts and contribute evidence-based insights to digital philology and historical linguistics.

Keywords: Banda Archipelago; digital philology; linguistics documentation; NLP; oral history; semantic analysis

check for updates

Academic Editor: Shmuel Tomi Klein

Received: 25 October 2025
Revised: 21 November 2025
Accepted: 2 December 2025
Published: 4 December 2025

Citation: Kersapati, M.I.; Perono Cacciafoco, F.; Sihite, B.; Wu, S.; Widyaningrum, K.P.; Atqa, M.; Toni, E.A.B. Unravelling Lexical and Narrative Patterns in the *Hikayat Lonthoir*: A Computational Linguistics Approach. *Information* **2025**, *16*, 1069. https://doi.org/10.3390/info16121069

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Debates regarding qualitative and quantitative research approaches have become increasingly prominent, particularly with the growing use of digital technology and data science in historical linguistics. A long-standing view among scholars emphasizes empirical studies, which traditionally rely on historical documents as their primary sources [1,2]. In fact, historical linguistics has been data-centric since its inception. Nevertheless, the adoption of quantitative methods in historical linguistics remains far from mainstream and still lags considerably behind the levels achieved in other branches of linguistics. However, the underuse of quantitative methods in historical linguistics opens a wide opportunity for a wide range of recent innovations. Instead of debunking theoretical hypotheses, it complements the qualitative investigation, particularly for the medium and large-scale datasets, allowing approaches by experts from multiple lenses [3].

Natural language processing (NLP) is central to this operation, developed as an immense motivation for automatic processing of historical texts, particularly in the emerging field of digital humanities [4]. Despite limited references, several previous works showcase notable examples of how the leverage of this automated approach becomes a concern for information technologies and data science experts, historic linguists, and archivists to advance and deepen the body of knowledge in historical linguistics, its documentation, and archival. Conathan (2011) examined an effective digital archival management of endangered language documentation as the fundamental functions in digital processing, such as appraisal, accession, arrangement, description, preservation, access, and use. He focuses on the collection repositories and the standardization of language documentation, particularly metadata, citation, and access [5].

Pessanha and Salah (2022) highlighted that the transformation of the field of archival science by computational technologies, such as automatic speech recognition and natural language processing, creates unique possibilities for analysis of oral history by prompting new approaches to process the extensive data in certain collections, which typically will be too time-consuming by the empirical approach [6]. These notions were reinforced by Nepal and Perono Cacciafoco (2024), who recently investigated the decipherment of the Bronze Age script, Linear A, used by the Minoan civilization, employing a combined computational approach, such as a feature-based similarity measure and consonantal approach. The results revealed a monumental finding that identified some possible word matches when comparing Linear A with each of the tested language families (Ancient Egyptian, Luwian, Hittite, Proto-Celtic, and Uralic) [7].

Building on this background, this paper explores the linguistic dimension, expanding the discourse on computational methods in examining historical linguistics and archival materials with a specific case of *Hikayat Lonthoir*, a saga from the Banda Archipelago, Maluku, written by M.S. Neirabatij (an influential figure called *Orang Kaya*) in 1922, considering its heritage significance of the Bandanese narrative (Figure 1). This manuscript also reflects rare and underrepresented Indigenous oral history, which has been silenced amidst the dominant colonial narrative [8]. Recently, this manuscript has been digitized and is available publicly by the National Maritime Museum in Amsterdam [9]. This marks a new stage of exploration into traditional knowledge.

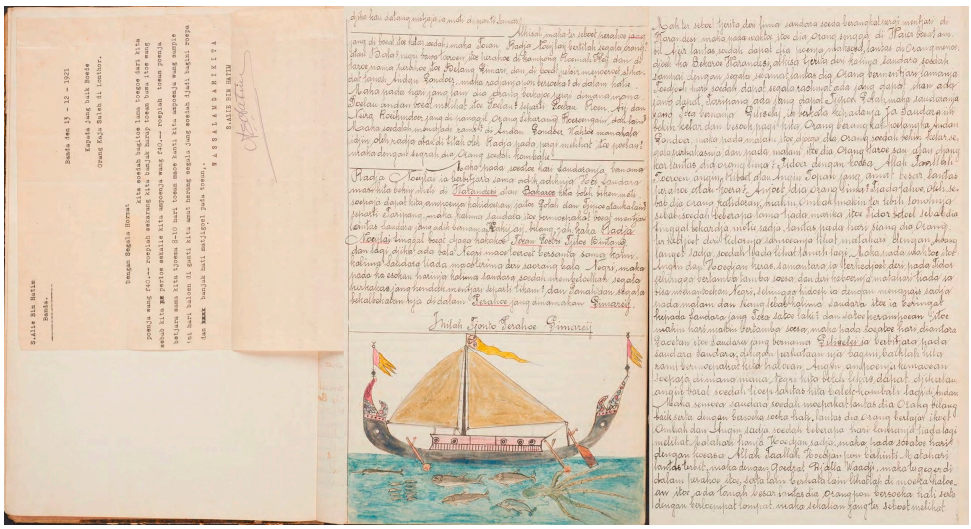


Figure 1. *Hikayat Lonthoir*, written by M.S. Neirabatij in 1922, curated from the National Maritime Museum collections, Amsterdam: The first page (**left**); *Belang* figure (Bandanese boat) (**middle**); and the narrative handwritten in cursive Malay (**right**).

Religious spirit significantly influences the stories in the saga, particularly Islamic taught sourced in the Al Qur'an in connection to the Banda Islands as the central narrative. The deep connection between the Bandanese people and God is evident from the opening pages of the *Hikayat Lonthoir*, recounting the story of the first inhabitants of the Banda Islands, beginning with the moment the sea receded after the great flood and the miracles of the prophet Noah (*Nuh*). Several other figures, such as Siti Gelsoen, Raja Noeilaj, and Neirabatij, appear as the main characters in multiple sections of the stories [10].

Exploration of this cultural narrative is essential, despite the massive technological development, particularly to document historical lexicons and terminologies [11]. The manuscript presents substantial challenges for linguistic and computational analysis due to its inconsistent elements, such as historical spelling, visual noise (for OCR operation), and hybrid orthographic features. These issues hinder reliable tokenization, dictionary matching, and motif identification. To date, no standardized procedures or reproducible workflows exist for processing Malay manuscripts exhibiting such pre-modern orthographic variation. This study addresses the problem by developing a transparent normalization framework, quantifying OCR and annotation reliability, and applying NLP-based modeling to uncover lexical patterns, narrative structures, and cross-cultural motifs within the text.

Hence, the primary questions seek to answer, such as how computational methods are used to deconstruct and analyze the characteristic linguistic features (e.g., lexicon, morphology, etc.) and the relation to entity and semantic groups found within the saga? How does this investigation explain and generate new insights on multiple scales into the narrative structure of the saga?

2. Materials and Methods

2.1. Workflow and Dataset

The main procedures are shown in Figure 2, involving operations such as transcription, vocabulary curation, data cleaning, validation, and analysis. Micro-context covers lexical analysis (*corpus linguistics*), constituting analyses around affixes, old spellings, and loanwords to uncover the vocabulary, morphology, and orthography. Meso-context includes semantic fields analysis using word embeddings, K-means clustering, and degree centrality to determine entities and their relationships in the narrative. In the macro-context, a review of ATU and Thompson folklore index is conducted to compare the narrative of the saga to global folklore systems (e.g., motifs, characters, and storylines).

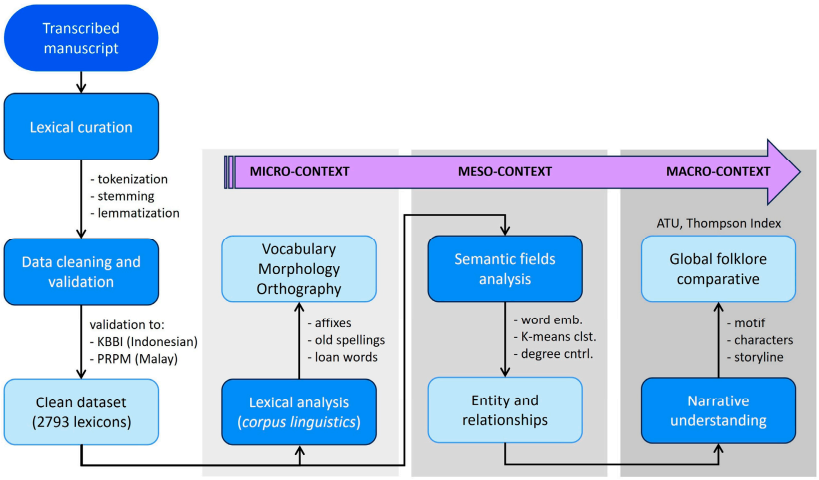


Figure 2. The research workflow constitutes three main contexts of investigation.

A transcription was carried out by automatic extraction using optical character recognition (OCR), then supervised and interpreted by a human user. To ensure quantifiable and reproducible transcription quality, OCR accuracy was evaluated using a 10% stratified sample of manuscript pages. Accuracy was computed using Levenshtein similarity at both the character and word levels. The OCR achieved a character-level accuracy of 86.1% and a word-level accuracy of 70.4%, reflecting the challenges posed by the historical *Van Ophuijsen* spelling system and hybrid handwritten–print features in the manuscript. The OCR output was subsequently corrected manually.

Historical spelling normalization was applied before tokenization to standardize the manuscript's pre-modern Malay orthography using *Van Ophuijsen's* systematic substitutions such as oe → u (*goeroe* → *guru*), tj → c (*tjoetji* → *cuci*), dj → j (*djalan* → *jalan*), and sj → sy (*sjarat* → *syarat*) [12,13]. After passing this operation, lexical data preparation was performed in two phases such as tokenization (breaking down the words), stemming (removing affixes and stop words), and lemmatization (deriving and learning from dictionary definitions) [14]. The primary pass addressed spelling errors, token boundary mistakes, and the removal of OCR artefacts. The secondary verification resolved residual inconsistencies and harmonized historical spelling variants. The raw OCR output consisted of 3410 tokens, of which 1148 tokens (33.7%) required manual correction. At the character level, 5143 characters (15.1%) required adjustment. A detailed correction log was maintained documenting substitution, deletion, insertion, split-word, and merged-word errors.

Two annotators independently corrected 20% of the corpus, yielding a token-level agreement of 92%, boundary agreement of 95%, and normalization agreement of 84% (Cohen's $\kappa = 0.92$, 0.95, and 0.84, respectively). All discrepancies were adjudicated to produce the final gold-standard transcription used for downstream analysis. Data cleaning and checking of related dictionaries were then carried out to validate the extracted lexicons in Indonesian and Malay, resulting in a list of 2793 words, as a ready-to-use dataset.

2.2. Data Cleaning and Validation

After transcription, the curated lexicons were then cleaned, towards a list of stop words, such as *yang* (which), *dan* (and), *di* (at/in/on), *ke* (to), *dari* (from), *ini* (this/these), *itu* (that/those), etc. These common words do not have much important information and need to be removed in the text preprocessing process for meaning retrieval. The extracted words were then validated against the Indonesian dictionary, *Kamus Besar Bahasa Indonesia* (KBBI), and the Malay dictionary, *Pusat Rujukan Persuratan Melayu* (PRPM), as shown in Figure 3.

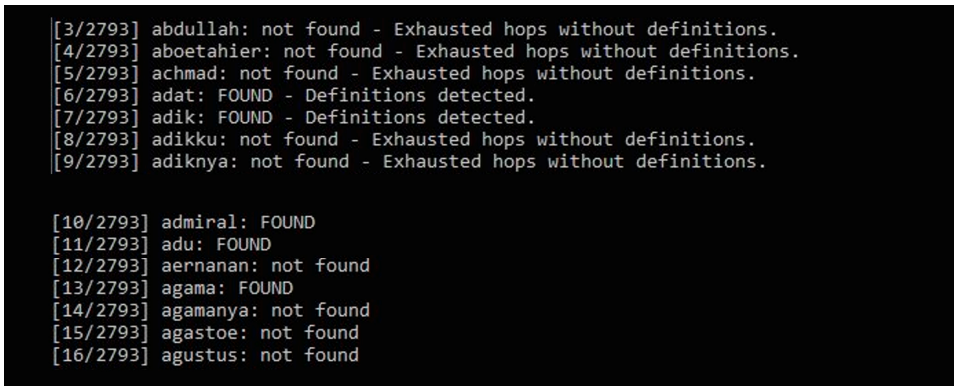


Figure 3. Word validation snippet towards the Indonesian dictionary (KBBI) https://kbbi.web.id (accessed on 13 September 2025) (**top**) and Malay (PRPM) https://prpm.dbp.gov.my (accessed on 13 September 2025) (**bottom**).

For the KBBI API, a customized prompt is required to scrape the words, whereas the PRPM has no official API but can be responsibly scraped for research purposes, provided rate limiting is observed. A hybrid validation pipeline provides the most practical workflow for this dataset by matching the words against modern Malay and Indonesian lexicons and flagging any "unknown" terms. These unknown words are then manually reviewed, with archaic or loan terms (e.g., "*Lonthoir*," "*Majapahit*," "*Andan*") added to the project dictionary, and finally, the analysis is re-run using the updated dictionary to ensure comprehensive coverage. Statistical measure was then set, constituting:

1. Observed Agreement (P_o) measures the actual agreement between two dictionaries. It specifies how often words were found in both KBBI and PRPM compared to all possible words.
2. Expected Agreement (P_e) quantifies the chance of agreement, or expresses how much overlap we would expect purely by random distribution, given the proportions of words found in each dictionary.
3. Cohen's Kappa (κ) adjusts the Observed Agreement by accounting for chance agreement (i.e., random overlap), giving a more accurate measure of how much the dictionaries actually agree beyond random chance.
4. Jaccard Index (for Found Sets) measures the similarity between two sets by calculating the ratio of their intersection over the union.
5. Phi Correlation (Φ) coefficient is a measure of association between two binary variables (e.g., whether a word is found in KBBI and PRPM).

To assess the robustness of dictionary agreement, confidence intervals were calculated using 1000 bootstrap iterations for the agreement metrics. To anticipate a large portion of the lexicon that appeared in neither dictionary, a random sample of 150 items was examined to classify their sources. This set consisted of historical spellings, affixed/derived forms, regional Banda–Maluku vocabulary, Arabic/Dutch loanwords with archaic orthography, proper names, and a small amount of OCR noise. This sampling confirmed that most unmatched items represent meaningful linguistic material rather than transcription errors.

2.3. Lexical Analysis (Corpus Linguistics)

This analysis primarily covers vocabulary, morphology, and orthography. Vocabulary and frequency identification are crucial not only as language documentation but also help to map the themes that are developed into semantic clustering. In terms of morphology, affixes were identified and decomposed from their tokens, including prefixes and suffixes (Table 1). Alongside this, foreign borrowings, particularly Dutch, Arabic, and local vernaculars, were annotated to trace their integration into the Indonesian lexicon. These terms were identified, classified, and analyzed for morphological adaptation, supporting semantic mapping and revealing cross-linguistic influences within the historical context of the texts [15]. The corpus was examined for loan words or the use of foreign terms from online repositories [16].

Morphological analysis relied on the *Sastrawi* stemmer, which is designed for modern Indonesian and therefore produces errors on historical or regionally inflected forms. To evaluate this, a small error analysis was performed on a stratified sample of 200 tokens. Mis-stemming occurred mainly in historical spellings (*bermatoe* → *ato*, *tjeritera* → *cerita*) and in productive affixation involving old forms (*diperbuatnja*, *terlaloeh*), with an estimated error rate of 12–15%. These errors were flagged but retained in context, as the aim of morphological profiling was descriptive rather than fully reconstructive. Normalized forms reduced some errors (e.g., *boenga* → *bunga*), but original forms were preserved for names and quotations.

Table 1. Affixes in the Indonesian language system primarily consist of prefixes and suffixes [15].

Affix Type	Affix Forms	Function & Examples
Prefixes (*Awalan*)	*ber-*, *di-*, *ke-*, *ter-*, *per-*, *se-*	Affixed to the beginning of the root word, *di-* forms the Passive voice (*dibaca* = is read); *ter-* often indicates the superlative (*terbesar* = biggest) or an unintentional action (*terjatuh* = accidentally fell).
	"*me-*" group: *me-*, *mem-*, *men-*, *meng-*, *menge-*, *meny-*	Variations in the Active voice prefix. These forms depend on the initial letter of the root word and are the most common way to form Active verbs (*menulis* = to write).
	"*pe-*" group: *pe-*, *pem-*, *pen-*, *peng-*, *penge-*, *peny-*	Variations used to form Nouns indicating the actor/agent (*penulis* = writer), the tool (*penggaris* = ruler), or the result.
Suffixes (*Akhiran*)	*-an*, *-kan*, *-i*, *-lah*, *-kah*, *-nya*	Affixed to the end of the root word. *-an* often creates a noun (*makanan* = food); *-kan* and *-i* often create causative or applicative verbs (*membelikan* = to buy for someone).

The italics indicate Indonesian terminologies used in affixes.

To identify the most frequent words, the normalized corpus was tokenized into word units, lowercased, and stripped of punctuation. Tokens belonging to a predefined Indonesian stopword list, numerals, and OCR noise fragments were excluded to avoid inflating frequencies with grammatical or non-lexical forms. A frequency distribution was computed over the remaining content words, and the results were ranked from highest to lowest.

2.4. Semantic Clustering

Word embeddings and K-Means clustering were employed to categorize the words in the dataset into predefined semantic themes. The word embeddings task converts words into numerical vectors that capture their meanings based on context [17]. Word2Vec was used using randomized word embeddings, which were generated for each word in the dataset. These embeddings represent the semantic properties of words in a multi-dimensional space, where words were clustered have similar meanings are close to each other using a neural network-based approach [18].

Because the corpus is relatively small, training Word2Vec from scratch would produce unstable vectors. To address this, the semantic analysis used pretrained Indonesian–Malay embeddings (FastText) as the base space, and vectors for the saga manuscript terms were obtained through subword composition. Clustering was performed using K-Means, and the number of clusters was selected using intrinsic criteria (silhouette score and Davies–Bouldin index). Stability was checked across 20 random seeds, with highly unstable terms excluded from interpretation. For transparency, each cluster is presented together with representative high-loading terms, and coherence was checked manually to ensure that cluster labels correspond to recognizable semantic themes (e.g., genealogy, religion, maritime lexicon).

The number of clusters (K) was set to 4, defining semantic themes: Maritime, Religious, Social/Genealogical, and Place/Geographic. After the clustering process, each word was assigned a cluster label. The clusters were then reviewed manually, and based on the words present in each cluster, a semantic group label was assigned. For instance, words related to family or kinship, e.g., *adik* (sibling), *ayah* (father), were assigned to the Social/Genealogical group. Words associated with religion, e.g., *Allah* (God), *salat* (prayer), were assigned to the Religious group. Words related to places, e.g., *laut* (sea), *pantai* (coast), were grouped under Place/Geographic.

The semantic clustering results should be interpreted with caution due to several validation limitations. Although silhouette scores and Davies–Bouldin indices were used to

select K = 4, the metric values were low, reflecting weak internal separation consistent with the small corpus size. Cluster labels were assigned post hoc and represent interpretive categories rather than independently verified semantic classes. The corpus contains only 2793 tokens, limiting the stability of Word2Vec representations even when initialized with pretrained FastText embeddings [19]. Sparse data increases sensitivity to hyperparameters and random seeds, and cluster robustness was not systematically evaluated across multiple runs. Consequently, these semantic groupings should be viewed as exploratory patterns rather than definitive semantic fields.

2.5. Named Entity Recognition and Network Analysis

A semi-automatic approach combining manual labelling and lexical pattern matching was used for named entity recognition (NER) within the narrative [20,21]. NER combined automatic extraction with a manually curated gazetteer that included person names, titles, toponyms, and clan names derived from the manuscript and regional sources. Matching used case-insensitive exact matching and fuzzy matching (Levenshtein distance ≤ 1) for orthographic variants. Co-occurrence networks were built using a $\pm$5-token window, and edge weights were proportional to co-occurrence frequency. Network measures included degree, weighted degree, betweenness, and eigenvector centrality.

To assess extraction quality, a manually annotated sample (n = 500 tokens, $\approx$18% of the corpus) was used to estimate NER precision and recall (see Table 2). Centrality estimates were further evaluated using bootstrap resampling (1000 iterations) to compute uncertainty ranges, ensuring that high-centrality entities represent stable patterns rather than sampling fluctuations. Entity extraction constitutes key proper nouns and thematic terms extracted from the corpus, focusing on characters, places, events, and objects central to the plot (e.g., *Raja Noeilaj*, *Banda*, *Belang Limareij*, *Islamisasi*).

Table 2. NER Precision, Recall, and Bootstrap Uncertainty (1000 iterations).

Metric	Mean	95% CI (Lower–Upper)
Precision	0.91	0.88–0.94
Recall	0.86	0.83–0.89
F1-Score	0.88	0.85–0.90
Centrality Stability (Top-10 Nodes)	0.30	0.26–0.34

The NER procedure has several limitations. The gazetteer (214 entries) was compiled from capitalized tokens, Banda historical sources, and manual inspection, but its construction remains heuristic. The Levenshtein threshold ($\leq$1) was chosen through pilot testing rather than formal optimization. Our precision–recall check covers only 18% of the corpus and does not include error analysis by entity type, so some categories may be more affected than others. In addition, Sastrawi's 12–15% stemming error rate likely propagates into downstream frequency counts, affix patterns, dictionary matching, and semantic clustering, although we have not quantified these effects due to the absence of a full gold-standard annotation. Future work will address these issues through comprehensive annotation, entity-specific evaluation, and more suitable historical Malay NLP tools.

Semantic grouping and quantification were then performed to organize the entities into a hierarchical structure for thematic analysis, and were measured by frequency, divided into Hierarchical grouping (Level 1). The specific semantic types (Level 2) were aggregated into four primary thematic domains (Level 1) that represent the core narrative components of the saga and historical chronicle:

1. Person (LIVB): Key individuals and characters.
2. Place (GEOG): Geographic locations and landmarks.

3. Object (DEVI/TISS): Physical items/artefacts.
4. Concept (CONC): Abstract ideas.

The quantification for the frequency of each Level 2 Semantic Type was estimated using data derived from the "Sum" row of a broader co-occurrence matrix analysis (e.g., LIVB, CONC) to quantify the prominence of each category within the overall corpus activity.

2.6. Ethics and Community Considerations

Although this study did not involve human participants, the *Hikayat Lonthoir* manuscript constitutes the cultural heritage of the Banda community. All transcription, normalization, and analytical procedures were conducted with respect for local cultural values and responsible handling of heritage materials. Only publicly accessible content was analyzed, and no sensitive or restricted cultural knowledge was disclosed.

3. Results and Discussion

3.1. Dictionary Entries: Statistical Exploratory

Word validation results toward KBBI (Indonesian) and PRPM (Malay), as shown in Table 3, detected the presence of the observed lexicons 50.3% (1405 words) in both dictionaries. While 4.8% (134 words) were found only in KBBI and 7.09% (198 words) appeared only in PRPM, showing a slightly higher number of words unique to the Malay lexicon compared to Indonesian. On the other hand, 37.81% (1056 words) were found in neither dictionary. This is a substantial portion, possibly words with affixes (both prefixes and suffixes) such as *di-*, *ke-*, *-nya*, etc., as well as archaic, regional, borrowed, or non-standard words that were not documented in either source.

Table 3. Contingency matrix for validation results towards both dictionaries.

	PRPM (Found)	**PRPM (Not Found)**
KBBI (Found)	1405	134
KBBI (Not Found)	198	1056

Figure 4 compares word-length distributions across four lexical categories: KBBI-only, PRPM-only, overlapped (shared), and neither. This measures how long the words are (in characters) within each group and helps reveal whether shared words tend to be short and simple (core vocabulary), whether unique dictionary entries are longer and more specialized, or whether non-dictionary words follow irregular length patterns. Statistically, each violin displays the kernel density estimate, showing where word lengths are most concentrated: wider sections represent frequent lengths, while narrow sections indicate rarity. The jittered points reveal the underlying data, confirming that no category contains extreme outliers. The mean marker, accompanied by its 95% confidence interval, summarizes the central tendency and the precision of the average length estimate for each category. Although the mean lengths are relatively similar, the violins differ in variance, with some groups exhibiting broader distributions (more heterogeneous word lengths) and others appearing more compact (more uniform length structure).

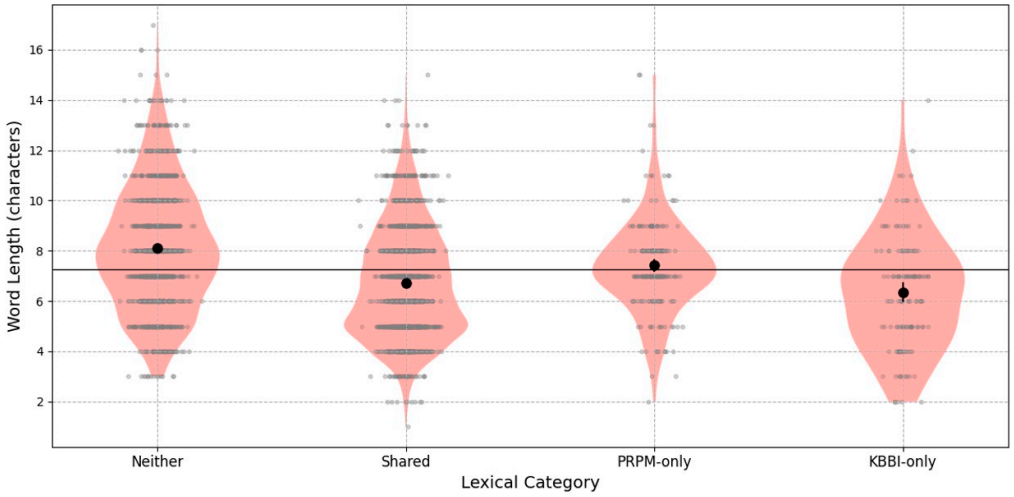


Figure 4. Word length distribution. Grey dots represent individual words, black dots indicate category means, and the horizontal black line marks the overall mean word length.

Agreement and similarity measures (Table 4) suggest that KBBI and PRPM were highly consistent and aligned well in terms of word inclusion. This measure also confirms the reliability of the data representation [22]. This was indicated by the observed agreement (88.11%) and Cohen's Kappa (0.7586). Meanwhile, the phi correlation (0.7594) and Jaccard index (0.8089) further confirm that there was a strong association and overlap between the two dictionaries. Conditional Likelihoods show that the conditional probabilities show that words found in one dictionary (especially KBBI) were very likely to also appear in the other dictionary (PRPM), with high probability when a word was present in KBBI or PRPM. On the other hand, differences in coverage reveal words missing from KBBI or PRPM have a low chance of appearing in the other dictionary, suggesting that there may be unique sets of words in each dictionary, but the overlap is strong.

Table 4. Agreement and similarity metrics constitute Jaccard (0–1): overlap of "found" sets (1 = identical); Cohen's κ agreement beyond chance: ≈0 means weak, >0.4 moderate, >0.6 substantial; Phi correlation between "found" in KBBI and PRPM; and Conditionals (e.g., P(PRPM | KBBI = 1)) probability that a KBBI-found word is also in PRPM and vice versa.

Metric	Value
observed_agreement	0.8811
expected_agreement	0.5075
cohens_kappa	0.7586
jaccard_found_sets	0.8089
phi_correlation	0.7594
P(PRPM | KBBI = 1)	0.9129
P(PRPM | KBBI = 0)	0.1579
P(KBBI | PRPM = 1)	0.8765
P(KBBI | PRPM = 0)	0.1126

"Neither" category (blue) depicts the greatest variability, with a range of 4–17 characters and a median of around 7 characters, indicating a mix of shorter and longer words. This also denotes the existence of affixes that influence word length. In contrast, the "Shared" category (orange) contains mostly shorter words, ranging from 5 to 15 characters and a median of around 6 characters. This group has the least variation, as shown by the narrow distribution, indicating that shared words are generally more consistent in length. The "PRPM-only" category (green) shows similar features between 2 and 15 characters,

with a median of 7 characters. This distribution is slightly more uniform compared to "Neither," reflecting less variation in word length, but still with some diversity in size.

The "KBBI-only" category (red) ranges from 2 to 14 characters (median 6 characters), similar to the "Shared" category. This indicates that words unique to KBBI are also generally of moderate length, with moderate variation. Overall, the plot reveals that words in the "Neither" category tend to have the longest and most varied lengths, while the "Shared", "PRPM-only", and "KBBI-only" categories feature words that are generally shorter and more consistent in length. The "PRPM-only" and "KBBI-only" words are slightly shorter on average than those in the "Neither" category but share a similar range of variability.

3.2. Vocabulary, Morphology, and Orthography

The clean list of extracted vocabularies from the manuscript constitutes unique tokens that are essential to be examined to reveal the most frequent words. This objective, in the bigger implication, will give the predominant theme depiction as the outcome of this study. The 50 most frequent words are identified and shown in Figure 5. This distribution can be described as having three main segments: the plot shows an extremely steep initial drop among the first few high-frequency words, with the most frequent word (*Raja*) reaching approximately 400, rapidly falling to around 200 to 250 by the fourth or fifth ranked word.

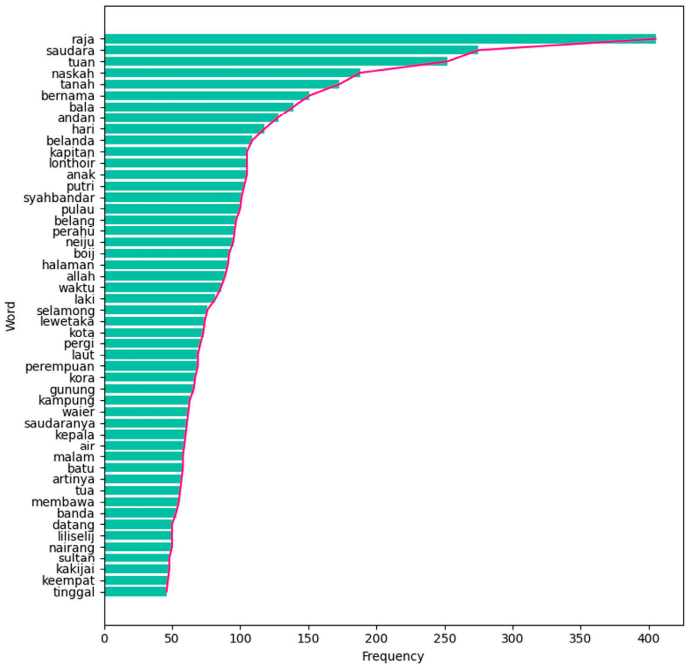


Figure 5. Top 50 most frequent words.

Following this head, the curve enters a moderate decline across the mid-frequency range, which spans roughly from 200 down to 100 (e.g., from *bernama* down to *perempuan*), where the frequency decreases steadily but at a much slower rate. Finally, the distribution ends in a long, flat tail where a vast number of unique words (e.g., *tuan*, *membawa*, *tinggal*) cluster around very similar, low frequencies, specifically centered around 50, confirming that the majority of the vocabulary contributes minimally to the total word count.

Raja (king) indicates that monarchy or royal figures are central to the saga, involving topics related to the kingdom, rulers, and governance. The second-ranked words, *saudara* (brother/sibling/relative), *tuan* (sir/master), and *bala* (troopers), suggest formal address or social hierarchy, showing interpersonal or structured societal roles. Other commonly used terms, such as *tanah* (land), show the essence of territory, while *belanda* (Dutch), *kapitan*

(captain), and *syahbandar* (harbor master) hint at a colonial or maritime context, emphasizing the period of European influence in the Banda region. *Anak* (child), *putri* (princess), and *perempuan* (woman) show family and gender elements in the narrative. *Perahu* (boat), *laut* (sea), and *banda* (Banda Islands) reinforce the idea that the setting involves coastal or island communities, typical of the Maluku archipelago, specifically Banda. Words such as *Allah* (God) and *malam* (night) suggest spiritual or poetic undertones, possibly reflecting the cultural and religious environment of the time.

In morphology, affix analysis shows that prefixation dominates the corpus, with action-oriented morphemes such as *me-* and *di-* most salient, signaling active and passive voice in narrative clauses (e.g., *mendengar, membawa, menjawab* versus *disebut, dibawa, diangkat, dijawab*), shown in Figure 6. The bar chart shows that prefixes are more frequent than suffixes in the corpus, with 906 prefixed words and 616 suffixed words in total. The three most frequent prefixes are *me-* with 290 tokens (32.0%), *ber-* with 179 tokens (19.8%), and *di-* with 172 tokens (19.0%), followed by *ke-* (73 tokens; 8.1%), *ter-* (52 tokens; 5.7%), *pe-* (47 tokens; 5.2%), *se-* (45 tokens; 5.0%), *per-* (42 tokens; 4.6%), and *ku-* (6 tokens; 0.7%). In contrast, suffix usage is more concentrated in a few forms. The most dominant suffix is *-an* with 324 tokens (52.6%), followed by *-nya* with 114 tokens (18.5%), *-lah* with 107 tokens (17.4%), *-i* with 51 tokens (8.3%), and the less frequent *-ku* and *-mu*, each appearing 10 times (1.6%).

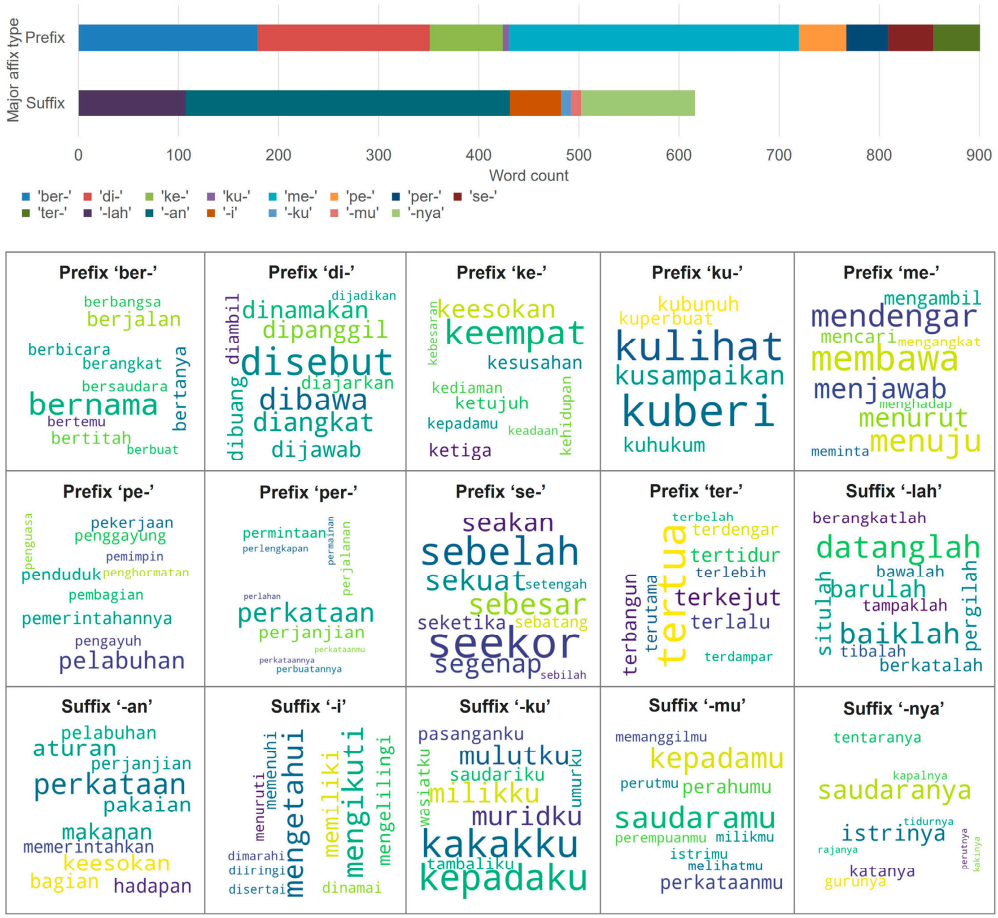


Figure 6. Bar charts (**top**) visualize the major affixed words and their frequency, clustered based on prefix and suffix. The word clouds (**bottom**) show the top 10 words for each cluster for prefixes (*'ber-', 'di-', 'ke-', 'ku-', 'me-', 'pe-', 'per-', 'se-', 'ter-'*) and suffixes (*'-lah', '-an', '-i', '-ku', '-mu', '-nya'*). The yellow-to-purple color gradient indicates word frequency, with yellow marking lower frequency and purple marking higher frequency.

Lexemes with *ber-* foreground intransitive or middle actions (*bernama*, *berjalan*, *berbicara*), while *pe-* and *per-* nominalize processes and roles (*pekerjaan*, *penduduk*, *pemimpin*; *perkataan*, *perjanjian*), helping build the social and institutional texture of the story world. Quantifying and delimiting meanings surface through *se-* (*sebelah*, *sebesar*, *seekor*, *segenap*), and *ke-* frequently marks ordinals and abstract states (*keempat*, *ketiga*, *keesokan*, *kesusahan*), structuring chronology and circumstance; *ter-* highlights stative or superlative readings and suddenness/emergence (*tentulah*, *terkejut*, *terdengar*), fitting moments of discovery or affect.

Suffixation is also prominent, led by *-nya*, which encodes definiteness and third-person possession (*saudaranya*, *istrinya*), indicating referential continuity across scenes; *-an* productively derives nouns and collectives (*perkataan*, *perjanjian*, *pakaian*), while *-i* marks applicative or locative verbal nuance (*mengetahui*, *mengikuti*). Possessive closeness and address appear in *-ku* and *-mu* (*mulutku*, *milikku*, *kepadamu*), and the pragmatic particle *-lah* (*datanglah*, *baiklah*) punctuates directive or emphatic turns. Together, these affix patterns show a narrative that is densely verbal, agentive, and referentially cohesive: voice alternations drive event progression, nominalizations scaffold themes and institutions, quantifiers order time and space, and possession/definiteness maintain character continuity.

In terms of spelling signals, the referentiality to dictionaries was examined to identify orthographic patterns. Figure 7 shows the orthographic skewness per-10k-chars frequency (i.e., sy, kh, ny, ng, ai, au, dz) across categories, revealing whether sy/kh skew PRPM-only and ny/ng or certain diphthongs skew elsewhere. Word length can provide valuable insights into linguistic patterns and help us understand the nature of the vocabulary within different categories. In the "Shared" category, shorter words (typically 4–6 letters) dominate, indicating that these words likely represent core vocabulary with common roots shared between Malay and Indonesian.

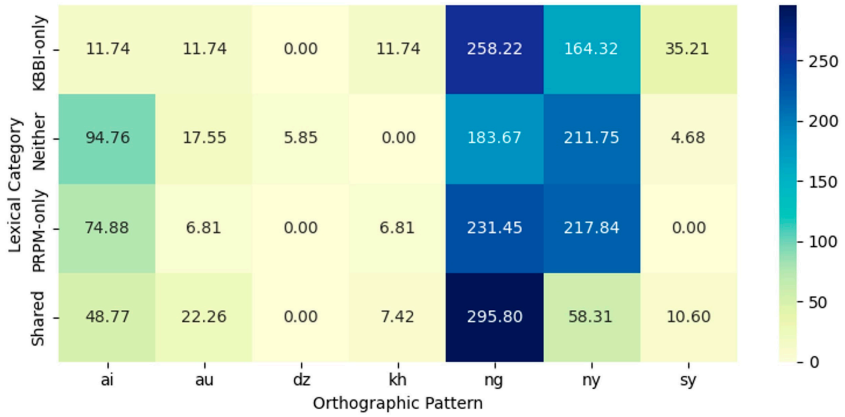


Figure 7. Orthographic pattern (normalized) frequency per 10k characters by category.

In contrast, the "KBBI-only" category contains longer words, which are often affixed or derived forms of Indonesian, such as *menyampaikan* or *peraturan*. These longer words suggest a higher degree of morphological complexity in Indonesian vocabulary. The "PRPM-only" category, with its relatively longer words, seems to reflect classical or Arabic-influenced spellings, seen in examples like *syahbandar* or *khidmat*. This points to the historical influence of Arabic and classical Malay on the lexicon of PRPM. Lastly, the presence of very long or short words in the "Neither" category could indicate the inclusion of archaic terms, compound words, or possibly misspelt tokens, reflecting the complexity or inconsistency in word usage that falls outside both dictionaries. These insights reveal how word length and structure can reflect the underlying linguistic, historical, and cultural influences on the language.

The orthographic footprint of the *Van Ophuijsen* system measurably shapes the corpus: graphemic correspondences such as oe → /u/ (*boelan*, *boenga*, *batoe*), dj → /j/ (*djakaria*, *djohor*), tj → /c/ (*tjakbeir*, *tjekalele*), and j → /y/ (*aij*, *ajeir*, *kelij*) systematically inflate the "Neither" bucket in dictionary matching and complicate downstream tasks (tokenization, lemmatization, and NER). Empirically, many tokens flagged as out-of-vocabulary resolve to regular Indonesian/Malay forms after orthography-aware normalization (e.g., *boelan* → *bulan*; *djohor* → *johor*; *tjakbeir* → *cakbir*), yet we deliberately preserve the pre-normalization forms for onomastic fidelity because the archaic layer is dominated by proper names, such as individuals, toponyms, boats, and ritual objects (e.g., *Raja Noeilaj*, *Gunung Oeloepitoe*, *Belang Limareij*).

This dual track (reversible normalization for lexical analytics and retention for named entity integrity) reduces false negatives in dictionary lookups while safeguarding historical signal for narrative mapping. Practically, we implement a deterministic rewrite table (regex rules in Table 5) before stemming/lemmatization, then restore original spellings at render time for quotations and entity graphs; this yields cleaner frequency profiles without erasing the diachronic character of the manuscript and explains why archaic lexicons appear chiefly as names rather than productive lemmas in the morphological inventory.

Table 5. Several examples of old-spelt words (The *Van Ophuijsen*).

Regex Patterns	Identified Word Examples
oe → /u/	aboetahier, agastoe, batoe, benoeh, boediman, boeij, boelan, boenga, boeroeng, coen
dj → /j/	djabar, djabarollah, djailin, djakaria, djakariang, djein, djo, djodjau, djohor, djopati
tj → /c/	koembangratji, pantjibei, prentji, ratjie, sepantjibai, tjakbeir, tjamara, tjekalele, tjekalelei
j → /y/	aij, ajeir, aulija, batij, boij, eij, kakijai, kelij, kijakbir, lekleij

In the loanwords (foreign terminology), the inventory reveals two principal streams of borrowing that intersect directly with the manuscript's religious and colonial chronicle: Arabic and Dutch. The Arabic stratum ranges from greetings and ritual terms (*assalamualaikum*, *azan*, *zuhur*), through theological lexis (*Allah*, *gaib*, *rahmat*, *illallah*), to religious onomastics (*Abdul-*, *Abdullah*, *Achmad*, *Auliya*, *Syekh*), signaling a strong Islamic register and functions as a discourse marker for prayer, legitimacy, and scholarly authority, while also forming dense clusters of named entities for NER (see Table 6). Orthographically, several occur in historical spellings, which can evade modern dictionaries. Hence, we apply reversible normalization during preprocessing and restore original forms at presentation to preserve historical fidelity.

By contrast, the Dutch layer (see Table 7), such as *compania/VOC*, *gubernur*, *gulden*, *kapitan*, *perk (plantation)*, *vandel (banner)*, along with administrative surnames and titles such as *Verhoven* and supervisory roles like *Velak*, encodes the colonial infrastructure of the spice economy, governance, and coercive power in Banda. Methodologically, the two tables complement the orthographic and corpus analyses by (i) providing controlled lists for domain-semantic labelling (Religion; Colonial/Political–Economic), (ii) informing entity-sensitive orthographic rules, and (iii) reducing false negatives in dictionary matching and named entity extraction. In sum, Arabic borrowings inscribe ritual–genealogical networks, Dutch borrowings inscribe institutional–economic networks, and together they explain why most "foreign" tokens surface as proper names or domain-specific terminology rather than productive lemmas, making cross-linguistic handling and historical orthography central to analytic accuracy in this manuscript.

Table 6. Several Arabic loan words and foreign terminologies, identified from the dataset.

Arabic Term	English Translation/Explanation
abdul	Servant of (usually part of a compound name, e.g., Abdul Rahman = Servant of the Most Merciful)
abdulgader	Servant of *Al-Qadir* (God, the All-Powerful)
abdullah	Servant of God
aboetahier	Father of Tahier/a personal name (*Abu* = father of)
achmad	A variant of Ahmad, a male given name meaning "most commendable"
ahad	One/Unique (as in "God is One"); Sunday
allah	God
assalamualaikum	Peace be upon you (Islamic greeting)
auliya	Saints/close friends of God
azan	Call to prayer
gaib	Unseen/hidden/supernatural
halal	Permissible/lawful (according to Islamic law)
idznillah	With God's permission
illallah	There is no god but God
insya	If God wills/God willing
joesoep	Variant of Yusuf/Joseph (personal name)
khatib	Preacher or imam delivering the sermon (usually on Friday)
kiblat	*Qibla*/direction of prayer toward Mecca
lailatulqadar	Night of Decree/Night of Power in *Ramadan*
malaikat	Angels
mufakat	Agreement/consensus
pikir	Thought/thinking
rahmat	Mercy/blessing (from God)
tarekat	Sufi spiritual path/order
ulama	Religious scholars
sahur	Pre-dawn meal before fasting (*Ramadan*)
syekh	*Sheikh*/Islamic teacher or elder
syiar	Islamic propagation/spreading the message of Islam
zuhur	Noon prayer (*Dhuhr* prayer)

Table 7. Several Dutch loan words and foreign terminologies, identified from the dataset.

Dutch Term	English Translation/Explanation
compania	Company (usually referring to the Dutch East India Company, VOC)
gubernur	Governor
gulden	Guilder (Dutch currency used in the colonial period)
kapitan	Captain/local leader (used in colonial administration, often "*Kapitan Cina*")
perk	Plantation/plot of land (especially spice plantations in the Banda Islands)
vandel	Banner/flag
verhoven	Surname/proper name (e.g., Dutch official or settler)
velak	Plantation overseer/foreman (historical colonial term in the Maluku context)

3.3. Semantic Fields Analysis

The profile of semantic patterns across four themes (Maritime, Religious, Social/Genealogical, and Place/Geographic), partitioned into 15 sequences ordered by descending word counts (see Figure 8). The class divisions were not exactly equal. Instead, this was only to map the sequential order of the persistent words based on their frequency. The sequential blocks show that Sequences **1–4** are led by maritime and social/political

governance vocabulary (e.g., *perahu*, *pelabuhan*, *kora*, *raja*, *tuan*, *kapitan*, *Belanda*), which consistently co-activate place/geographic anchors (*Banda*, *pulau*, *kota*); together, these establish seaborne movement and authority relations as the narrative spine.

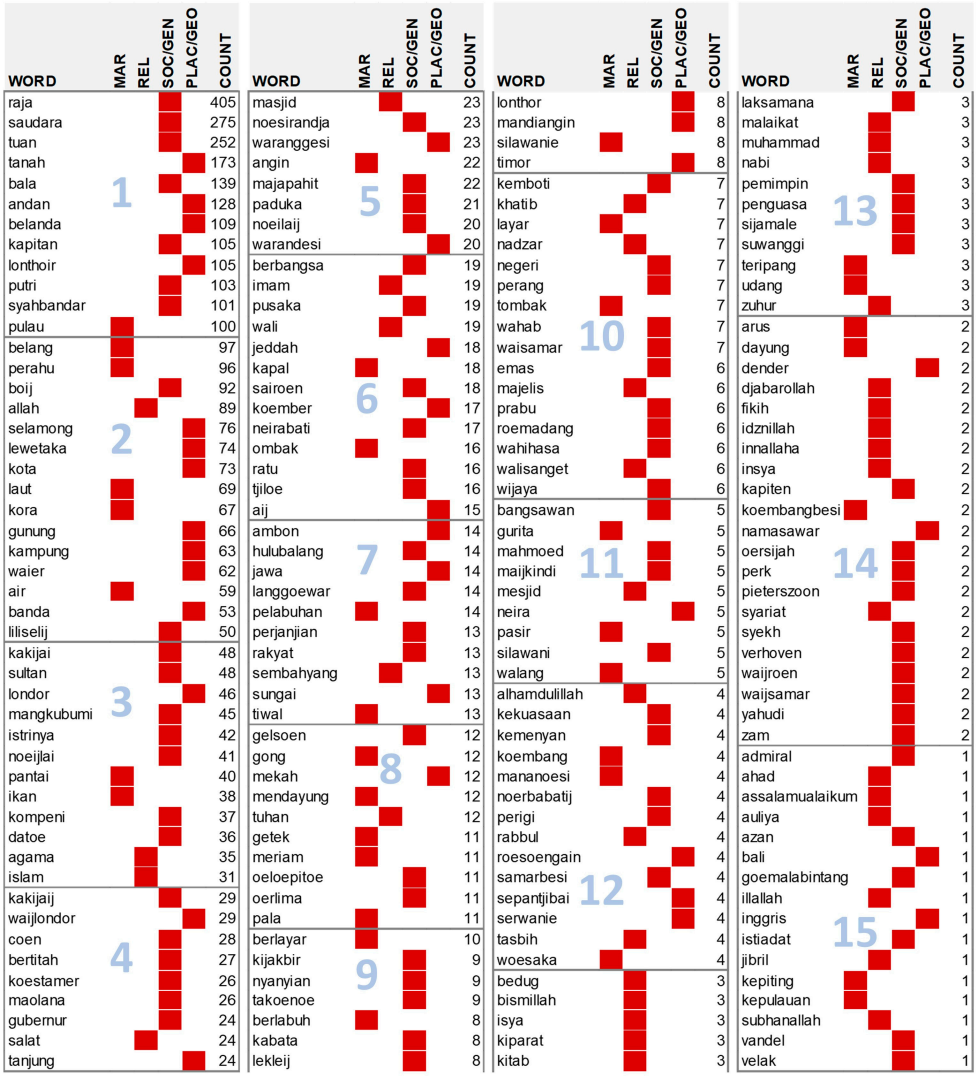


Figure 8. Semantic patterns for identified themes: Maritime, Religious, Social/Genealogical, and Place/Geographic, divided into 15 sequences according to the order of the word counts.

Sequences **5–9** thicken around geospatial nodes and mobility (*Neira*, *Ambon*, *Java*, *Timor*; *ombak*, *layar*), while rank/household terms (*bangsawan*, *rakyat*, *saudara*) knit actors into hierarchies; many tokens here light up multiple columns, acting as hinges that connect voyage → port → magistrate → kin network. Sequences **10–13** introduce concentrated religious bursts, such as institutions, roles, and formulae (*masjid*, *khatib*, *syekh*; *sembahyang*, *bismillah*, *innallah*), which appear as spikes rather than a steady background, marking oaths, blessings, and legitimizing passages. The tail (**14–15**) carries colonial–administrative/Dutch items (*laksamana*, *perk*, *Pieterszoon*, *vandel*, *velak*) plus occasional fauna/objects, signaling episodic detail within governance scenes. Overall, the **15** sequences reveal a skewed but coherent ecology: maritime and politico-social terms persist longest and structure the storyline; religious lexicon punctuates key ritual or moral junctures; and place terms scaffold movement throughout.

Principal Component Analysis (PCA) was performed on the word-context frequency matrix to plot the spatial distribution of words and assess their underlying usage patterns

(see Figure 9). The initial frequency data was processed using one-hot encoding (or a similar binary representation), where each word forms an observation vector and its presence or absence across different contexts forms the features [23].

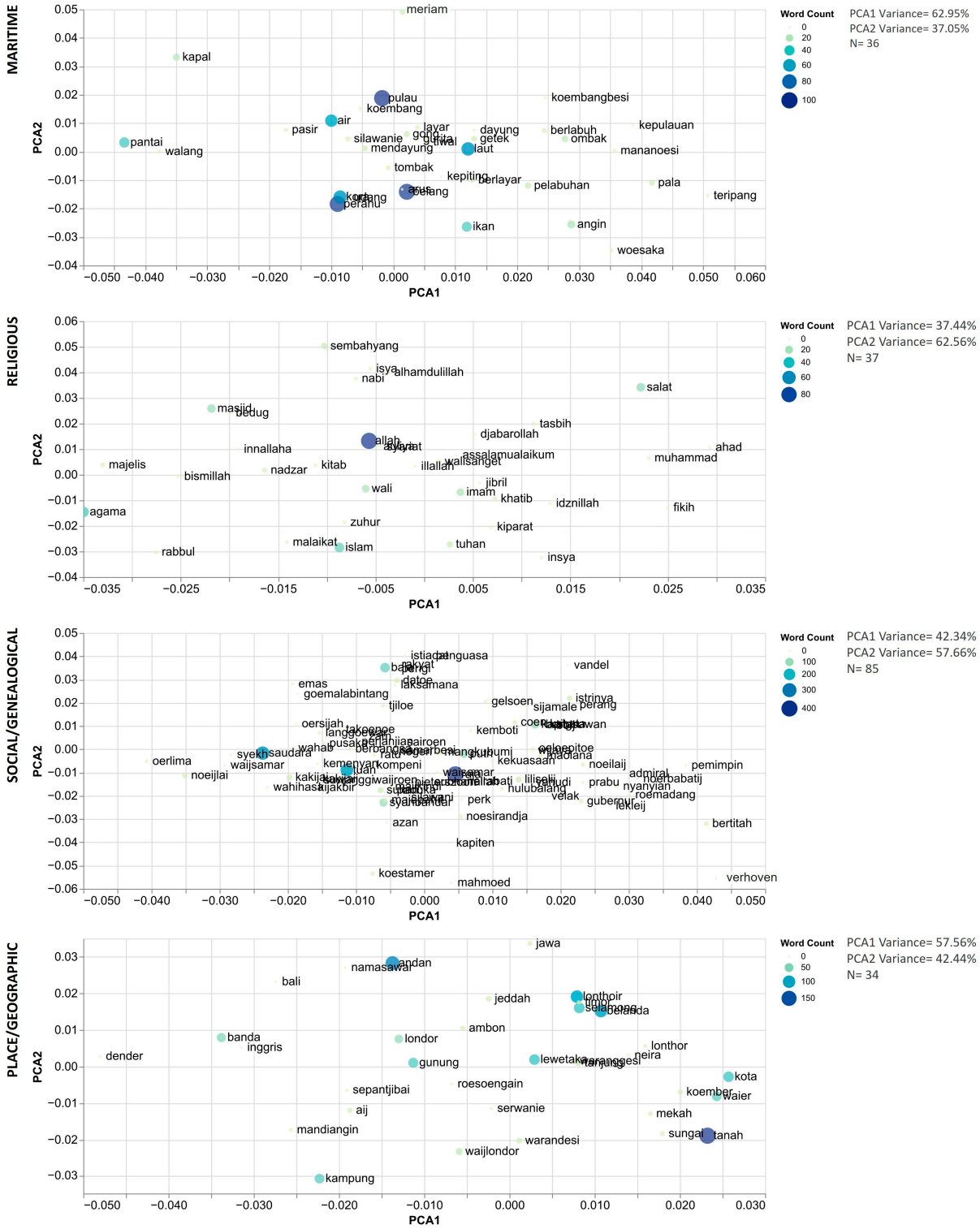


Figure 9. PCA of one-hot encoded word frequencies. Scatter plots show word distribution across themes; closer words have similar usage patterns in the manuscript.

Each point represents a single word, and the points are color-coded according to their pre-assigned semantic groups. Proximity between points indicates a high degree of similarity in the words' contextual usage patterns, suggesting a shared semantic or thematic

meaning, while distant points represent words with distinct frequency distributions. This visual grouping is crucial, as it provides immediate confirmation that the dimensions extracted by the PCA effectively differentiate the words according to their labeled thematic categories, validating the dimensionality reduction process.

The Maritime group forms a noticeable cluster, generally positioned near the central vertical axis, confirming internal heterogeneity, which is dominated by PCA1 (62.95%). In contrast, the Social/Genealogic group, heavily driven by PCA2 (57.66%), tends to form a large, central, and often dense cluster, indicating that while these terms are frequent, their primary separation may be driven by lower-ranked components or that they exhibit a usage pattern closer to the overall average. The Religious (dominated PCA2 = 62.56%)and Place/Geographic group (PCA1 = 57.56%) form well-defined, generally tight clusters, suggesting internal consistency in how these terms are used relative to all other words.

3.4. Key Entities Identification and Relationships Mapping

Figure 10 shows the hierarchical clustering of entity types based on their estimated frequency, visualized in a dendrogram. In this context, it groups entities with similar frequencies closer together using Ward's method, which minimizes the variance of the clusters being merged. The Euclidean distance was used on the scaled frequency values as the distance metric [24]. The vertical lines (or branches) represent the distance between the clusters (or individual entities) being merged. Several clusters were identified:

1. Cluster 1 (Low Frequency): This includes Place (GEOG) and Social/Genealogical group (PGRP), which are grouped first due to their very similar and low frequencies.
2. Cluster 2 (Mid-Frequency): This cluster consists of Maritime (PROC), Concept (CONC), Object/Physical Entity (DEVI/TISS), and Event (ACTI), all of which have frequencies clustered in the low hundreds.
3. Outlier: Person (LIVB) stands alone, merging with the main clusters at a much greater distance, as its estimated frequency is significantly higher than all the other entity types.

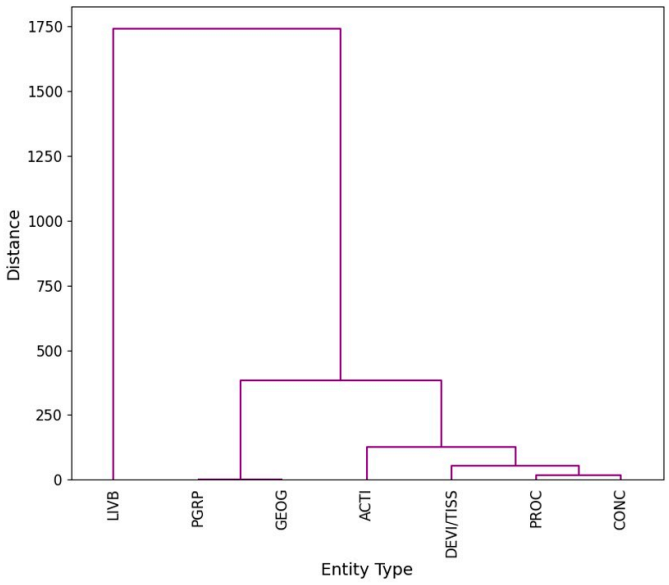


Figure 10. Entity types based on estimated frequency.

Table 8 presents a list of identified entities and their specified relationship types, highlighting how different entities are interconnected within the historical and cultural framework. For example, *Nabiullah Nuh* and *Andara (Banda)* are connected by a Geographic relationship, indicating that Andara is considered the first land to emerge after Noah's flood.

Jailin and *Siti Gelsoen* share a Genealogical relationship, marking them as the ancestral couple and the first major figures in the narrative.

Table 8. Identified entities and specified relationship types. The list has been simplified to increase the legibility.

Source (Node A)	Target (Node B)	Relationship Type	Notes
Nabiullah Nuh	Andara (Banda)	Geographic	Andara was the first land to emerge after Noah's flood.
Jailin	Siti Gelsoen	Genealogical/Person	The ancestral couple and the first major figure.
Siti Gelsoen	Gunung Oeloepitoa	Meaning/Geographic	Siti Gelsoen is buried there (a revered place).
Raja Noeilaj	Putri Cilu Bintang	Genealogical/Person	Siblings (the most important relationship).
Raja Noeilaj	Banda	Power/Person	The first and most important king in Banda/Andara.
Raja Noeilaj	Liliselij	Genealogical/Person	Siblings and major leaders who have separated.
Liliselij	Lautaka (Lewetaka)	Power/Person	King in Lautaka (regional power relationship).
Liliselij	Makkah	Religion/Geographic	One of four brothers who sailed and studied Islam.
Kakijaij	Jeddah	Journey/Geographic	Kakijaij was left in Jeddah (a crucial point in the journey).
Kakijaij	Nūr Al-Mubīn	Religion/Object	Kakijaij brought the Nūr Al-Mubīn (Holy Book) back to Banda.
Kakijaij	Makkah	Religion/Geographic	Following his brothers to study Islam there.
Perjalanan ke Warandesi	Belang Limareij	Transport/Object	The main voyage was on the Belang Limareij.
Perjalanan ke Warandesi	Warandesi	Geographic	The main purpose of the brothers' voyage.
Journey ke Mekah	Islamisasi	Implication/ Geographic	This journey implies the spread/introduction of Islam.
Majapahit	Prabu Wijaya	Geographic	A journey to Banda
Prabu Wijaya	Mahar	Marriage/Object	Marriage to Putri Cilu Bintang.
Mahar	Putri Cilu Bintang	Marriage/Event	Marriage to Putri Cilu Bintang.
Tuan Coen	Banda	Conflict/Concept	Central figures in the Dutch colonial conflict with Banda.
Tuan Coen	Perjanjian	Politics/Concept	The Dutch attempted to make a treaty with Banda.
Boij Ratang	Raja Noesniwi	Conflict/Person	The relationship began with a rejected proposal.
Pohon Kayu Lima	Siti Gelsoen	Concept/Person	The magical pomegranate that led to Siti Gelsoen's pregnancy.
Portugis	Gunung Tabaliku	Battle/Geographic	Geographical representation of Portuguese/Dutch battles and bases.
Neirabatij	Lonthoir	Power/Person	Heads of State/local rulers under colonial rule.

Other relationships are more complex, such as *Raja Noeilaj* and *Putri Cilu Bintang*, who share a strong Genealogical connection as siblings. *Raja Noeilaj* also has significant Power connections, notably with Banda, where he is identified as the first and most important king. *Liliselij*, another key figure, has multiple relationships, such as with *Lautaka (Lewetaka)* (regional power) and *Makkah* (religious journey), demonstrating a blend of political and religious ties. Other entries in the table also reveal the interconnectedness of religion, such as *Kakijaij* and *Makkah*, where his journey to study Islam. *Perjalanan ke Warandesi* shows the voyage's significance, linking to the *Belang Limareij* (a transport vessel) and the geographic destination of *Warandesi*. In addition, there are entries related to conflict and colonial influence, such as *Tuan Coen*'s relationship with *Banda*, signifying a Conflict/Power dynamic, and the *Perjanjian* (treaty), illustrating the Dutch colonial attempts.

Figure 11 maps two dominant clusters, as the general theme introduced by the saga. First, Genealogical/Local cluster highlights several entities such as *Raja Noeilaj*, *Siti Gelsoen*, *Putri Cilu Bintang*, and *Jailin*, showcasing the story's focus on the origin and kinship in Banda. The second cluster (Migration/Religion) highlights agglomerated entities such as *Kakijaij*, *Liliselij*, *Makkah*, and *Jeddah*, exhibiting the theme of journey and Islamic introduction to Banda. Several characters are also identified as bridging entities that link one story fragment with another story, such as *Kakijaij* and *Liliselij*, connecting the Genealogical cluster (*Banda*, *saudara*) with Religion/Migration (*Makkah*, *Jeddah*, *Nūr Al-Mubīn*). These entities reflect their prophetic role, which brought religious and cultural changes to Banda.

Another character, *Tuan Coen*, also constitutes an isolated node (connected to *Banda* and *Perjanjian* only). This reflects his status as an external actor and the primary source of conflict, interacting directly with the Banda rulers.

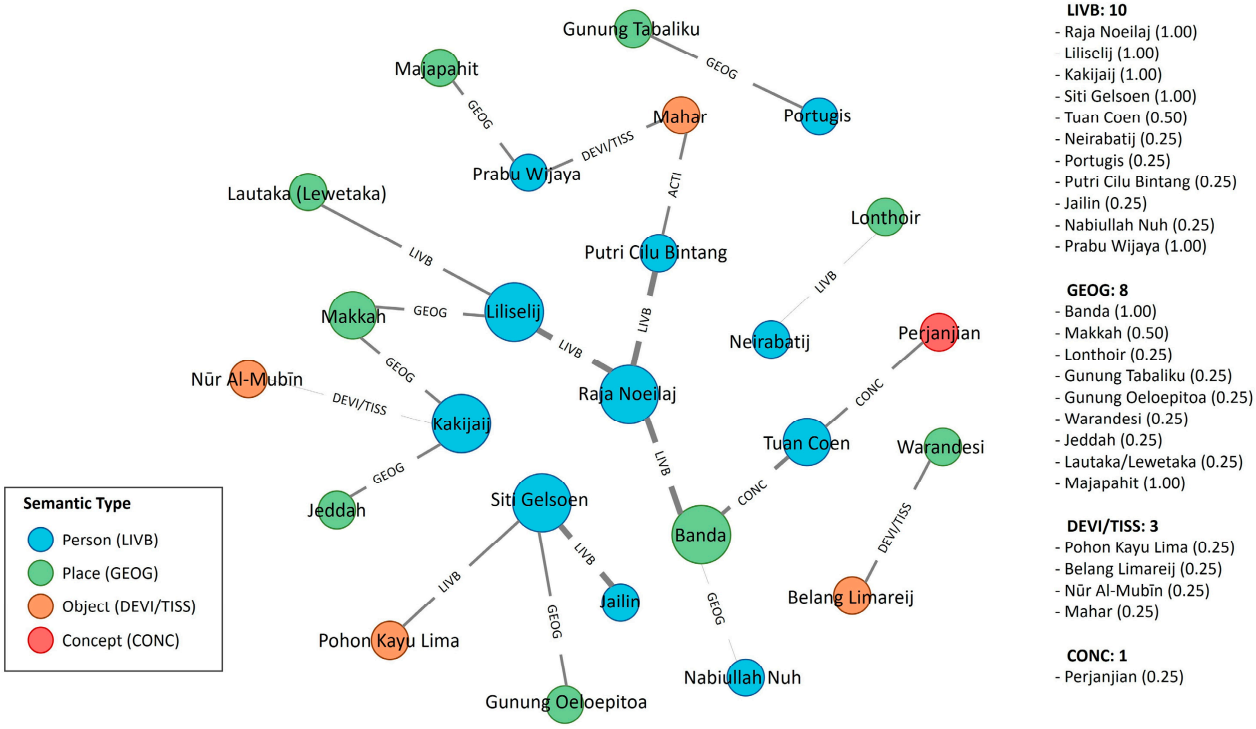


Figure 11. Map of the main entity relationship network from four main semantic types.

Degree centrality measures the number of direct connections (edges) a node has in a network or graph. It is used in network analysis to quantify the importance or influence of a node based on how many other nodes it is directly connected to. The degree values range from 0 to 1. The higher the degree centrality, the more connections an entity has. For example, the entities *Raja Noeilaj*, *Banda*, *Liliselij*, *Kakijaij*, and *Siti Gelsoen* each have a degree centrality of 1.00, indicating that they are each directly connected to roughly the same number of other entities in the network. In contrast, entities such as *Gunung Oeloepitoa* and *Belang Limareij* have less connection and are more peripheral, with a value of 0.25. The entities are categorized into four types. Person (LIVB) category tends to have a higher centrality value, suggesting they might play a more central role in the network compared to Place (GEOG), Object (DEVI/TISS), and Concept (CONC), which generally have lower centrality. This provides a clear explanation, strengthened by quantitative and visual representations of how entities in the saga interact with each other, shaping a complex cultural narrative.

Raja Noeilaj, marked by its size and central location, forms the nucleus of the network [25]. This position shows the thickest, strongest relational links to key Person entities *Liliseij*, *Kakijaj*, and *Siti Gelsoen*, and the primary Place node, *Banda*, signifying these as the most frequent and important associations in the underlying narrative. The graph further reveals interconnected sub-networks, with the core group of people linking to important geographical hubs like *Jeddah* and *Makkah*, suggesting themes of trade or pilgrimage, while other connections involve political entities like *Portugis* and abstract elements such as *Perjanjian* (Treaty) and *Mahar* (Dowry). Overall, the visualization effectively maps the complex historical and social landscape surrounding *Raja Noeilaj*, identifying not only the central figures and locations but also the critical objects and concepts (like the single CONC node, *Perjanjian*) that define the political and historical narrative being represented.

3.5. Comparative Folkloristics: Reflection on Our Findings

3.5.1. Regional Comparison and Transferability

The linguistic and narrative patterns observed in *Hikayat Lonthoir* align with broader features of Eastern Indonesian and Islamic maritime literary traditions while retaining several Banda-specific characteristics. Within Maluku, parallels may be drawn with Ternate–Tidore chronicles such as the *Hikayat Ternate* (1859–1864), the *Hikayat Hitu*, and *Bobato* genealogies, such as the *Hikayat Ternate* and *Bobato* genealogies, which similarly intertwine sacred genealogy, Islamic legitimation, and royal authority. Like *Hikayat Lonthoir*, these chronicles deploy Arabic-derived religious lexicon and embed origin narratives within Qur'anic cosmology. However, compared with the highly court-centered, Islamized, and politicized Ternate narratives, the Banda saga displays a stronger emphasis on migration, seafaring mobility, and regional dispersal of kin groups, consistent with Banda's historical role as a multi-island maritime network rather than a centralized sultanate [26].

Javanese *babad* traditions offer an additional point of comparison. Although *babad* texts also integrate miraculous births, political legitimation, and encounters with foreign powers, their narrative texture is shaped by classical Javanese court culture, Sanskrit heritage, and a cyclical conception of kingship. In contrast, *Hikayat Lonthoir* exhibits a lexical core built upon Malay nautical terminology, *Van Ophuijsen* orthography, and Arabic-Islamic narrative structuring, with little Sanskrit or Old Javanese influence. This suggests that while miracle motifs and genealogical legitimacy are regionally shared, the Banda corpus's specific configuration of maritime mobility, island-to-island dispersal, and interaction with Dutch colonial administration marks it as distinct within the eastern archipelago [27].

Within the broader Malay world, the manuscript shares affinities with Islamic maritime narratives such as the *Hikayat Raja-raja Pasai*, *Hikayat Aceh*, and localized *mi'raj* adaptations. These traditions commonly feature Qur'anic intertextuality, supernatural intervention, and Arabic technical vocabulary relating to ritual, cosmology, and kinship [28]. The loanword profile in *Hikayat Lonthoir* similarly reflects the Islamization processes typical of the 17th–19th centuries in maritime Southeast Asia. However, the Dutch lexical layer is unusually dense compared with mainland Malay manuscripts, underscoring Banda's entanglement with spice monopoly governance and VOC administrative structure.

In relation to Islamic literary traditions and standardization, Malay manuscripts from Aceh, Malacca, and Ternate often exhibit parallel Arabic-derived lexicon and narrative structures, supporting the view that *Hikayat Lonthoir* belongs to a transregional Islamic manuscript ecosystem. Meanwhile, the orthographic profile reflects early 20th-century *Van Ophuijsen* standardization, situating the manuscript within the broader evolution of Indonesian orthography prior to the 1972 improved spelling (EYD) reform. Case-specific components, on the other hand, address the unique context of Banda, encompassing Banda-specific orthographic rules, a gazetteer for local entities, a Dutch colonial lexical layer, and the mapping of folklore motifs tied to Banda oral histories [29,30].

The processing pipeline demonstrates a combination of transferable and case-specific components. Transferable elements include OCR with human-supervised correction, historical spelling normalization, dictionary-based validation, corpus linguistics profiling, semantic embedding with clustering, and named entity recognition alongside co-occurrence analysis. To address these issues, historical text normalization pipelines were established, such as rule-based orthographic mapping, OCR error modelling, and hybrid human-in-the-loop correction, as essential for processing pre-modern manuscripts in low-resource languages [31,32].

3.5.2. Global Folklore Classification Systems

Folklorists seek to arrange the different knowledge from across literature around the world and connect them, and their recurring patterns, such as motifs, tropes, and other storytelling elements, serve as a 'biological genes' analogy in cultural heritage [33]. This helps folklorists understand how similar value systems are expressed in diverse cultures and how motifs evolve across regions [34–36]. To investigate this common thread, a cross-check was conducted against global folklore classification systems such as the Aarne-Thompson-Uther (ATU) [37] and Stith Thompson's Motif-Index of Folk-Literature [38] (see Figure 12).

The various motifs identified from previous findings were matched through searches of corresponding keywords, and then discussed qualitatively to obtain a broader understanding of the universal values adopted. For instance, the motif of Supernatural Ancestry and Origin highlights "supernatural birth" (ATU 705A and T511.7) involves miraculous births through supernatural or divine means, by eating food; in this context pomegranate was eaten by Siti Gelsoen and got pregnant. As quoted from the normalized manuscript:

> "*Akhirnya, pada suatu hari, ada seorang Bernama Jailin yang mendapat buah itu, lalu membawanya kepada Siti Gelsoen. Maka buah itu dimakanlah oleh Siti Gelsoen. Dengan kuasa Allah Ta'ala, ia pun mengandung dan melahirkan dua anak laki-laki.*"
>
> Translation:
>
> "Finally, one day, someone named Jailin found the fruit and brought it to Siti Gelsoen. Siti Gelsoen ate the fruit (pomegranate). By the power of Allah Ta'ala, she became pregnant and gave birth to two sons."

A similar concept of eating a child born from a magical or divine source, like a tree, also occurred to Boij Ratang Princess, who unintentionally swallowed a small fish. Other themes in this supernatural cause theme are found in Thompson's Index, such as "miraculous provider" (D1470), showcasing the power of Wali Allah's prayer to fulfil his wife's wish, and "mythological place origin" (A900–A999), showing the story of the Banda Archipelago origin as the aftermath of Noah's flood, as shown in the quote:

> "*Pada permulaan zaman Nabiullah Nuh… setelah selesai peristiwa banjir besar… tanah yang pertama kali muncul di sebelah timur adalah tanah Andara (artinya Banda).*"
>
> Translation:
>
> "At the beginning of the era of Prophet Nuh… after the great flood… the first land to appear in the east was Andara (Banda)."

This reflects that the *Hikayat* interpretation provides God's existential discourse [39] through an implicit Qur'anic narrative where diverse descriptions of it are found in the hadith literature, as well as vernacular genres [40] as interpretive keys for cultural transmission of religious philology [41].

The Hero's Journey and Spiritual Power (Kakijaij): "neglected younger brother" (ATU 500–559 and L10), "gaining divine knowledge" (ATU 750–849 and D1720), "magic flight/rescue" (E341), "magic transformation" (D100–D199 and D1960). Political & Ritualistic Motifs (Agreement & Conflict): "covenant/oaths" (M185), "broken vows and punishment" (M200), "the false accusation plot" (ATU 705–712 and K2110.1). Separation, Rescue, and Return: "separation by storm" (R175), "miraculous seafaring" (D2120), "the abandoned/lost hero" (L10 and R320), "post-mortem transportation" (E341 and B557.7). Prabu Wijaya's Marriage to Putri Cilu Bintang: "high-status marriage to a foreign king" (ATU 870 and T111), "test/condition for marriage" (H300–H499 and H36), "wife's supernatural worth" (ATU 465 and T11.2), "the royal covenant" (P14.12), "marriage by decree" (T50.2). This narrative documentation helps contextualize the identified motifs in *Hikayat Lonthoir*, reinforcing the connection between this story and global folklore traditions [42].

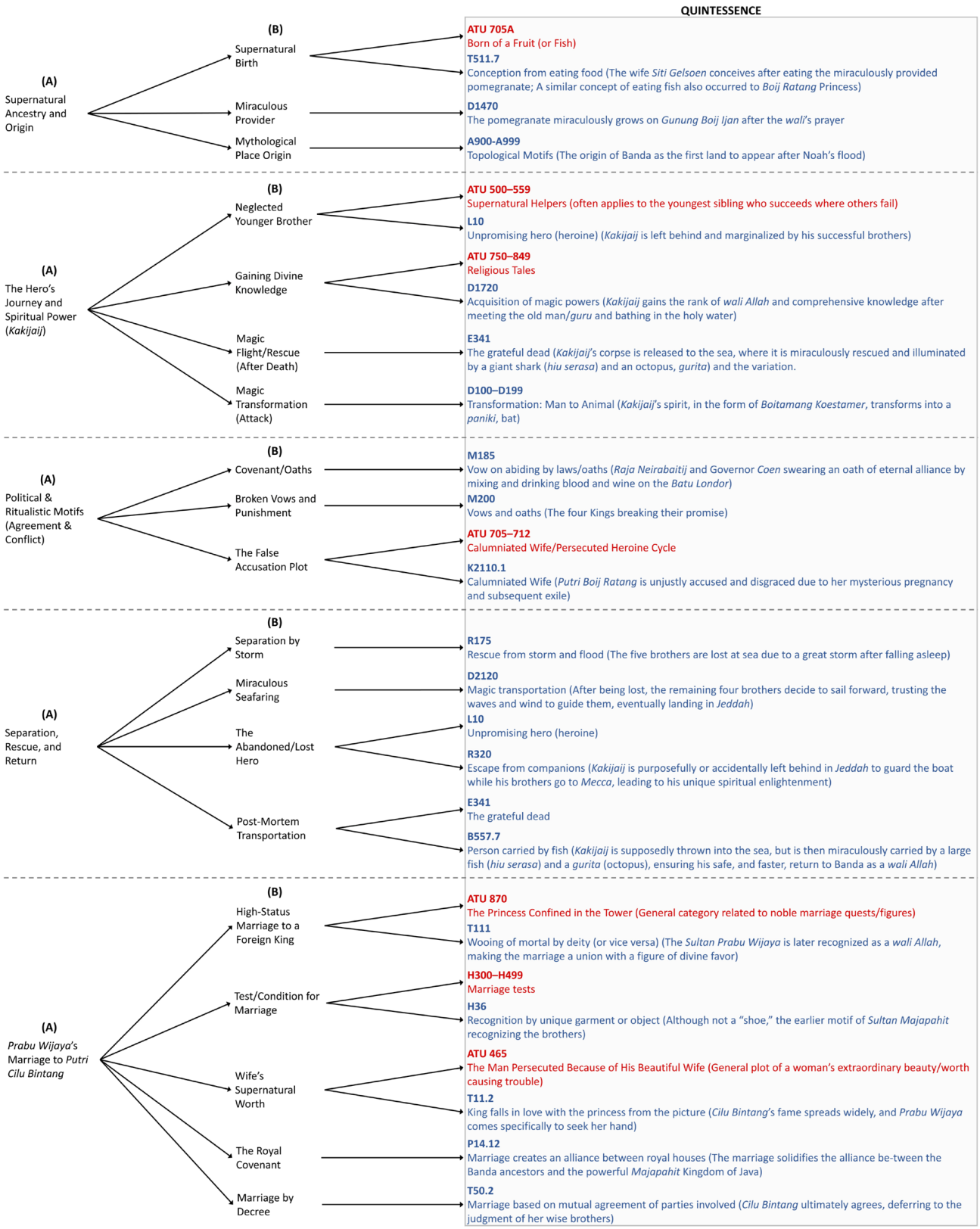


Figure 12. Motif analysis: (A) identified motif, (B) story elements, according to the Aarne-Thompson-Uther (ATU) classification (red-colored) and Stith Thompson's Motif-Index (blue-colored).

The folklore motif analysis is interpreted as a preliminary, single-analyst exploratory exercise rather than a formal coding result. No inter-coder reliability assessment was conducted, since the narrative features are better explained through Indigenous Maluku epistemologies, cosmologies, and oral-formulaic conventions rather than universalist or Euro-

American folkloristic frameworks. To reduce the bias caused by the absence of multiple annotators, only motifs with strong structural or functional correspondence were included. Future work will incorporate multiple coders, clearer inclusion criteria, and engagement with Bandanese cultural experts to ensure that motif classification respects local narrative logic and avoids over-reliance on external typologies.

3.6. Limitations

This study has several limitations. It is based on a single 1922 manuscript authored by an elite Bandanese figure, and therefore reflects only one social perspective rather than the full diversity of Banda oral traditions. Dictionary matching and morphological validation are also affected by more than a century of language change, as many historically valid forms in the manuscript do not appear in contemporary Indonesian or Malay lexicons. In addition, the manuscript's preservation within Dutch colonial archives highlights the cultural politics of archival provenance, since colonial collecting practices determined which Bandanese voices were recorded and which were excluded. Finally, the study was conducted without consultation from contemporary Bandanese community members, meaning that interpretive insights are not grounded in local perspectives; future work should incorporate community engagement to address this gap.

4. Conclusions

In the last decade, computational linguistics has experienced stagnant progress in historical linguistics, exacerbated by a conservative point of view of empirical research. At the same time, this becomes an opportunity to develop novel contributions. *Hikayat Lonthoir*, a rare collection from the National Maritime Museum of Amsterdam, retains the heritage significance of the Indigenous oral history of the Banda Archipelago, one of the underrepresented communities amidst the colonial narrative. By examining this manuscript, this research highlights its importance by incorporating corpus linguistics with automation, allowing big data processing for an empirical study. Deconstruction of its linguistic features shows lexical and morphological richness in micro-context, demonstrating that the lexicon is dominated by themes of monarchy and governance (*Raja*), social hierarchy (*saudara*, *tuan*, *bala*), and maritime setting (*tanah*, *laut*, *Banda*).

In morphological terms, action-oriented prefixes like *me-* (Active Voice) and *di-* (Passive Voice) show a densely verbal and agentive plot. The manuscript also retains the *Van Ophuijsen* old spelling (oe, dj, tj, j) and is influenced by Arabic and Dutch loanwords, explaining tokens flagged as proper names or domain-specific terminology. Semantic clustering indicates that the narrative is structured by maritime/politico-social themes and Place/Geographic anchors, while religious lexicon appears as concentrated spikes marking oaths, blessings, and legitimizing junctures. Network analysis quantified the centrality of key individuals, with Person entities (LIVB) holding the highest degree centrality, forming a Genealogical/Local cluster, while external figures like Tuan Coen acted as peripheral nodes linking the local narrative to conflict and power (Agreement).

Comparative analysis validates the saga to global folkloric systems in the macro-context, highlighting universal motifs such as Supernatural Ancestry and Origin (ATU 705A), the Hero's Journey (ATU 500–559; D1720), and the Persecuted Heroine Cycle (ATU 705–712). These motifs reinforce the narrative as a legitimizing cultural text that roots the Bandanese lineage in divine events and establishes political order through myth. This multidisciplinary approach generates new insights into the underappreciated narrative of the Banda Islands, confirming its value for historical linguistics, cultural heritage, and the global discourse on folklore.

Author Contributions: Conceptualization, M.I.K. and F.P.C.; methodology, writing—original draft preparation, software, M.I.K.; formal analysis, investigation, S.W.; resources, data curation, B.S.; writing—review and editing, F.P.C.; validation, visualization, K.P.W.; project administration, funding acquisition, M.A.; supervision, writing—review and editing, E.A.B.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Culture of the Republic of Indonesia, under contract 472/F5/KB.18.05/PPK II/2024.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the University Research Ethics Review Panel of Xi'an Jiaotong-Liverpool University (protocol code: ER-LRR-110001024202312021600001; approval Date: 1 December 2023).

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated or analyzed during the study are available at https://doi.org/10.6084/m9.figshare.30444374 (accessed on 16 November 2025).

Acknowledgments: We greatly appreciate the support of the community representatives in Maluku and North Maluku Province, the Ministry of Culture of the Republic of Indonesia, the Language Office of Maluku Province, the Department of Applied Linguistics, Xi'an Jiaotong-Liverpool University, and other involved parties who helped us with this research.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NLP	Natural Language Processing
OCR	Optical Character Recognition
KBBI	*Kamus Besar Bahasa Indonesia* (Great Dictionary of the Indonesian Language)
PRPM	*Pusat Rujukan Persuratan Melayu* (Centre for Malay Language Reference)
API	Application Programming Interface
NER	Named Entity Recognition
VOC	*Vereenigde Oost-Indische Compagnie* (Dutch East India Company)
PCA	Principal Component Analysis
LIVB	Entity code for "Person"
GEOG	Entity code for "Place"
DEVI/TISS	Entity code for "Object"
CONC	Entity code for "Concept"
PGRP	Entity code for "Social/Genealogical"
PROC	Entity code for "Maritime"
ACTI	Entity code for "Event"
ATU	Aarne-Thompson-Uther

References

1. Fischer, O. What counts as evidence in historical linguistics? *Stud. Lang.* **2004**, *28*, 710–740. [CrossRef]
2. Penke, M.; Rosenbach, A. What counts as evidence in linguistics? An introduction. In *What Counts As Evidence in Linguistics*; Penke, M., Rosenbach, A., Eds.; John Benjamins: Amsterdam, The Netherlands, 2007; pp. 1–49. [CrossRef]
3. McGillivray, B.; Jenset, G.B. Quantifying the quantitative (re-)turn in historical linguistics. *Humanit. Soc. Sci. Commun.* **2023**, *10*, 37. [CrossRef]
4. Piotrowski, M. NLP and digital humanities. In *Natural Language Processing for Historical Texts: Synthesis Lectures on Human Language Technologies*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 23–52. [CrossRef]
5. Conathan, L. Archiving and language documentation. In *The Cambridge Handbook of Endangered Languages*; Austin, P.K., Sallabank, J., Eds.; Cambridge University Press: Cambridge, UK, 2011; pp. 235–254.
6. Pessanha, F.; Salah, A.A. A computational look at oral history archives. *ACM J. Comput. Cult. Herit.* **2022**, *15*, 1–16. [CrossRef]

7. Nepal, A.; Perono Cacciafoco, F. Minoan cryptanalysis: Computational approaches to deciphering Linear A and assessing its connections with language families from the Mediterranean and the Black Sea areas. *Information* **2024**, *15*, 73. [CrossRef]
8. van Donkersgoed, J. Shifting the historical narrative of the Banda Islands: From colonial violence to local resilience. *Wacana J. Humanit. Indones.* **2023**, *24*, 500–514. [CrossRef]
9. Wildeman, D. Het Banda-manuscript. *Zeemagazijn* **2021**, *48*, 32–35.
10. van Donkersgoed, J.; Farid, M. Belang and Kabata Banda: The significance of nature in the adat practices in the Banda Islands. *Wacana J. Humanit. Indones.* **2022**, *23*, 415–450. [CrossRef]
11. Wu, S.; Perono Cacciafoco, F. Healing plants from Alor Island: A data paper for language documentation. *Ann. Univ. Craiova Ser. Filol. Lingvistica* **2025**, *46*, 413–435. [CrossRef]
12. Kersapati, M.I. Spatio-historical data enrichment for toponomastics in Bali, The Island of Gods. *GeoJournal* **2023**, *88*, 5489–5510. [CrossRef]
13. Al Azkaf, A.Z.; Yannuar, N.; Basthomi, Y.; Febrianti, Y. Connecting texts and thoughts: How translanguaging and multilingual writings reflect hybrid identities in colonial times. In *Applied Linguistics in the Indonesian Context: Engaging Indonesia*; Stroupe, R., Roosman, L., Eds.; Springer: Berlin/Heidelberg, Germany, 2025. [CrossRef]
14. Chai, C.P. Comparison of text preprocessing methods. *Nat. Lang. Eng.* **2023**, *29*, 509–553. [CrossRef]
15. Badudu, J.S.; Lesmanesya; Lubis, L.; Muchtar; Wijayakusumah, H. *Morfologi Bahasa Indonesia (Lisan)*; Pusat Pembinaan dan Pengembangan Bahasa: East Jakarta, Indonesia, 1984. Available online: http://repositori.kemendikdasmen.go.id/id/eprint/3155 (accessed on 21 October 2025).
16. Mi, C. Loanword identification based on web resources: A case study on Wikipedia. *Comput. Speech Lang.* **2023**, *81*, 101517. [CrossRef]
17. Sabharwal, N.; Agrawal, A. Introduction to word embeddings. In *Hands-on Question Answering Systems with BERT*; Apress: New York, NY, USA, 2021; pp. 35–58. [CrossRef]
18. Ayyadevara, V.K. Word2vec. In *Pro Machine Learning Algorithms*; Apress: New York, NY, USA, 2018; pp. 147–168. [CrossRef]
19. Mars, M. From Word Embeddings to Pre-Trained Language Models: A State-of-the-Art Walkthrough. *Appl. Sci.* **2022**, *12*, 8805. [CrossRef]
20. Song, Y.; Kim, H. Semi-automatic construction of a named entity dictionary based on active learning. In *Computer Science and Its Applications*; Park, J., Stojmenovic, I., Jeong, H., Yi, G., Eds.; Springer: Berlin/Heidelberg, Germany, 2015; Volume 330, pp. 109–120. [CrossRef]
21. Ehrmann, M.; Hamdi, A.; Pontes, E.L.; Romanello, M.; Doucet, A. Named entity recognition and classification on historical documents: A survey. *ACM Comput. Surv.* **2023**, *56*, 1–47. [CrossRef]
22. Walter, S.R.; Dunsmuir, W.T.; Westbrook, J.I. Inter-observer agreement and reliability assessment for observational studies of clinical work. *J. Biomed. Inform.* **2019**, *100*, 103317. [CrossRef]
23. Dharmawan, A.; Masithoh, R.E.; Amanah, H.Z. Development of PCA-MLP Model Based on Visible and Shortwave Near Infrared Spectroscopy for Authenticating Arabica Coffee Origins. *Foods* **2023**, *12*, 2112. [CrossRef]
24. Ogasawara, Y.; Kon, M. Two clustering methods based on Ward's method and dendrograms with interval-valued dissimilarities for interval-valued data. *Int. J. Approx. Reason.* **2021**, *129*, 103–121. [CrossRef]
25. Baek, S.I.; Bae, S.H. The effect of social network centrality on knowledge sharing. *J. Serv. Sci. Res.* **2019**, *11*, 183–202. [CrossRef]
26. Rubaidi, R.; Wachidah, H.N.; Sham, F.M. Moloku Kie Raha and the Legacy of Cultural Islam: The Enduring Influence of Ternate and Tidore. *Islam. J. Studi Keislam.* **2025**, *19*, 373–404. [CrossRef]
27. Widodo, W.; Burhanudin, M.; Kumala, S.A.; Tobing, S.H.; Lestariningsih, A.D. Classical Javanese Manuscripts as Identity Memory that Speaks Cultural Diaspora. *Abjad J. Humanit. Educ.* **2023**, *1*, 102–112. [CrossRef]
28. Burhanuddin, J. Global Networks and Religious Dynamics: Reading the Hikayat Raja Pasai of Pre-Colonial Malay-Archipelago. *Stud. Islam.* **2025**, *32*, 347–372. [CrossRef]
29. McGregor, K.; Dragojlovic, A. Songs from Another Land: Decolonizing Memories of Colonialism and the Nutmeg Trade. *Mem. Stud.* **2024**, *17*, 599–612. [CrossRef]
30. Lukito, Y.N.; Alkadri, M.F.; Adiswari, N.; Romadona, M.R. Revisiting Fort Heritage in Ternate, North Maluku: Nostalgia, Tourism, and Community Engagement. *Built Herit.* **2025**, *9*, 12. [CrossRef]
31. Nguyen, Q.-D.; Phan, N.-M.; Krömer, P.; Le, D.-A. An Efficient Unsupervised Approach for OCR Error Correction of Vietnamese OCR Text. *IEEE Access* **2023**, *11*, 58406–58421. [CrossRef]
32. Zoph, B.; Yuret, D.; May, J.; Knight, K. Transfer Learning for Low-Resource Neural Machine Translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–4 November 2016; Association for Computational Linguistics: Austin, TX, USA, 2016; pp. 1568–1575. [CrossRef]
33. Mayhew, A. The Dragonslayer: Folktale Classification, Memetics, and Cataloguing. *Proc. Doc. Acad.* **2020**, *7*, 3. [CrossRef]

34. Goldberg, C. Some suggestions for future folktale indexes. In *Die Heutige Bedeutung Oraler Traditionen/The Present-Day Importance of Oral Traditions*; Heissig, W., Schott, R., Eds.; VS Verlag für Sozialwissenschaften: Wiesbaden, Germany, 1998; Volume 102, pp. 187–198. [CrossRef]
35. Declerck, T.; Lendvai, P. Linguistic and semantic representation of the Thompson's Motif-Index of Folk-Literature. In *Research and Advanced Technology for Digital Libraries: TPDL 2011*; Gradmann, S., Borri, F., Meghini, C., Schuldt, H., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; Volume 6966, pp. 251–264. [CrossRef]
36. Svelstad, P.E. Allying with Beasts: Rebellious Readings of the Animal as Bridegroom (ATU 425). *Humanities* **2025**, *14*, 51. [CrossRef]
37. University of Missouri Libraries. ATU-AT-Motif: Explanation of Pages. Research Guides. 2025. Available online: https://libraryguides.missouri.edu/c.php?g=1039894 (accessed on 16 October 2025).
38. Thompson, S. *Motif-Index of Folk-Literature: A Classification of Narrative Elements in Folktales, Ballads, Myths, Fables, Medieval Romances, Exempla, Fabliaux, Jest-Books, and Local Legends*; Indiana University Press: Bloomington, IN, USA, 1955. Available online: https://www.ruthenia.ru/folklore/thompson/ (accessed on 16 October 2025).
39. Fadhilla, I.; Alim, S. Prayer poems for reaching happiness during COVID-19 pandemic in Indonesia. In *Managing Disruption and Developing Resilience for a Better Southeast Asia*; Nawawi, Alami, A.N., Bautista, J., Aung-Thwin, M., Simandjuntak, D., Prasetyo, Y.E., Eds.; Springer: Berlin/Heidelberg, Germany, 2022; pp. 551–562. [CrossRef]
40. Millie, J. The Malay Hikayat Miʿrāj Nabi Muḥammad. The Prophet Muḥammad's nocturnal journey to Heaven and Hell. *Bijdr. Tot De Taal-Land-En Volkenkd./J. Humanit. Soc. Sci. Southeast Asia* **2015**, *171*, 587–589. [CrossRef]
41. Paranjape, J.; Walvekar, M.R. Reading rules across traditions: Paribhāṣā in the Abhidhānappadīpikāṭīkā and early Amarakoṣa commentaries. *Lang. Hist.* **2025**, 1–14. [CrossRef]
42. Nopriyasman, N.; Asnan, G.; Fauzi, A.; Hastuti, I.P.; Ritonga, A.H.; Kurniawan, V.; Mairiska, R. Reading indigenous signs: The wisdom of nagari communities toward natural disaster in Pasaman Barat. *Int. J. Disaster Risk Reduct.* **2024**, *107*, 104497. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.