E-ISSN: 2583-9667
Indexed Journal
Peer Reviewed Journal
https://multiresearchjournal.theviews.in

Received: 01-07-2025
Accepted: 05-09-2025
Published: 23-10-2025

INTERNATIONAL JOURNAL OF ADVANCE RESEARCH IN MULTIDISCIPLINARY

Volume 3; Issue 4; 2025; Page No. 131-133

The Data-Mined Canon: A Computational Analysis of Thematic Evolution in 20th Century British Poetry

Dr. Bala Rani

Assistant Professor, Saraswati Vidhya Mandir Law College, Rajju Bhaiya Sarswati Vihar Shikarpur, Bulandshahr, Uttar Pradesh, India

DOI: https://doi.org/10.5281/zenodo.17964308

Corresponding Author: Dr. Bala Rani

Abstract

This study employs computational text mining and Natural Language Processing (NLP) techniques to map and analyze the thematic evolution of 20th-century British poetry. Moving beyond traditional, subjective literary historiography, we construct a quantitative "data-mined canon" from a corpus of over 50,000 poems by 120 canonical and marginalized poets, sourced from digital archives. Using Latent Dirichlet Allocation (LDA) for topic modeling and diachronic word embedding alignment, we identify dominant thematic clusters (e.g., War & Trauma, Urban Modernity, Nature & Ecology, Myth & Archetype, Domestic & Introspective) and trace their flux across four temporal periods: Edwardian/Georgian (1900-1918), Modernist (1919-1945), Post-War (1946-1979), and Late-Century (1980-1999). Our analysis reveals: 1) a sharp, quantifiable thematic rupture caused by World War I, with the War & Trauma cluster displacing Pastoral Idealism; 2) the persistent, albeit transforming, presence of nature poetry, shifting from romantic escapism to environmental anxiety; 3) the rise of a distinct Domestic & Introspective cluster post-1950, correlating with the "Movement" poets and late-century explorations of identity; and 4) evidence of thematic "echoes," where earlier themes re-emerge in mutated forms. This computational approach challenges rigid periodization, demonstrating a more fluid and recursive model of literary change. It also surfaces overlooked thematic continuities in the work of women and post-colonial poets, prompting a re-evaluation of the canonical narrative. The paper argues for a complementary partnership between distant reading and close reading, where data-driven patterns generate new questions for qualitative interpretation.

Keywords: Computational Literary Studies, Digital Humanities, Topic Modeling, 20th Century British Poetry, Thematic Evolution, Distant Reading, Literary History, Natural Language Processing (NLP), Canon Formation

1. Introduction

The literary history of 20th-century British poetry is traditionally narrated as a sequence of dominant movements: the twilight of Victorianism, the Georgian interlude, the seismic shock of Modernism, the reactive empiricism of the Movement, and the pluralistic expansions of post-war and post-colonial voices. This narrative, constructed through selective close reading and critical consensus, inevitably emphasizes rupture and champions specific aesthetic ideologies. However, the advent of large-scale digitized corpora and computational methods offers an unprecedented opportunity to test, refine, or challenge this narrative through a macroscopic, data-driven lens-a practice Franco Moretti termed "distant reading."

This paper presents a computational analysis of thematic evolution across the century. We ask: Can quantitative methods map the thematic "DNA" of the poetic field? How do themes-abstract, recurrent constellations of subject matter-emerge, peak, fade, and mutate over time? Does the data confirm the sharp breaks of traditional periodization, or suggest more subtle, continuous transformations? By constructing a "data-mined canon," we aim not to replace qualitative criticism but to generate new, empirically-grounded hypotheses about literary change, potentially surfacing patterns invisible to localized reading.

2. Literature Review

Traditional Literary Histories: Foundational works like The New Apocalypse (Hynes, 1976) [4] and The Oxford English Literary History, Vol. 12 (Crawford, 2015) [2] provide period-based narratives. They often centre on canonical figures (Eliot, Yeats, Auden, Hughes, Larkin) and formal-aesthetic shifts.

Digital Humanities & Distant Reading: Moretti (2013) [7] advocates for analyzing literary systems rather than individual texts. Underwood (2019) [10] demonstrates the tracking of thematic trends over centuries using ML. So and Long (2013) [9] apply network analysis to literary influence.

Computational Literary Studies: Jockers (2013) [5] uses macroanalysis to study style and theme in novels. Piper (2018) [8] employs ML for conceptualizing "themes" in novels. For poetry, specific studies include Heuser & Le-Khac (2012) [3] on Learning to Read Data, but a comprehensive computational thematic history of 20th-century British poetry remains underexplored.

Critiques of the Canon: Feminist (Montefiore, 1994) [6] and post-colonial (Bhabha, 1994) [1] critiques have challenged the traditional canon, urging inclusion of marginalized voices. A computational approach can quantitatively assess their thematic integration or divergence.

3. Methodology & Data

3.1 Corpus Construction: We compiled a corpus of approximately 50,000 poems from 120 poets, aiming for balanced representation across gender, era, and canonical status. Primary sources included Poetry Magazine archive, Project Gutenberg, the Poetry Foundation, and scanned anthologies. Poets were categorized into four periods: P1 (1900-1918), P2 (1919-1945), P3 (1946-1979), P4 (1980-1999).

3.2 Text Preprocessing: Poems were cleaned, lowercased, and lemmatized using spaCy. A custom stopword list removed common English words and poetic archaisms, retaining content-rich lexical items.

3.3 Analytical Techniques

Topic Modeling (LDA): Applied to the entire corpus and per-period sub-corpora to identify stable thematic clusters. Optimal topic number (k=40 for full corpus) was determined via coherence scores. Topics were labeled interpretively (e.g., Trenches, Body, Memory → "War & Trauma").

Diachronic Analysis: Topic prevalence was calculated for each period to trace rise/fall. Keyness analysis (Log-Likelihood) identified words statistically over-represented in each period.

Word Embedding Evolution: Continuous Bag-of-Words (CBOW) models were trained on each period's text. Using orthogonal Procrustes alignment, we traced semantic shift vectors for seed words (e.g., "nature," "machine," "self," "empire").

4. Results and Analysis

4.1 Thematic Landscape: LDA revealed five dominant, persistent macro-clusters:

1. **War & Trauma:** Battle, blood, ghost, scream, silent, mud.
2. **Urban Modernity & Alienation:** City, street, machine, crowd, noise, glass.
3. **Nature & Ecology:** Tree, leaf, river, bird, stone, green, (later:) plastic, waste.
4. **Myth, History & Archetype:** Stone, king, word, time, bone, ancient.
5. **Domestic & Introspective Self:** Room, hand, mother, father, child, memory, skin.

4.2. Diachronic Thematic Flux

P1 (1900-1918): Dominated by a bifurcation between Pastoral Nature (Georgians) and early Urban/Mythic clusters (Imagists, early Modernists). War emerges explosively post-1914.

The WWI Rupture (P1→P2): The most significant quantitative shift. War & Trauma prevalence jumps from ~8% to ~28% of topical content, permanently altering the thematic palette. Pastoral Idealism sharply declines.

P2 (1919-1945): High Modernism. Myth/Archetype peaks (Eliot, Yeats), serving as a framework to order traumatic chaos. Urban themes mature, blending with anxiety (waste land, crowd).

P3 (1946-1979): Post-war settlement. The Domestic & Introspective cluster rises significantly (~22% prevalence), aligned with the Movement's anti-romantic ethos. Nature returns, but stripped of grandeur (Larkin's "unfenced existence"). War persists as memory and private guilt.

P4 (1980-1999): Pluralism and expansion. Domestic themes intensify, often exploring gender, race, and sexual identity (e.g., Duffy, Armitage). Nature mutates into explicit Ecology and pollution anxiety. Myth is reclaimed and subverted by post-colonial and feminist poets.

4.3 Semantic Shifts: Embedding analysis showed

"Nature": Associated vectors shift from beauty, divine, peace (P1) to threat, isolated, detail (P3) to fragile, system, activism (P4).

"Self": Moves from soul, universal (P1/P2) to observer, social (P3) to body, politicized, performative (P4).

5. Discussion

Our data-mined analysis supports and complicates traditional historiography. It confirms the catastrophic thematic impact of WWI, quantitatively validating the notion of a "broken tradition." However, it also shows that nature poetry never disappeared; it adapted, demonstrating remarkable thematic resilience. The rise of the Domestic cluster post-1950 offers quantitative evidence for a major re-centring of poetic concern towards the everyday and psychological, a shift often attributed anecdotally to the Movement.

Crucially, the analysis reveals thematic recursions: the Myth cluster of late-century poetry shares surface features with Modernist myth but is deployed for different (decolonizing, feminist) ends. This suggests evolution is less a linear progression than a spiral, where themes return transformed by new contexts.

Furthermore, including marginalized voices in the corpus shows they often intensify core thematic clusters (e.g., Jackie Kay amplifying the Domestic, Caribbean poets engaging intensely with Nature), while also introducing distinct new lexical fields (diaspora, border), expanding the thematic periphery of the canon.

6. Limitations & Future Work

Limitations include corpus bias (digitization favours published, out-of-copyright works), the interpretative leap in labeling LDA topics, and the difficulty of capturing tone, irony, and form-poetry's essence-through bag-of-words models. Future work will integrate metrical and stylistic features, employ more advanced transformer-based models (e.g., BERT) for deeper semantic understanding, and expand into comparative transnational analysis.

7. Conclusion

This study demonstrates the utility of computational methods for modeling large-scale literary change. The "data-mined canon" provides a dynamic, quantitative map of 20th-century British poetry's thematic evolution, revealing both confirmed ruptures and unexpected continuities. It moves us from a history of "isms" to a history of preoccupations-tracking how collective poetic attention shifts in response to historical trauma, social change, and aesthetic fatigue. Ultimately, the patterns identified do not close interpretation but open new avenues for inquiry: Why did domesticity surge post-war? How does the ecological theme differ in its late-century incarnation? By posing such questions, computational analysis serves as a powerful hypothesis generator, fostering a new, empirically-informed dialogue between distant and close reading in the ongoing project of understanding literary history.

8. References

1. Bhabha HK. The Location of Culture. London: Routledge; c1994.
2. Crawford R, editor. The Oxford English Literary History. Vol. 12, 1960–2000: The Last of England? Oxford: Oxford University Press; c2015.
3. Heuser R, Le-Khac L. Learning to read data: bringing out the humanistic in the digital humanities. Victorian Stud. 2012;54(1):79–86.
4. Hynes S. The Auden Generation: Literature and Politics in England in the 1930s. London: Bodley Head; c1976.
5. Jockers ML. Macroanalysis: Digital Methods and Literary History. Urbana (IL): University of Illinois Press; c2013.
6. Montefiore J. Feminism and Poetry: Language, Experience, Identity in Women's Writing. London: Pandora; c1994.
7. Moretti F. Distant Reading. London: Verso; c2013.
8. Piper A. Enumerations: Data and Literary Study. Chicago: University of Chicago Press; c2018.
9. So RJ, Long H. Network analysis and the sociology of modernism. boundary 2. 2013;40(2):147–182.
10. Underwood T. Distant Horizons: Digital Evidence and Literary Change. Chicago: University of Chicago Press; c2019.