Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 18 June 2026
Revised 27 July 2026
Accepted 06 August 2026

Natural Resources for Human Health 2026, Vol. 6 (6): 286–300

KEYWORDS:
Journey to the West; Odyssey; Comparative literature; Computational text analysis; Personal adversity; Principal component analysis

Adversity and Response in Journey to the West and the Odyssey: A Computational Comparative Analysis

Xinyan Lu

Nanjing Foreign Language School, Nanjing, 210022, Jiangsu, China
Correspondence Email: luxy0806@outlook.com
https://orcid.org/0009-0004-9656-3066

ABSTRACT: How do classical literary works represent personal adversity and responses to it? This study compares Journey to the West with the Odyssey through an interpretable computational framework that connects thematic measurement with literary interpretation. The Chinese text of Journey to the West and the English translation of the Odyssey are segmented within chapter boundaries into 910 and 625 text blocks, respectively. A bilingual dictionary organized around challenges, responses, goals, and values converts textual expressions into measurable features. Coverage rates and mean match counts describe their prevalence, while joint principal component analysis (PCA), following logarithmic transformation, inverse-coverage weighting, and normalization, examines variation in expression profiles. The results reveal both shared patterns and differentiated emphases. Violence and coercion occur in approximately half of the blocks in each work. Journey to the West shows broader coverage of task commitment, rescue and care, and compassion, whereas the Odyssey gives greater prominence to homecoming and courageous endurance. PCA projections display substantial overlap between the works alongside differences in their average profiles, suggesting that distinct narrative emphases coexist with shared expressive resources. The first three components explain 35.4% of feature variance, indicating that low-dimensional visualizations capture only part of the texts' heterogeneous structures. The study provides a transparent, reproducible approach to cross-cultural literary comparison grounded in explicit bilingual retrieval rules. Its findings concern dictionary-detected expressions in two specific textual versions and should be interpreted through narrative context rather than treated as direct measurements of character virtue or general differences between Eastern and Western cultures.

1 INTRODUCTION

The spiritual strength displayed by the protagonist in a literary work is often closely tied to the tribulations they endure. Tribulations serve to drive the progression of the plot, bringing the character's inner desires, responsibilities, and behavioral choices into focus, and making them observable and open to discussion. When characters confront danger, temptation, and fierce conflict, how they perceive their plight, what actions they take, and what consequences they suffer often serve as crucial metrics for evaluating the artistic quality and ideological value of a literary work.Between Eastern and Western literature, this phenomenon is worthy of attention: different plot configurations and protagonists' responses to adversity reflect philosophical thoughts, spiritual pursuits, and shared values of East and West. As seen in classical works such as Notre-Dame de Paris and War and Peace, a protagonist's choices amid the turbulence of fate bridge textual narrative and philosophical concerns.

As monumental classics in Eastern and Western literary traditions, Journey to the West and the Odyssey hold irreplaceable positions within their respective cultural landscapes. Journey to the West, consisting of around 860,000 words and written

during China's Ming Dynasty (1368–1644), is the first romanticist chaptered novel of gods and demons in Chinese literature. It narrates the journey of a four-member pilgrimage led by the monk Xuanzang, who brave numerous hardships on their trek west to retrieve Buddhist scriptures. Combining traditional Chinese Buddhist and Taoist philosophies with political satire and metaphor, the prolonged journey intertwines external adventure with inner spiritual pursuit. The Odyssey, along with the Iliad, constitutes the ancient Greek Homeric epics (c. 8th–7th century BCE). Its main narrative recounts the ten-year wandering and tribulations of Odysseus following the Trojan War as he struggles to return home after angering the sea god Poseidon. Formed during the transition from primitive communal society to slavery in ancient Greece, the epic vividly reflects the social landscape, family ethics, and pursuit of personal heroism and rational intelligence of that era. Numerous interpretations of both works emphasize the deep connection between the protagonists' actions, their social relationships, and their underlying value pursuits. Building upon this foundation, this paper poses the following core research questions: How do Journey to the West and the Odyssey differently portray individual tribulations and coping mechanisms? How do these differences help us understand the distinct qualities and value orientations of East and West reflected in these two works?

To address these questions, this paper adopts an interpretable computational linguistic approach to analyze the texts of Journey to the West and the Odyssey. First, the digital texts of both works were cleaned and segmented into hundreds of semantically coherent text blocks based on chapter structures and internal plot turns. Next, a multi-tiered dictionary consisting of main categories, subcategories, and keyword rules was constructed to convert abstract research questions into quantifiable indicators of textual expression. Building on this, the paper compares the coverage and distribution of various themes and expressions across both works and employs PCA to examine the underlying structures of co-varying expressions. While statistical charts present overall distributional patterns, PCA projections and loadings provide critical clues for multi-thematic comparison. This allows for a deeper context-aware discussion of the connections between tribulations, actions, and values in each work. Finally, the paper discusses the convergences and divergences between these computational linguistic findings and classical literary interpretations.

In the text preprocessing stage, the original chapter and book structures of both works were segmented into finer-grained text block units based on two core considerations. First, the original number of chapters in both texts is relatively small, which would render statistical inference based solely on original chapters susceptible to small-sample bias. Second, the word count per chapter varies significantly across both works, and individual chapters often compress a multitude of distinct narrative features. Consequently, analyses reliant on raw chapter divisions operate on text units of non-comparable lengths, which complicates the clean separation of textual features and compromises statistical credibility. Specifically, taking the original chapters as natural hard boundaries, adjacent paragraphs were concatenated within each chapter to maintain a target chunk length of approximately 600 Chinese characters or 200 English words. This process yielded 910 text blocks for Journey to the West and 625 text blocks for the Odyssey. These text blocks preserve coherent narrative context while providing a substantially larger sample size to facilitate robust statistical inference.

At the measurement stage, a three-tiered bilingual dictionary was constructed to establish a transparent set of expression indicators. High-level categories delineate core research concepts, while subcategories correspond to the feature columns in the subsequent analysis matrix. For each subcategory, a retrieval set consisting of Chinese keywords and English regular expressions was populated based on close text reading and retrieval. The search and matching procedure recorded the positions and surrounding contexts of all matches while resolving multiple hits within the same subcategory, ultimately producing a text-block-by-subcategory count matrix. Preliminary findings from frequency distributions and histograms reveal that Journey to the West and the Odyssey exhibit similar frequencies in most of the categories, whereas a substantial disparity is observed in the rest, with both works displaying notably high frequencies in the Violence and Coercion category.

To further examine how different themes co-vary within both texts, the raw count matrix was subjected to log transformation, coverage weighting, and row-wise normalization. Text blocks containing non-zero matches from both works were merged to conduct PCA. The results indicate that the first three principal components account for approximately 40% of the total variation. By projecting the PCA results into two-dimensional and three-dimensional spaces, I observed the degree of similarity among text blocks and identified the thematic central tendencies of both works. Key findings from this spatial projection highlight several compelling clusters, demonstrating whether and to what extent the two works concentrate within distinct thematic clusters.

This paper makes two major contributions to the literature. First, it proposes an interpretable and easily accessible computational framework, combining word frequency statistics with PCA as an unsupervised learning pipeline, applied directly to the thematic analysis of literary texts. This approach bridges quantitative text analysis and qualitative literary criticism, offering a scalable methodology for examining personal tribulations and individual character traits across comparative literary traditions. Second, our empirical findings corroborate existing scholarly insights regarding the dominant presence of violent tribulations and physical crises in both works. Furthermore, the identification of novel thematic clusters (e.g., the co-occurrence of violent adversity, emotional distress, and character fortitude) offers fresh analytical perspectives for the comparative study of Eastern and Western literary classics.

The rest of this paper is organized as follows. Section 2 reviews the related literature on comparative analysis, tribulations in classical literature, and computational text processing. Section 3 outlines the methodological pipeline, detailing text preprocessing, bilingual dictionary construction, and dimension reduction techniques. Section 4 presents the main empirical findings alongside thematic comparisons. Section 5 concludes the paper with a discussion of implications and potential avenues for future research.

2 LITERATURE

This study draws on three strands of literature: religious allegory in Journey to the West, heroic agency and narrative structure in the Odyssey, and computational approaches to literary analysis.

The literature on the religious dimensions of Journey to the West provides a foundation for understanding the relationship between physical journeys and spiritual transformation. (Yu 1983) situates the novel alongside Dante's Commedia within a comparative framework of religious pilgrimage. (Bantly 1989) examines its Buddhist allegory while highlighting the interpretive challenges posed by the interplay of Buddhist, Taoist, and Confucian elements with humor and irony. This perspective cautions against reducing the novel's narrative complexity to a single doctrinal framework. (Lam 2019) further explores the relationship between allegory and world-formation, showing that correspondences between bodily landscapes and cosmic order are neither consistent nor self-contained. Together, these studies illuminate the religious meanings of adventure while preserving tensions among competing interpretations. Their emphasis on contextual interpretation, however, leaves room for a complementary examination of how relevant expressions are distributed across the full text and how those distributions differ between works. Building on these insights, I develop traceable dictionary-based indicators of adversity and responses to it. These indicators allow us to examine the textual prevalence of expressions associated with literary interpretations, without treating their frequency as a direct test of those interpretations.

A second strand of literature examines heroic agency, social relationships, and narrative form in the Odyssey. (Pache 2020) connects Odysseus's cunning intelligence (metis) with his homecoming and the restoration of household order and legitimate rule. (Griffin 2003) emphasizes how the epic subjects heroism to scrutiny through varied adventures and social encounters. Approaching the poem through its ethnographic imagination, (Dougherty 2001) relates Odysseus's resourcefulness to the experiences of ancient Greek audiences adapting to uncertain conditions. This account places heroic action within a broader social world rather than attributing it solely to individual qualities. From a narratological perspective, (Jong 2001) examines narration, characterization, focalization, and type-scenes, demonstrating how narrative techniques shape the presentation of characters and their actions. These studies offer detailed accounts of the relationships among strategy, social order, and storytelling. Our study complements them by comparing the Odyssey and Journey to the West through a shared set of bilingual expression indicators. A three-tiered dictionary and joint principal component analysis (PCA) provide an overview of cross-text patterns, while close reading remains necessary to interpret their narrative significance.

A third strand of literature uses computational methods to investigate literary narratives and characters. (Reagan et al. 2016) identifies recurring emotional arcs through sentiment analysis, while (Bamman et al. 2014) develops a statistical model of literary character types. These approaches demonstrate how computational analysis can reveal patterns across texts that are difficult to establish through close reading alone. Nevertheless, emotional trajectories and statistical groupings do not directly establish how works represent adversity, evaluate conduct, or endorse particular qualities. The methodological discussion in (Grimmer and Stewart 2013) likewise stresses the need to validate automated textual measures against the specific questions they are intended to address. Our study responds to this measurement challenge by combining an explicit bilingual dictionary with PCA: the dictionary makes retrieval rules inspectable, and PCA summarizes variation in the resulting expression profiles. Linking

these features to their source passages provides a reproducible basis for moving between quantitative patterns and literary interpretation, while maintaining a distinction between detected expressions, character qualities, and narrative judgments.

3 METHODOLOGY

3.1 Data Sources and Corpus Preparation

This study employs a comparative case study design to examine the thematic structure of tribulations and protagonist responses in Journey to the West and Homer's Odyssey. The primary textual corpus consists of the classical Chinese source text of Journey to the West (comprising 100 chapters and approximately 860,000 characters) and the Samuel Butcher and Andrew Lang English prose translation of the Odyssey (comprising 24 books and approximately 300,000 words). For the Odyssey, the introductory preamble preceding Book 1 was excluded from analysis to maintain narrative focus, while poetic passages, songs, and quoted speeches in both works were retained.

3.2 Text Block Segmentation

The original structural divisions of both texts present two methodological challenges for quantitative analysis: individual chapters and books vary substantially in word count, and the small number of original divisions in the Odyssey (24 books) severely limits statistical power. To address these limitations, I implemented a text-segmentation algorithm that constructs a large sample of units of comparable length while respecting original chapter and book boundaries as hard constraints. As a baseline specification, target chunk sizes were set to approximately 600 Chinese characters for Journey to the West and 200 English words for the Odyssey (both inclusive of punctuation).

The segmentation procedure prioritizes preserving natural paragraph integrity without allowing text blocks to cross chapter or book boundaries. The algorithm follows a two-step process involving the splitting of excessively long paragraphs and the concatenation of shorter text fragments.

First, paragraph integrity is preserved whenever possible. A paragraph is split only if adding it causes the cumulative length of the current text block to exceed 1.5 times the target threshold (900 characters for Chinese; 300 words for English). In such cases, the algorithm searches for the last sentence-ending punctuation mark within the first 600 characters (or 200 English words). If splitting at this mark produces a text block that is at least 55% of the target length (330 characters or 110 words), the split occurs at the punctuation mark. Otherwise, a strict hard split is enforced immediately following the 600th character or 200th word, and the remainder is processed recursively under the same rules.

Meanwhile, following paragraph splitting, shorter paragraphs and fragments are merged sequentially in their original narrative order. A new text block is exported once the cumulative length reaches the target threshold (600 characters / 200 words). If the current block remains below target length but incorporating the subsequent fragment would exceed 1.5 times the target threshold, the accumulated text is exported as a standalone block, and a new block is initiated from the next fragment.

To ensure clear narrative differentiation, any remaining text at the end of a chapter or book that falls short of the target length is preserved as an independent text block rather than merged across chapter boundaries. This segmentation procedure produced approximately 910 text blocks for Journey to the West and 625 text blocks for the Odyssey. Although these text blocks do not strictly correspond to complete, self-contained plot episodes, they serve as fine-grained, semantically coherent narrative units. The empirical distributions of text block lengths (visualized via histograms and kernel density plots) are presented in Section 4.

3.3 Dictionary Construction and Rule Matching

To measure thematic concepts, I constructed a three-tiered bilingual dictionary structured into main categories, subcategories, and keyword rules. Main categories define overarching conceptual domains (e.g., tribulations, actions, virtues), subcategories specify distinct attributes within those domains (corresponding to feature columns in our empirical matrix), and keyword rules define text-matching patterns.

For the English text, keyword rules utilize regular expressions to account for morphological variations, tense inflections, and synonymous expressions. To prevent inverted semantic meanings from contaminating affirmative metrics, explicit negation handling was incorporated: Chinese matching checks for immediate pre-target negation terms, while English matching evaluates negation markers within a four-word context window preceding the target match.

Pattern matching was conducted on the raw paragraph level. A single paragraph was permitted to trigger matches across multiple distinct subcategories. However, multiple keyword hits falling under the same subcategory within a single paragraph were aggregated and recorded as a single occurrence. Finally, all paragraph-level occurrences were summed at the text block level to construct the primary text-block-by-subcategory count matrix.

3.4 Coverage Rate and Frequency Metrics

Let c_{ij} denote the count of matches for subclass j in text block i, recorded after deduplication within each paragraph and local negation filtering. The full set of counts yields an $N \times J$ non-negative occurrence matrix $\mathbf{C}$, where N is the total number of text blocks across both texts and J is the total number of subcategories in the dictionary.

Let $\mathcal{I}_b$ represent the set of text blocks belonging to work $b \in \{\text{Journey to the West,Odyssey}\}$, with $N_b = |\mathcal{I}_b|$ denoting the number of blocks in work b. The coverage rate p_{bj} of subcategory j in work b is defined as:

$$p_{bj} = \frac{1}{N_b} \sum_{i \in \mathcal{I}_b} \mathbb{I}(c_{ij} > 0)$$

where $\mathbb{I}(\cdot)$ is an indicator function that equals 1 if $c_{ij} > 0$ and 0 otherwise. Note that text blocks with zero matches ($c_{ij} = 0$) are retained in the denominator N_b. The coverage rate p_{bj} quantifies the breadth of thematic presence across the narrative. The distributions of these coverage rates are visually presented in Section 4 using histograms and kernel density estimations.

3.5 Rescaling, Inverse-Coverage Weighting, and Normalization

To reduce the disproportionate influence of high-frequency baseline terms and emphasize distinctive thematic expressions prior to dimension reduction, I apply an inverse-coverage weighting transformation. We compute the unweighted average coverage rate $\bar{p}_j$ for each subclass j across both works:

$$\begin{equation} \bar{p}_j = \frac{1}{2} \sum_{b \in \{\text{\textit{JW}}, \text{\textit{O}}\}} p_{bj} \end{equation}$$

Based on $\bar{p}_j$, the inverse-coverage weight w_j for subclass j is calculated as:

$$w_j = 1 + \ln\left(\frac{1}{\bar{p}_j + 0.01}\right)$$

where 0.01 serves as a smoothing constant. Averaging coverage rates across works ($\bar{p}_j$) rather than applying work-specific weights prevents the work with a larger text-block sample size from dominating the weighting scheme.

Applying w_j alongside a logarithmic transformation yields the adjusted thematic occurrence score x_{ij} for subclass j in text block i:

$$x_{ij} = \ln(1 + c_{ij}) \cdot w_j$$

where adding 1 handles zero occurrences ($\ln(1) = 0$). This transformation down-weights ubiquitous terms while amplifying rare, distinctive expressions that characterize the specific narrative identity of each work.

Next, to focus Principal Component Analysis on relative thematic composition rather than variation in absolute text-block word count, I apply L_2 normalization to each text block that contains at least one non-zero match:

$$z_{ij} = \frac{x_{ij}}{\sqrt{\sum_{l=1}^{J} x_{il}^2}}$$

Normalization projects each non-zero feature vector onto a unit sphere ($\| \mathbf{z}_i \|_2 = 1$). Text blocks with zero matches across all subcategories ($c_{ij} = 0, \forall j$) cannot be normalized and are excluded from the PCA matrix; however, they remain in descriptive summary statistics (the counts of zero-match text blocks for both works are reported in Section 4).

3.6 Joint Principal Component Analysis (PCA)

To project both texts into a shared coordinate space for comparative analysis, I perform a joint PCA on the merged matrix of normalized non-zero text blocks of the two fictions combined (i.e., the joint part). Let n denote the total number of retained non-zero text blocks, and let μ_j represent the sample mean of normalized feature j across all n blocks:

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} z_{ij}$$

The score of text block i on the k-th principal component t_{ik} is computed as:

$$t_{ik} = \sum_{j=1}^{J} v_{jk}\,(z_{ij} - \mu_j)$$

where v_{jk} represents the loading (coefficient) of subclass j on component k. The loading v_{jk} reflects both the directional impact and magnitude of subclass j in shaping the k-th component. A positive loading ($v_{jk} > 0$) indicates that a higher-than-average presence of subclass j increases the component score, whereas a negative loading reduces it. Subclasses with higher absolute loading magnitudes $|v_{jk}|$ exert greater influence in defining the thematic axis, highlighting primary thematic dimensions within the comparative corpus.

For instance, let the first principal component (PC1) be the following:

$$t_{i1} = 0.60 \cdot (z_{i,\text{violence}} - \mu_{\text{violence}}) - 0.50 \cdot (z_{i,\text{hospitality}} - \mu_{\text{hospitality}}) - 0.40 \cdot (z_{i,\text{responsibility}} - \mu_{\text{responsibility}}) + \cdots$$

In this example, a higher value for violence-related expressions increases a text block's PC1 score, whereas higher values for rescue- or responsibility-related expressions decrease it. The axis therefore captures a contrast between these expression profiles. Estimated coefficients are obtained from all retained text blocks and define a common coordinate system for both works; differences in the distributions of their scores indicate differences in their expression profiles.

PCA is implemented using scikit-learn and visualized on selected pairs of principal components based on their explanatory power of the total variance. All views use scores and coefficients from the same jointly fitted PCA model. Each point represents a text block with at least one non-zero dictionary feature and each retained block receives equal weight. Consequently, averaging the two works' coverage rates when constructing feature weights does not give the two works equal influence on the PCA fit when their numbers of retained blocks differ.

To summarize each work's average expression profile, we calculate its mean score on each principal component. For a plane defined by components a and b, the centroids are $(\bar{t}_{Xa}, \bar{t}_{Xb})$ and $(\bar{t}_{Oa}, \bar{t}_{Ob})$, where X denotes Journey to the West and O denotes the Odyssey. Their differences are defined consistently as the Odyssey minus Journey to the West:

$$\Delta_a = \bar{t}_{Oa} - \bar{t}_{Xa}, \qquad \Delta_b = \bar{t}_{Ob} - \bar{t}_{Xb}.$$

To relate this difference to specific expression indicators, we project the centroid displacement back into the transformed feature space using the two selected component directions:

$$d_j^{(a,b)} = v_{ja}\Delta_a + v_{jb}\Delta_b,$$

where v_{ja} and v_{jb} are the coefficients of feature j on components a and b, respectively. The quantity $d_j^{(a,b)}$ is the reconstructed mean difference for feature j retained by the selected two-dimensional subspace. A positive value indicates a higher reconstructed value for the Odyssey, whereas a negative value indicates a higher reconstructed value for Journey to the West. Larger absolute values identify features that are more prominent in the centroid difference represented in that plane. Different component pairs may therefore highlight different features because they retain different parts of the overall difference between the works.

Natural Resources for Human Health
eISSN: 2583-1194

The direction of feature j's arrow in the selected plane is determined by (v_{ja}, v_{jb}). For example, a feature with a positive PC1 coefficient and a negative PC3 (the third principal component) coefficient points toward the lower-right quadrant in a PC1–PC3 plot. Its displayed endpoint is

$$\mathbf{e}_j^{(a,b)} = s\begin{pmatrix} v_{ja} \\ v_{jb} \end{pmatrix},$$

where $s > 0$ is a common display-scaling factor for all arrows within the figure. This scaling preserves their relative directions and lengths. In three dimensions, the corresponding direction is given by (v_{j1}, v_{j2}, v_{j3}). It is worth noting that these arrows describe the directions of the transformed expression indicators; their lengths do not measure raw frequency or literary importance.

4 RESULTS

4.1 Empirical Distribution of Text Block Lengths

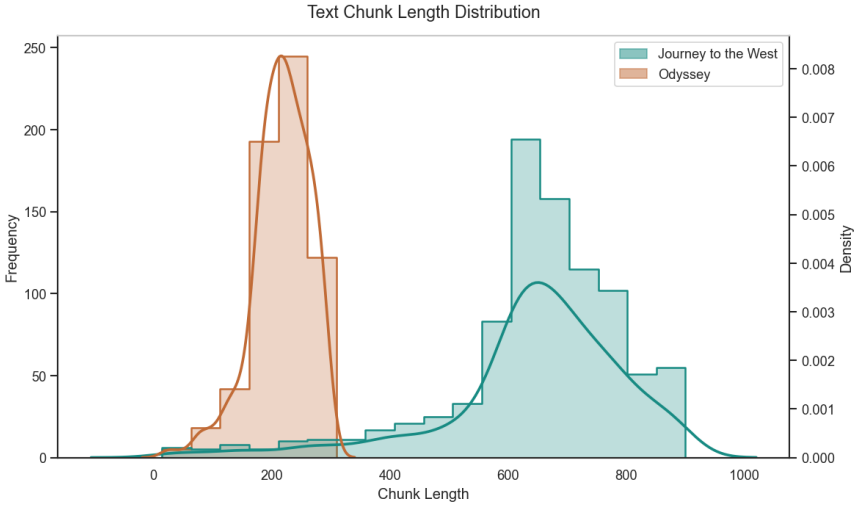


Figure of histogram on the word counts of text blocks.

The text segmentation procedure yielded $N_{\text{\textit{O}}} = 625$ text blocks for the Odyssey and $N_{\text{\textit{JW}}} = 910$ text blocks for Journey to the West, providing a substantial sample size to ensure statistical power for subsequent PCA.

Figure 1 illustrates the empirical length distributions and kernel density estimates for the segmented text blocks across both works. As shown in the plot, the chunk lengths closely align with our a priori baseline specifications. For the Odyssey, the text block lengths strongly concentrate around the target threshold of 200 English words, with a sharp upper truncation at 300 words. For Journey to the West, block lengths primarily cluster around 600 Chinese characters, truncated strictly at 900 characters. These distribution patterns confirm that the segmentation algorithm successfully adhered to the upper limit constraint ($1.5 \times$ target length), generating standardized narrative units of comparable semantic scale across the two corpora.

4.2 Conceptual Measurement Framework and Dictionary Design

To operationally bridge qualitative literary interpretations and quantitative computational measurement, I constructed a structured, three-tiered bilingual dictionary. Table [tab:dictionary-classes] presents the consolidated dictionary schema, specifying the four macro Themes, nineteen analytical Classes, and their corresponding bilingual keyword patterns. The selection of themes, classification of subclasses, and extraction of representative keywords were meticulously informed by both established comparative literature scholarship and the empirical linguistic properties of the two corpora.

@lL2.5cmL6.4cmL5.5cm@

Natural Resources for Human Health
eISSN: 2583-1194

Theme & Class & Chinese keywords & English keywords / expressions
Theme & Class & Chinese keywords & English keywords / expressions

& Desire and Temptation & 贪心、贪欲、贪恋、色欲、贪色、贪财、贪吃、贪嘴、动了凡心、招赘、欲念、美色、爱上我、痴迷、心痒难挠、骨软筋麻、肚子、塞饱、暴食者、狼吞虎咽、战利品、饥餐、渴饮、花言巧语、痴心 & temptation; lust; greed; desire; pleasure; romantic infatuation; craving; appetite and gluttony; booty; seductive speech & Emotional Distress & 大怒、发怒、怒气、忧愁、忧虑、恐惧、惊恐、害怕、心慌、愁苦、悲伤、嗔怒、大惊失色 & wrath; anger; terror; fear; grief; anxiety; indignation; rage; alarm

& Deception and Misrecognition & 欺骗、哄骗、诓骗、受骗、假扮、疑心、真假、假冒、骗、疑惑 & deceive; deception; false tale; falsehood; disguise; doubt

& Separation and Loss of Belonging & 离乡、离家、流落、思乡、失散、生离死别、别离、离别、流浪 & far from home; longing for home; wandering; bereavement; exile

& Suffering and Hardship & 害人、吃人、百姓受、生灵涂炭、伤害百姓、童男童女、小儿性命、受害、灾难、受苦、受难、遭殃、苦难、困苦、苦楚 & devour companions; companions slain; suffering people; distress; suffer; affliction; misery; evil plight & Task and Command & & charge laid; command

& Restraint and Endurance & 忍耐、忍住、忍辱、忍气、按下怒气、强忍、耐心、克制 & restrain; endure; patient; hold or check anger; refrain; bear patiently & Strategy and Disguise & 计策、计谋、计赚、诡计、使诈、假扮、假冒、装作、计议、计较、变作 & stratagem; trick; cunning; crafty; disguise; Noman; craft; device; plan & Rescue and Care & 救出、搭救、解救、救人、救他、救师父、救师傅、保护、照顾、照料、供养、奉养、接济 & rescue; save; care for; protect; give food or clothing; hospitality; shelter; nurse; tend; help; aid; defend; deliver; succour; guard & Violence and Coercion & 杀死、打死、斩杀、杀了、大战、恶战、威胁、捆绑、绑缚、打斗、砍、捉、劈、行凶、烧杀、剥了皮、剥皮、绳缠、索绑、厮杀、唬杀、捆、屠杀、被擒、捉住、囚禁、性命难保、性命不保、遭难、丧命、受伤、被杀、锁、痛 & slay; kill; smite; fight; threaten; battle; capture; detain; bind; restrain; slaughter; imprison; wound; death; pain

& Reception and Hospitality & 款待、管待、留宿、相留、设宴、摆宴、斋饭、茶饭 & kindly entreat; welcome guests or strangers; give lodging or raiment; hospitality; entertain; provide food or drink; clasp hands in welcome & Seeking Help & 求救、救命、求助、相助、助我、助你、帮忙 & help me, us, him, or them; aid; pray to; take counsel; counsel together; assist

& Homecoming and Reunion & 归家、回家、还乡、回乡、返乡、团圆、团聚、故乡、还去、家去、回去、相认、标记 & homecoming; return home; return to one's country or family; homeward journey; reunion; recognition; marker & Task and Commitment & 取经、取真经、使命、奉旨、誓愿、誓言、不负、许诺、承诺、西天 & oath; swear; promise; fulfil; covenant; pledge; vow

& Seeking the Way and Cultivation & 求道、学道、修行、修心、悟道、寻仙、访仙、参禅、修持、修炼、历练、求知、漂泊 & spiritual discipline; cultivate or purify the mind; self-knowledge; self-control; training; quest for knowledge; wandering
& Honor and Renown & 功名、名声、扬名、威名、荣誉、名望、荣耀、闻名、光荣、尊崇、尊敬、正果、功成行满、功满行完 & glory; renown; fame; honor; reputation; title; prestige; respect; acclaim; merit and achievement & Revenge and Retribution & 报仇、复仇、雪恨、报恨、讨还、报复、仇报 & revenge; vengeance; avenge; requital; repay; retaliation; retribution; redress

& Character and Discipline & 心性、心猿、意马、定性、存神、清心、禅心、看破、戒律、持戒、守戒、斋戒、无欲、嫉妒、酒戒、紧箍咒、待客之道、不成文的习俗、宙斯法则 & self-control; self-restraint; discipline the mind; pure heart; jealousy; abstinence; temperance; hospitality customs; Zeus's law

& Compassion and Mercy & 慈悲、怜悯、慈心、恻隐、仁慈、不忍心、可怜、不杀、莫杀、休杀、戒杀、饶他性命、饶了他、不可伤生、休伤、收降 & compassion; mercy; pity; kindness; refrain from killing; spare life; forgive; non-harm; leniency

& Courage and Perseverance & 勇敢、勇气、英勇、坚忍、坚毅、坚心、忍苦、胆量、勇猛、雄豪、坚强、咬定牙关、果敢、大胆 & courage; bravery; much-enduring; steadfast; valiant; fearless; determined; resilient; persevere

4.2.1 Theoretical Foundations and Comparative Literature Grounding

In the dictionary, the high-level categorization into four macro themes—Challenge, Response, Goal, and Value—is directly anchored in classical narrative theories of hero journeys and comparative epic studies. In heroic epic and pilgrimage literature, narratives are fundamentally propelled by a dynamic sequence: characters encounter external or internal obstacles (Challenge), deploy cognitive, physical, or social agency to overcome them (Response), pursue teleological destinations or obligations (Goal), and embody implicit ethical or philosophical principles (Value). This four-part structure directly operationalizes comparative insights from scholars such as Anthony C. Yu (Yu 1983, 2012) and Jasper Griffin (Griffin 2003). In comparative pilgrimage literature, external tribulations (Challenge) constantly force protagonists to make behavioral choices (Response) that bridge physical geographic movement toward a destination (Goal) with spiritual cultivation or heroic self-realization (Value). By structuring the dictionary around these four pillars, I ensure that quantitative feature extraction maps directly onto the narrative arc of comparative epic traditions.

To capture the nuanced narrative dynamics of both texts, the four macro themes were refined into nineteen distinct analytical classes. This division was designed to reflect specific cultural, structural, and ethical dilemmas highlighted in comparative scholarship.

Following Bantly (Bantly 1989) and Dougherty (Dougherty 2001), challenges are divided into physical threats (Suffering and Hardship), internal psychological trials (Desire and Temptation, Emotional Distress), epistemological obstacles (Deception and Misrecognition), and existential crises (Separation and Loss of Belonging). Additionally, the Task and Command class accounts for divine or imperial mandates that initiate heroic journeys.

Drawing on narratological commentaries (Jong 2001), responses reflect the multi-faceted mechanisms of agency. They encompass overt physical confrontation (Violence and Coercion), strategic ingenuity (Strategy and Disguise, e.g., Odysseus's "Noman" ploy), moral/social engagement (Rescue and Care, Reception and Hospitality), self-restraint (Restraint and Endurance), and external reliance (Seeking Help, reflecting Tang Sanzang's dependence on divine intervention).

The goal classes capture the fundamental ideological divergence between Western epic teleology and Eastern pilgrimage pursuits. While the Odyssey centers on Homecoming and Reunion (nostos) (Pache 2020) and Revenge and Retribution, Journey to the West is driven by Task and Commitment (retrieving scriptures) and Seeking the Way and Cultivation (Lam 2019). Reflecting ethical codes of both traditions, these classes operationalize core virtues: Buddhist/Daoist mind-discipline and Greek codes of host-guest relations (Character and Discipline, including Zeus's Law and Tight-Fillet Spell), moral restraint (Compassion and Mercy), and heroic endurance (Courage and Perseverance).

4.2.2 Keyword Selection and Stylistic Adaptation

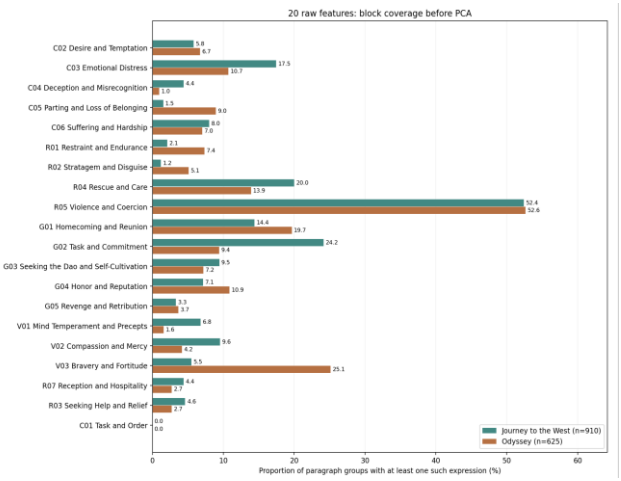
Keyword selection was governed by a rigorous iterative process combining close reading, domain expertise, and linguistic feature matching tailored to the distinct literary styles of Ming Dynasty vernacular Chinese and Victorian English epic prose.

First, stylistic and vernacular adaptation was essential. In Journey to the West, keywords were selected to capture Ming vernacular idioms, religious allegories, and recurring stock phrases (e.g., "心猿"、"意马" for mental discipline, "心痒难挠"、"骨软筋麻" for temptation, and "功成行满" for spiritual achievement). In the Butcher-Lang English translation of the Odyssey, keywords were extracted to reflect archaic, Homeric register and formulaic expressions (e.g., "seductive speech", "much-enduring", "kindly entreat", and "guest-friendship").

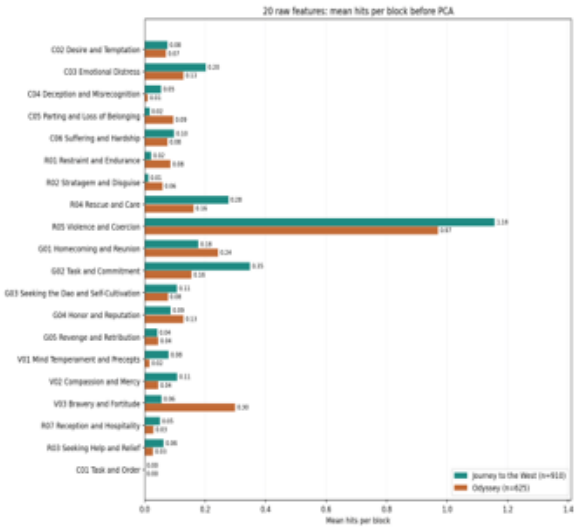
Second, regular expressions and morphological expansion were deployed for the English corpus to capture inflections, archaic conjugations, and syntactic variations without omitting relevant passages (e.g., matching slay, slain, slewest, slaying under Violence).

Third, context-aware negation filtering was applied to eliminate false positives. In classical Chinese narrative prose, moral keywords frequently appear in negated contexts (e.g., "莫杀"、"休伤" under Compassion and Mercy). The matching algorithm explicitly evaluated local pre-target negation markers in Chinese and a four-word preceding window in English to distinguish affirmative thematic presence from prohibited or negated acts.

4.3 Histogram of Class Frequency



Histogram of unweighted class frequency



Histogram of weighted class frequency

Histogram of Class Frequency

As illustrated in the distribution figures 4, the Violence and Coercion category exhibits a significantly higher occurrence frequency than all other thematic categories. In the unweighted raw counts, the proportion of text blocks containing violence-related keywords reaches approximately 50% in both Journey to the West and the Odyssey, indicating that at least half of the text blocks in both works depict violent confrontations or combat scenes. This quantitative pattern aligns closely with the narrative structures of both epics: in Journey to the West, the four pilgrims encounter eighty-one perilous tribulations, nearly

every chapter featuring life-threatening crises and physical violence; similarly, the narrative of the Odyssey centers on Odysseus's arduous homeward journey, during which he repeatedly relies on violent force to overcome obstacles and ensure survival.

Conversely, the Seeking the Way and Cultivation category covers a relatively small proportion of text blocks, accounting for only about 7% of the total corpus across both works. This low frequency does not imply an absence of spiritual themes, as both texts describe the protagonists' journey of enduring hardships, eliminating worldly distractions, and conquering personal desires. Rather, it reflects an implicit narrative style: neither work explicitly proclaims that the protagonist is undergoing "spiritual cultivation"; instead, this process of self-transcendence is subtly embedded within the descriptions of physical tribulations.

Furthermore, the figures highlight significant cross-cultural divergences between the two works. In the Courage and Perseverance category, the proportion of text blocks containing relevant keywords in the Odyssey is markedly higher than that in Journey to the West. This divergence stems from distinct characterization techniques: traditional Eastern narrative style in Journey to the West favors subtlety, portraying heroic character traits indirectly through actions, dialogue, and narrative details rather than explicit virtue terms; in contrast, Western epic tradition, as exemplified by the Odyssey, frequently utilizes direct adjectives and explicit vocabulary to shape and praise heroic character, resulting in a higher keyword match rate.

In addition, Appendix [Coverage of Themes Indexed by the Sequence of Text Blocks] provides another visualization of the coverage of themes indexed by the sequence of text blocks. It manifests that Violence and Coercion is almost uniformly distributed in the two fictions from the early to the late stages, while other topics are scattered.

4.4 Principal Component Variance Decomposition

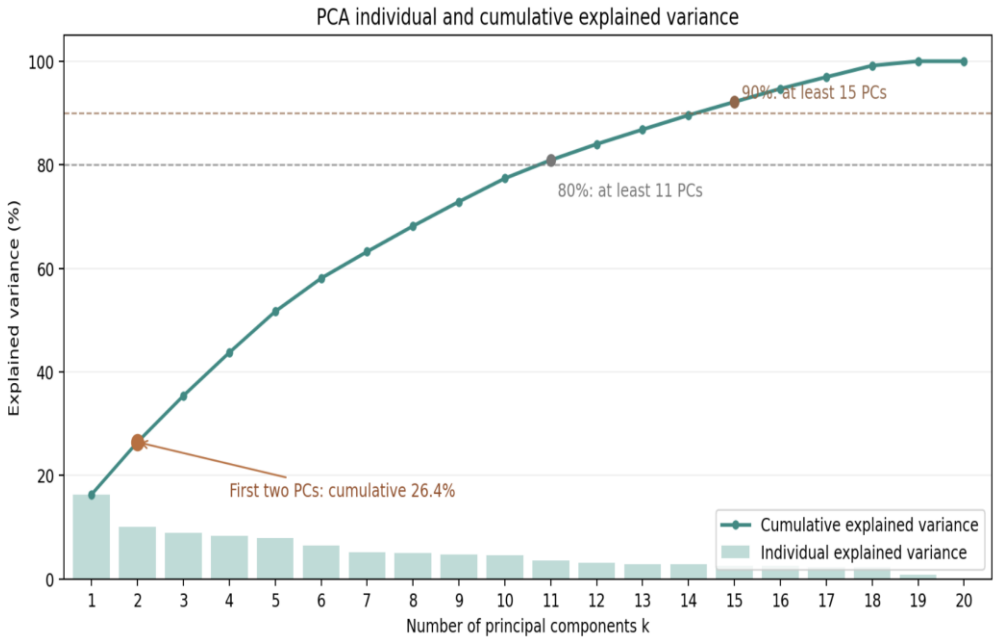


Figure of PCA variance decomposition.

Figure 5 illustrates the proportion of variance explained by the individual principal components. As shown in the plot, the first five principal components account for a relatively high share of the variance, with each individual component explaining no less than 10%. Apart from the first principal component, which exhibits significantly higher explanatory power than the rest, the marginal contribution of subsequent components begins to show a clear diminishing trend starting from the sixth principal component. Reaching a cumulative explained variance threshold of above 80% requires retaining at least eleven principal components.

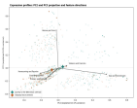
For our textual analysis, this pattern indicates a high degree of heterogeneity and diversity in the thematic combinations embedded across different text blocks, suggesting that the narrative structure of the full text cannot be reduced to merely a few

dominant components. Given the inherent complexity, thematic density, and narrative variability of classic long-form epics, this variance decomposition profile is both empirically reasonable and theoretically expected.

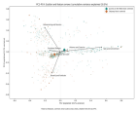
4.5 PCA Score and Loading Projections



Projection Plot of PC1 and PC2



Projection Plot of PC2 and PC3



Projection Plot of PC2 and PC4



Projection Plot of PC4 and PC5

Biplot Projections of Scores and Feature Loadings Across the First Five Principal Components

Figure 6 and Figure 7 illustrate the biplot projections of text block scores and subcategory feature loadings across the first three principal components. From an overall perspective, the centroids of text block distributions for both Journey to the West and the Odyssey cluster near the origin, indicating a substantial macro-level similarity in their narrative foundations and core plot structures.

As shown in Figure 6, the feature loadings for Violence and Coercion and Task and Commitment form distinctly divergent vector directions, with very few text blocks falling within the angle between these two arrows. This indicates that these two narrative themes rarely co-occur within the same individual text block. In contrast, feature vectors such as Desire and Temptation and Courage and Perseverance form a prominent vector cluster. Within this region, the density of green dots (Journey to the West) and orange dots (Odyssey) is virtually equal, demonstrating that both works frequently integrate these themes simultaneously within their narrative episodes. Furthermore, a substantial cluster of interwoven green and orange dots is distributed in the angle between Violence and Coercion and the four vectors pointing toward the bottom-left. This reflects a shared narrative mechanism across both epics: when protagonists confront violent adversity, they are simultaneously exposed to internal turmoil and temptations, which in turn forge and reveal their courage and endurance.

In Figure 7, the loading vectors for Rescue and Care and Task and Commitment likewise diverge significantly. Unlike in the previous biplot, the arrow representing Violence and Coercion is noticeably shortened due to the shift in projection axes. Text block projections concentrate primarily around the vectors for Homecoming and Reunion and several adjacent categories, confirming the prominent presence of distinct episodes corresponding to these themes in both texts. Notably, around the Homecoming and Reunion vector, orange scatter points (Odyssey) are markedly denser than green points (Journey to the West). This pattern captures a fundamental structural difference: the Odyssey is strictly centered on the teleological theme of homecoming (nostos) as its overarching narrative line, whereas Journey to the West focuses on an outward spiritual pilgrimage (journey), devoting comparatively less emphasis to physical return. Along the Rescue and Care vector, points from both works are evenly intermingled, suggesting a similar quantitative emphasis on acts of succor and care. However, the sparse density within the inter-vector angles indicates that these three thematic dimensions do not exhibit strong co-occurrence relationships within either text.

Regarding the spatial distribution around Rescue and Care, a intuitive question arises: why are there virtually no text blocks situated in the angle between Rescue and Care and Violence and Coercion? Intuitively, when protagonists encounter violent crises, acts of rescue and care should naturally follow. A plausible explanation lies in our text segmentation preprocessing: within a single macro-level episode, paragraphs depicting violent confrontation and those depicting subsequent rescue were frequently split into adjacent but distinct text blocks, preventing their direct co-occurrence within the same fine-grained analysis unit.

In Figure 8, three primary loading vectors exhibit distinct and prominent directions, indicating that these themes represent relatively independent and highly characteristic text block types in both epics.

Specifically, text block scatter points concentrate densely along both sides of the vectors for Homecoming and Reunion and Task and Commitment. Along the Task and Commitment vector, green points representing Journey to the West are markedly more prevalent than orange points representing the Odyssey, accompanied by a dense concentration of Suffering and Hardship expressions. Conversely, along the Courage and Perseverance vector, orange points representing the Odyssey predominate. This spatial distribution closely aligns with the findings from the earlier histogram analysis, reflecting fundamental differences in narrative style between Eastern and Western literary traditions: Journey to the West relies primarily on detailed behavioral and situational descriptions rather than explicitly proclaiming or praising character virtues, whereas the Odyssey frequently utilizes direct expressions to portray heroic character.

Furthermore, a prominent scatter cluster emerges between the vectors for Violence and Coercion and Rescue and Care. This spatial aggregation indicates that within narrative episodes centered on Rescue and Care, themes of Emotional Distress and Courage and Perseverance frequently co-occur as interconnected narrative elements.

In Figure 9, the three primary loading vectors exhibit distinctly divergent directions. Notably, scatter points representing Journey to the West (in green) concentrate more densely around the vector for Emotional Distress. This pattern likely reflects the character arcs and spiritual progression within the text: the protagonists in Journey to the West (such as Tang Sanzang's disciples) begin their pilgrimage as mortals or beings with mortal attachments, rendering them more prone to displaying anxiety, apprehension, and emotional vulnerability when confronting harrowing tribulations. Along the vector for Courage and Perseverance, the relative proportion of scatter points across both works remains highly consistent with our earlier findings. Meanwhile, scatter points representing the Odyssey continue to predominate near the Homecoming and Reunion vector, further highlighting the structural divergence between the two epics regarding their narrative teleology and ultimate goals.

This section illustrates the projections and feature loading directions among the first five principal components. In Appendix [Additional graphs on joint PCA], we also present the projections and loading distributions across other principal component combinations, such as PC1 versus PC3 and three-dimensional spatial perspectives. These results consistently yield highly similar conclusions, thereby demonstrating the robustness of our empirical findings.

5 DISCUSSION AND CONCLUSION

5.1 Discussion

While the interpretable computational framework proposed in this study offers scalable and verifiable macro-level insights into the comparative narrative structures of Journey to the West and the Odyssey, several methodological and conceptual limitations warrant further refinement. First, regarding dictionary construction and metric design, the specification of thematic classes and keyword entries remains an area open to continuous optimization. The current nineteen analytical classes represent a preliminary attempt to operationalize complex literary themes; however, the granularity and scope of these categories could be further calibrated through deeper alignment with domain-specific literary taxonomies. Furthermore, keyword sets within each subclass—though carefully curated based on close reading and corpus characteristics—can be expanded or truncated to further reflect subtle regional or archaic stylistic variations.

Second, from an analytical perspective, this paper relies primarily on rule-based keyword matching and frequency counting, which focuses on explicit surface-level lexical occurrences rather than deeper contextual semantics. Future research could incorporate advanced semantic analysis models, such as word or sentence embeddings (Mikolov et al. 2013; Pennington et al. 2014) and pretrained language models such as ChatGPT or Claude, to elevate the analysis from the lexical level to the semantic level. Semantic representation models can better capture implicit narrative themes, metaphorical expressions, and contextual nuances that keyword matching might overlook, such as complex allegorical subtexts or indirect character motivations. Combining dictionary-based transparency with modern semantic embedding spaces will offer a more comprehensive approach to mapping the alignment between quantitative textual structures and qualitative literary interpretations.

5.2 Conclusion

This study has presented an interpretable, unsupervised computational pipeline to examine and compare the thematic structures of personal tribulations, narrative responses, and value orientations in Journey to the West and the Odyssey. By segmenting the original texts into standardized text blocks, constructing a transparent three-tiered bilingual dictionary, and applying inverse-

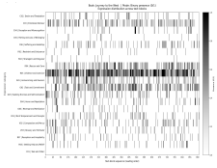
coverage weighting alongside joint PCA, Our empirical analysis reveals key thematic similarities and divergences between the two epics.

Specifically, the most prominent thematic similarity lies in the high prevalence of violent conflict and physical adversity across both works, a pattern consistent with the structural nature of epic narratives where heroic protagonists continuously confront life-threatening challenges. Conversely, dimensions such as Task and Commitment and Rescue and Care reflect distinct narrative teleologies and character dynamics. These observed disparities, may partially stem from dictionary keyword constraints and the limitation of dictionary-based methods in capturing complex semantic relationships—a nuance that provides a promising agenda for future research incorporating advanced language models.

Methodologically, this paper demonstrates the validity, rationality, and robustness of applying unsupervised computational models to comparative literature. By bridging quantitative text-as-data techniques with qualitative literary interpretation, our framework offers a scalable, reproducible, and transparent methodology that complements traditional close reading. Ultimately, this study provides a fresh analytical perspective and a systematic toolset for future scholars seeking to leverage statistical models and computational pipelines in the exploration of classic literary texts.

APPENDIX

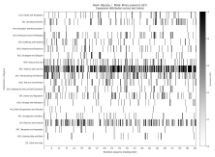
Coverage of Themes Indexed by the Sequence of Text Blocks



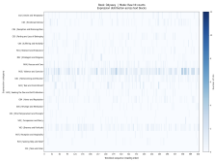
Journey to the West: unweighted expression distribution across text blocks



Journey to the West: weighted expression distribution across text blocks



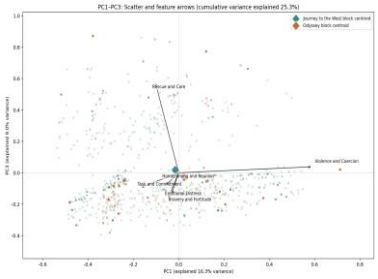
Odyssey: unweighted expression distribution across text blocks



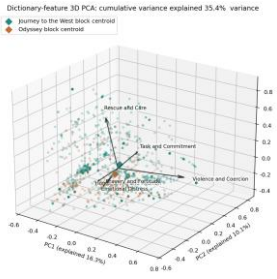
Odyssey: weighted expression distribution across text blocks

The figure shows both unweighted and weighted expression distribution across text blocks of two books.

Additional graphs on joint PCA



Histogram of unweighted class frequency



Histogram of weighted class frequency

Figure 16 exhibits a structural morphology closely resembling that of Figure 6. Because both biplots share the same horizontal axis while the vertical axis shifts from the second principal component (PC2) to the third principal component (PC3), the vector pointing toward the top-left changes accordingly from Task and Commitment to Rescue and Care. Except for Rescue and Care and Violence and Coercion, the directional loadings of the remaining categories remain largely consistent between the two projections.

Within this coordinate system, we observe a scatter distribution pattern highly congruent with that in Figure 16: substantial concentrations of text blocks aggregate along the vectors for Task and Commitment and Emotional Distress. Crucially, the densities of green points (Journey to the West) and orange points (Odyssey) within these clusters are virtually identical, demonstrating that under our dictionary metric, both works frequently and simultaneously combine these narrative themes

within their episodes. Furthermore, the angle between Violence and Coercion and the four vectors pointing toward the bottom-left features a dense intermingling of green and orange points, reinforcing the consistency of these cross-textual thematic co-occurrence patterns.

The three-dimensional scatter plot provides a novel interpretative perspective, the significance of which has already been fully characterized in the preceding discussion.

REFERENCES

[1] Bamman, David, Ted Underwood, and Noah A. Smith. 2014. "A Bayesian Mixed Effects Model of Literary Character." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 370–79. https://doi.org/10.3115/v1/P14-1035.

[2] Bantly, Francisca Cho. 1989. "Buddhist Allegory in the Journey to the West." The Journal of Asian Studies 48 (3): 512–24. https://doi.org/10.2307/2058637.

[3] Dougherty, Carol. 2001. The Raft of Odysseus: The Ethnographic Imagination of Homer's Odyssey. Oxford University Press. https://doi.org/10.1093/oso/9780195130362.001.0001.

[4] Griffin, Jasper. 2003. "Preface." In Homer: The Odyssey. Cambridge University Press. https://doi.org/10.1017/CBO9781139165334.001.

[5] Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." Political Analysis 21 (3): 267–97. https://doi.org/10.1093/pan/mps028.

[6] Jong, Irene J. F. de. 2001. A Narratological Commentary on the Odyssey. Cambridge University Press. https://doi.org/10.1017/CBO9780511482137.

[7] Lam, Ling Hon. 2019. "Allegory and 'World' Formation in the Journey to the West." In A Companion to World Literature, edited by Ken Seigneurie. Wiley. https://doi.org/10.1002/9781118635193.ctwl0080.

[8] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." arXiv Preprint arXiv:1301.3781.

[9] Pache, Corinne Ondine. 2020. "Odysseus." In The Cambridge Guide to Homer, edited by Corinne Ondine Pache, Casey Dué, Susan Lupack, and Robert Lamberton. Cambridge University Press.

[10] Pennington, Jeffrey, Richard Socher, and Christopher D Manning. 2014. "Glove: Global Vectors for Word Representation." Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–43.

[11] Reagan, Andrew J., Lewis Mitchell, David Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. "The Emotional Arcs of Stories Are Dominated by Six Basic Shapes." EPJ Data Science 5: 31. https://doi.org/10.1140/epjds/s13688-016-0093-1.

[12] Sarzynska-Wawer, Justyna, Aleksander Wawer, Aleksandra Pawlak, et al. 2021. "Detecting Formal Thought Disorder by Deep Contextualized Word Representations." Psychiatry Research 304: 114135.

[13] Yu, Anthony C. 1983. "Two Literary Examples of Religious Pilgrimage: The 'Commedia' and 'the Journey to the West'." History of Religions 22 (3): 202–30. https://doi.org/10.1086/462922.

[14] Yu, Anthony C. 2012. The Journey to the West, Revised Edition, Volume 1. University of Chicago Press.