3L: Language, Linguistics, Literature® The Southeast Asian Journal of English Language Studies
Vol 30(2), June 2024 http://doi.org/10.17576/3L-2024-3002-01

The Language of AI and Human Poetry: A Comparative Lexicometric Study

AFENDI HAMAT
Center for Research in Language and Linguistics
Universiti Kebangsaan Malaysia, Malaysia
fendi@ukm.edu.my

ABSTRACT

This study conducts a lexicometric analysis to compare the lexical richness and diversity in poetry generated by AI models with that of human poets. Employing a robust dataset that includes 1,333 AI-generated poems and 517 human-authored poems across seven distinct poetic eras, six key lexical metrics—Maas Index, MTLD, MATTR, HD-D, Hapax Legomenon Ratio, and Lexical Density—were applied for comparative analysis. The lexical characteristics of the poems were studied through a series of statistical tests and machine learning techniques, including Mann-Whitney U tests, Cliff's Delta, and Random Forest classification. The findings reveal a marked lexical superiority in human poetry, evidenced by significant differences and large effect sizes in all metrics except Lexical Density. HD-D emerged as the most discriminating factor, adeptly differentiating human poetry from its AI-generated counterpart. Further analysis identified the GPT-4 model as exhibiting the closest alignment to human poetry in terms of lexical attributes. The study discusses these outcomes in the context of AI's evolving linguistic competencies, shedding light on the inherent challenges and future prospects of AI in creative writing. Thus, this research provides an empirical framework for assessing AI's language generation abilities and sets the stage for further interdisciplinary exploration into the frontiers of artificial creativity.

Keywords: artificial intelligence; lexicometry; machine learning; lexical analysis; poetry

INTRODUCTION

Poetry, an artistic domain steeped in the nuanced fabric of human emotion and creativity, has traditionally been an emblem of our cultural and emotional expression. The evocative power of poetry to elicit profound emotional responses and conjure vivid mental imagery has been well-documented by researchers such as Wassiliwizky et al. (2017), who explore its neurological impact, and Belfi et al.(2018), who underscore the correlation between the aesthetic appeal of poetry and individual experiences of vividness. These insights affirm the longstanding view of poetry as a sanctuary of human ingenuity—a medium through which words transcend their literal meanings to resonate emotionally and imaginatively with readers. Kubi (2018) further reinforces this sentiment, positing poetry as a heartfelt reflection of our deepest passions and sentiments.

Yet, in the digital age, the emergence of artificial intelligence as a non-human author presents a paradigm shift, challenging the notion that the creation of poetry is an exclusively human endeavour. AI-generated poetry, though lacking innate emotion, has the capacity to mimic the structural and stylistic elements of poetry authored by humans. This raises intriguing questions about the nature of creativity and the potential for machines to replicate or even enhance literary art forms. Given the expanding capabilities of AI, it becomes imperative to examine its poetic productions not just qualitatively but also through a quantitative, lexicometric lens.

Lexicometry allows for the quantifying of lexical attributes and opens the possibility of rigorously evaluating how AI poetry stands in relation to human poetry within these specific dimensions. Such an approach demystifies the artistic capabilities of AI, allowing us to appreciate the intricate patterns of language that AI algorithms generate and to discern whether these patterns bear the hallmarks of what is traditionally considered poetic. It also equips us with a methodical framework to compare the output of various AI models, shedding light on the extent to which they can emulate the nuanced artistry of human poets.

As AI continues to evolve and take on more sophisticated creative tasks, our understanding of its potential and limitations in the realm of poetry must also advance. A lexicometric analysis provides a solid, exploratory foundation for such understanding, offering empirical insights that can inform both the development of AI technology and the broader conversation on the intersection of technology and the arts.

With this in mind, the present study embarks on a lexicometric investigation into AI-generated poetry. By applying rigorous statistical methods to compare the lexical features of poetry crafted by both human and artificial poets, it aims to unravel the quantitative aspects of poetic expression and explore the evolving landscape of literary creativity in the age of AI.

REVIEW OF LITERATURE

LEXICOMETRY

Lexicometry, the statistical study of lexicon within texts, serves as an important foundation in the field of corpus linguistics. It offers insights into the complexity and richness of human language use, and as McCarthy and Jarvis (2010) note, lexicometry allows researchers to quantify and analyse linguistic phenomena, transcending subjective interpretations and providing objective measures of language diversity and richness.

Quantitative language analysis began in the early 20th century, starting with readability research by Thorndike (1921). Johnson's introduction of the Type-to-Token Ratio (TTR) in 1944, measuring the ratio of unique words to total words in a text, spurred the development of various metrics like CTTR, MATTR, Yule's K, and Maas's Index (Maas). These metrics offer diverse statistical approaches to vocabulary measurement.

Lexicometry, like its parent discipline of corpus linguistics, benefits greatly from advances in information technology (Meng, 2021). Though its foundations lie in the early 20th century, lexicometry's recent resurgence has been directly fuelled by its close-knit dependence on information technology, particularly the rapid growth of accessible text corpora and powerful computational and analytical tools and methods. From its initial utility in measuring a language learner's performance (Abu-Rabiah, 2023; Zhang & Wu, 2021), it has been increasingly used in other disciplines, covering topics such as political studies (Benoit, 2020), economics (Attak, 2023), transportation (Mandják et al., 2019), climate and environmental issues (Richter et al., 2019) as well as healthcare (Fergadiotis et al., 2013). It also forms a foundational discipline in Natural Language Processing (NLP), a multi-disciplinary field of study that gives computers the ability to understand, manipulate and reproduce human language.

There are three important concepts often discussed in vocabulary studies and lexicometry: lexical density, lexical diversity, and lexical richness. Lexical density is the easiest to describe. It refers to the "packing" of content words (nouns, verbs, adjectives, adverbs) compared to function words (Sujatna et al., 2021; Vitta et al., 2023). Lexical diversity encompasses the range and variety

of vocabulary used in a sample (regardless of modality), indicating a broad spectrum of different words.

However, lexical richness has a more contentious history with regard to its usage in the literature. Jarvis (2013) argues that while the term 'lexical richness' has an original and rather simplistic meaning of a person's range of lexicon, it has increasingly been used by researchers as a superordinate term to include lexical diversity, density, and sophistication. However, this research argues that the term 'lexical richness' in its encompassing and evolved form is only relevant to vocabulary studies in the L1 and L2 contexts. The late 20th century saw a significant shift in studies on language learning and acquisition. Researchers began to view vocabulary not just as a set of isolated words but as a dynamic component of language proficiency (Uccelli et al., 2015). This led to a more holistic approach, where lexical richness came to include aspects of density and diversity (Jarvis, 2013). In language learning and acquisition, contextual factors like the learner's age, language exposure, and education level significantly influence lexical development (Collentine, 2004). Therefore, treating lexical richness as an overarching term allows for a more comprehensive assessment of these influences on a learner's vocabulary.

Lexicometry, which focuses on the quantitative analysis of texts, maintains a more distinct separation between the terms density and diversity/richness (Brglez & Vintar, 2022; Heng et al., 2023). This distinction allows for more precise and targeted analyses, which are essential in fields like computational linguistics, literary analysis, and corpus linguistics. However, it should be kept in mind that this is a field that has so far been unable to come to a satisfactory agreement on the terms of lexical diversity and lexical richness (McCarthy & Jarvis, 2010) and would commonly see overlapping definitions of the two terms (Jarvis, 2013). In a sense, to the earlier short definition of lexical diversity, we could also add the idea of lexical sophistication, i.e. the use of words that are not common as a marker of advanced vocabulary knowledge (Jarvis, 2013). This paper will use the term lexical richness, which is defined as the range, variety, and sophistication of vocabulary. It is separate from the idea of lexical density, which is simply the ratio of content words to the total words in a text.

It is well known to researchers that traditional indices used in assessing vocabulary are sensitive to text lengths (Zenker & Kyle, 2021). Their investigation of minimum text lengths based on data samples of L2 written texts with lengths of 50-200 tokens yielded the recommendation of MATTR (Moving Average TTR), MTLD (Measure of Textual Lexical Diversity), and MTLD-MA-Wrap for the token range, with HD-D being mentioned as stable as well. The issue of text length presents a rather confounding problem for the current research as human poetry could possibly be as short as one word or one character (works by Aram Saroyan). The issue of minimum text length will be discussed further in the methodology section.

Bestgen (n.d.) suggested that the problem of text lengths has been dealt with by the more recent matrices using parametrised length. In doing so, however, the matrices encountered another problem related to the parameter itself. The parameter could affect the results and should, therefore, be made known in reporting. She recommends the use of HD-D with the parameter set to the shortest text, as well as MATTR (with a parameter set to 50 tokens) and MTLD. In a study validating MTLD, McCarthy and Jarvis (2010) recommended employing a combination of MTLD, vocd-D (or HD-D), and Maas. While the study did not specifically focus on poetry and, therefore, did not address text length concerns, it demonstrated the efficacy of these metrics in capturing distinctive lexical information.

Interest in the intersection of artificial intelligence (AI) and poetry has permeated the field since its inception. Though not explicitly targeting poetry, early AI researchers were captivated by the notion of machines utilising language to form concepts, which is a core element of the poetic form. Notably, Alan Turing's 1950 proposal of the Turing Test, suggesting sonnet composition as a potential metric of machine intelligence, served as a seminal marker in the evolution of AI-generated poetry (Rockmore, 2020).

AI systems are primarily rule-based or based on neural networks and machine learning. Rule-based systems, deterministic and useful in stable, well-defined situations, are limited by their inability to learn or adapt (Swett et al., 2021). In contrast, neural networks and machine learning systems learn from data, adapt over time, and excel in complex, changing situations despite their opaque decision-making processes (Chen & Liu, 2014; Meng, 2021). Generative AI, like the GPT families, is trained on large datasets to generate new data, produce creative content and push machine capabilities (Labaca-Castro, 2023; Sennrich et al., 2016). These systems, capable of learning from vast data and handling complex tasks, have improved several computational linguistics tasks due to advancements in neural network architectures and machine learning techniques (Lo et al., 2022; Van de Cruys, 2020).

The mainstreaming of Generative AI also opens up questions and discussions regarding human creativity, what constitutes ‘arts’ and issues about copyright and originality (Atkinson & Barker, 2023; Hong & Curran, 2019; Lee, 2022; Rezwana & Maher, 2023). While recognising the broader importance of Generative AI's ethical and technical implications, this review deliberately focuses on linguistic and language-based analyses of poetry by Generative AI, leaving aside other considerations for future exploration.

AI-generated poetry, while impressive in its ability to mimic human style and creativity, has several limitations when compared to human-authored poetry. AI-generated poetry often lacks the emotional depth and complexity that human poets bring to their work. A study comparing reactions to a sonnet by Shakespeare and an AI-generated sonnet found that students favoured Shakespeare's work due to its complex language and greater emotional resonance (Rahmeh, 2023). The study looked at respondents' perceptions in terms of satisfaction, emotional engagement, and perceived linguistic complexity to form its findings. AI models currently struggle to fully comprehend and express emotions, which are integral to the art of poetry (Hutson & Schnellmann, 2023; Yi et al., 2018). Hutson and Schnellmann (2023) claim that ChatGPT 3 comes close to mimicking human writing in terms of vocabulary and word choices without giving quantification as to how close.

AI-generated poetry can sometimes lack authenticity and a deep understanding of context. For instance, a study comparing AI and human translations of Arabic poems to English found that the AI translations failed to capture the cultural context and nuances of the original poems. AI models often translate text word-for-word, which can be problematic for poetry that relies on figurative language, wordplay, and cultural references (Alowedi & Al-Ahdal, 2023).

AI-generated poetry often struggles with coherence in meaning, theme, or artistic conception for a poem as a whole (Hakami et al., 2021). For example, a study on Chinese poetry generation found a distinct gap between computer-generated poems and those written by poets, with the former often producing incoherences and inconsistencies (Yi et al., 2018). The studies presented so far gave rise to the question of what constitutes ‘quality’ in a poem. Poetic language is multifaceted (Sugunan, 2022) and non-systematic (Pulvirenti & Gambino, 2022). The use of lexical metrics to measure poetry is not perfect, but in the case of AI, it could serve as an empirical

starting point.

While AI models can generate convincing imitations of human writing, their creativity is often limited to the data they have been trained on. A study on co-creative writing experiences found that the AI-based application used in the study was evaluated as the weakest in support and idea quality (Kantosalo & Riihiaho, 2019). This suggests that AI models may struggle to generate truly original and creative ideas, a key aspect of poetry. Despite its present limitations, AI-powered poetry generation is rapidly advancing, with growing capabilities to mimic human-written works (Köbis & Mossink, 2021). As training data expands and diversifies, further sophistication in AI-generated poetry can be expected. This progress underscores the call by Köbis and Mossink (2021) for thorough, empirical methods to critically examine and compare AI-generated poetry, particularly for social science research. The study described in this paper offers a detailed quantitative analysis contrasting AI-generated poetry with its human-crafted counterparts. This methodology bridges a significant knowledge gap regarding AI's proficiency in creative text generation, particularly in empirically describing its capacity to imitate the unique lexical expressiveness intrinsic to human poets. The research described in this paper is guided by the following research questions:

1. How do the following lexical metrics (Maas Index, HD-D, MTLD, MATTR, Lexical Density and Hapax Legomenon Ratio) compare between AI-generated and human poetry, and what do these comparisons reveal about the lexical richness and diversity in both forms?
2. Which specific lexical metric most effectively distinguishes AI-generated poetry from human poetry?
3. Which of the AI models shows the closest alignment with human poetry in terms of the identified key lexical metrics?

METHODOLOGY

DATA COLLECTION AND PROCESSING

IDENTIFICATION OF POETIC ERAS

The first phase of the data collection process involved the identification of seven distinct poetic eras spanning a significant historical and literary timeline:

i. Elizabethan Era (1558-1603)
ii. Jacobean Era (1603-1625)
iii. Restoration & Augustan Era (1660-1745)
iv. Romantic Era (1780-1830)
v. Victorian Era (1837-1901)
vi. Modernist Era (1900-1945)
vii. Post-Modernist Era (1945-present)

SELECTION OF HUMAN POETS

The study used AI models (GPT-4, GPT-3, and PaLM2, accessed via the ChatGPT and Google Bard chatbots) to guide the selection of poets representative of their eras and distinct in style and theme. A preliminary list was compiled from these AI-assisted discussions. Two poets per era were then selected from this list, considering factors like style diversity, historical significance, and work availability. This approach balanced AI suggestions with human judgment for academic rigour.

COMPILATION OF HUMAN POEMS

Subsequently, an extensive collection of poems was gathered from the works of each chosen poet within their respective eras. The selection is based on available and accessible online sources, mostly from PoetryFoundation.

GENERATION OF AI POEMS

In the next phase, three artificial intelligence (AI) models, GPT-4(AI1), GPT-3(AI2) and PaLM2 (AI3), were employed to generate poems in the style of each human poet. The AI models were instructed to generate approximately 30-40 poems for each poet. The prompts used were:

a) *Do you know of XXX?* (XXX is the name of the human poet). If the answer is affirmative, then it will be prompted further.
b) *Please write a poem in the style of XXX.*

All the AI models recognised the names of the selected human poets and were therefore instructed to generate poems based on the style of the poet as described by the prompt in (b). The writing prompt is purposely kept simple so as to not guide the generation process too much. In simpler terms, the AI models were given 'freedom' to write as long as they emulate the style of the poet.

The AI-generation process took place in May and June of 2023. This is important to note as all the AI models are continuously updated, and this may affect the quality of their output. For example, Google Bard replaced their Large Language Model (LLM), PaLM2, with Gemini in December 2023. The collected poems were then transferred into an Excel file for further processing and analysis. The data breakdown is shown in Table 1:

TABLE 1. Data Breakdown

Heading	Heading	No of Poems	Tokens
Elizabethan	Human	74	34388
	AI	186	86206
Jacobean	Human	69	113832
	AI	189	67138
Restoration	Human	40	57039
	AI	194	64972
Romantic	Human	92	48928
	AI	186	66186
Victorian	Human	119	84044

	AI	198	70770
Modernist	Human	94	28723
	AI	185	77669
Post-Modernist	Human	29	16674
	AI	195	76540
Total	Human	517	383628
	AI	1333	509481
Grand Total		1850	893109

DATA PROCESSING

In the processing stage of the data analysis, two fundamental Natural Language Processing (NLP) techniques were used: tokenisation and Part-Of-Speech (POS) tagging. These procedures were executed utilising the *NLTK* and *SpaCy* libraries, renowned for their efficiency and accuracy in linguistic processing within the Python environment.

During tokenisation, the methodology focused on retaining lexical items while excluding punctuations. Punctuation marks are undeniably significant stylistic elements in poetry; however, their removal is warranted in this study due to its focus on lexical analysis. This approach aligns with the objective of quantitative lexical analysis of the poetic texts rather than their stylistic composition.

In this study, the Python library *lexical richness* was employed for computing key linguistic metrics: the Maas Index, Measure of Textual Lexical Diversity (MTLD), Hypergeometric Distribution D (HD-D), and Moving-Average Type-Token Ratio (MATTR). These metrics were selected based on their ability to capture unique lexical information (McCarthy & Jarvis, 2010) and their suitability for texts of varied lengths, an important aspect in the analysis of poetry.

Among the collected poems, 22 – all human-authored – consist of fewer than 50 tokens each. Initially, the researcher considered excluding these shorter poems to prevent potential skewing of the dataset. However, upon further reflection, it was recognised that these works hold significant value in the context of poetic expression. Poetry, inherently diverse and unbounded, is not constrained by length except in certain formal structures. The brevity of these poems does not undermine their linguistic and creative significance; instead, it illustrates the broad spectrum of human poetic expression. After all, no one would accuse the minimalist Aram Saroyan of not being a 'real' poet.

Therefore, the researcher decided to retain these poems in the analysis. This decision aligns with a commitment to a comprehensive and inclusive examination of poetic works, acknowledging that the essence of poetry transcends mere length. By including these shorter pieces, the study embraces a more holistic view of the poetic form, ensuring that the analysis reflects the richness and diversity of human poetic creativity, regardless of length.

The two other metrics, Lexical Density and Hapax Legomena Ratio, were calculated using custom Python scripts as the were no suitable libraries available for use. The results were then saved into a separate Excel file for analysis.

EXPLORATORY DATA ANALYSIS

Once the data is processed, the researcher carried out the exploratory phase to identify surface patterns, outliers, and other data features. This process is initially carried out using the Python library *ydata-profiling*. The library is chosen for its simplicity and utility in identifying patterns of a dataset. Strong correlations were discovered between the metrics as shown in Table 2.

TABLE 2. Correlations between the metrics (N=1850)

		maas	mtld	mattr	hdd	hapax_ratio	lexical-density
maas	Correlation Coefficient	1.000	-.831**	-.721**	-.915**	-.894**	-.408**
	Sig. (2-tailed)		0.000	0.000	0.000	0.000	0.000
mtld	Correlation Coefficient	-.831**	1.000	.936**	.908**	.625**	.360**
	Sig. (2-tailed)	0.000		0.000	0.000	0.000	0.000
mattr	Correlation Coefficient	-.721**	.936**	1.000	.841**	.522**	.313**
	Sig. (2-tailed)	0.000	0.000		0.000	0.000	0.000
hdd	Correlation Coefficient	-.915**	.908**	.841**	1.000	.718**	.308**
	Sig. (2-tailed)	0.000	0.000	0.000		0.000	0.000
hapax_ratio	Correlation Coefficient	-.894**	.625**	.522**	.718**	1.000	.409**
	Sig. (2-tailed)	0.000	0.000	0.000	0.000		0.000
lexical-density	Correlation Coefficient	-.408**	.360**	.313**	.308**	.409**	1.000
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	
	Correlation Coefficient	1.000	-.831**	-.721**	-.915**	-.894**	-.408**

**Correlation is significant at the 0.01 level (2-tailed)

Notably, the Maas Index, an inverse measure of lexical complexity, displayed strong negative correlations with MTLD, MATTR, HDD, and Hapax Legomenon Ratio. These relationships suggest that as the simplicity of language increases (as indicated by a higher Maas Index), metrics that directly measure lexical richness and diversity tend to decrease. Such correlations underscore a consistent pattern in the lexical attributes captured across poems, regardless of the originating source (AI or human). These findings pave the way for deeper comparative analysis (such as via statistical tests for correlation) to discern the quantitative nuances of language employed by AI and human poets.

For the next phase, the surface features of the dataset were explored using the SPSS statistical software. Surface-level comparisons for the metrics of human poems and AI poems were analysed using descriptive statistics. The results for each metric comparison are shown in their respective table.

TABLE 3. Maas Index

Human				AI			Heading
Mean		.015556	Std. Error .00020	Mean		.028595	Std. Error .00021042414
95% Confidence Interval for Mean	Lower Bound	.015159		95% Confidence Interval for Mean	Lower Bound	.028183	
	Upper Bound	.015954			Upper Bound	.029001	
5% Trimmed Mean		.015229		5% Trimmed Mean		.028423	
Median		.014845		Median		.027463	
Variance		.000		Variance		.000	
Std. Deviation		.00460		Std. Deviation		.007682	

Minimum	.00000		Minimum	.011978	
Maximum	.040429		Maximum	.053861	
Range	.040429		Range	.04188	
Interquartile Range	.00430		Interquartile Range	.011097	
Skewness	1.531	.107	Skewness	.382	.067
Kurtosis	5.154	214	Kurtosis	-.483	.134

Descriptive statistics for the Maas Index show human poems have lower mean and median values than AI poems, indicating greater lexical complexity. Human poems also exhibit less variability and a smaller range and interquartile range, suggesting consistent complexity and narrower distribution. Both distributions are positively skewed, with human poems more so, and human poems have a higher kurtosis, indicating a peak around the mean. These findings highlight the richness of vocabulary and complexity in human poetry. The broader range in AI poems may suggest language experimentation or inconsistent lexical sophistication across different AI models. The significant and substantial difference in complexity is confirmed by confidence intervals, standard deviations, and interquartile ranges.

TABLE 4. Measure of Textual Lexical Diversity

Human				AI			
			Std. Error				Std. Error
Mean		107.85271	2.16446	Mean		43.34253	.52937
95% Confidence Interval for Mean	Lower Bound	103.6004		95% Confidence Interval for Mean	Lower Bound	42.30402	
	Upper Bound	112.10496			Upper Bound	44.38103	
5% Trimmed Mean		105.83852		5% Trimmed Mean		41.85947	
Median		105.41176		Median		40.15250	
Variance		2422.104		Variance		373.561	
Std. Deviation		49.21487		Std. Deviation		19.32772	
Minimum		13.000		Minimum		12.02086	
Maximum		515.28542		Maximum		161.93847	
Range		502.28542		Range		149.91760194	
Interquartile Range		64.13675		Interquartile Range		21.939508236	
Skewness		1.471	.107	Skewness		1.484	.067
Kurtosis		8.950	214	Kurtosis		3.918	.134

Descriptive statistics for MTLD show human poems have higher mean and median values, indicating greater lexical diversity. Human poems also exhibit greater variability, range, and interquartile range, suggesting a broader spread of MTLD values and greater lexical diversity. Both distributions are positively skewed, with human poems less so, and human poems have a higher kurtosis, indicating a peak around the mean. These findings highlight the wider vocabulary and diversity in human poetry, with a tendency for clustering around higher MTLD values, while AI poetry shows a narrower, more uniform distribution.

TABLE 5. Moving Average Type-Token Ratio (MATTR)

Human				AI			
			Std. Error				Std. Error
Mean		.88487	.001867	Mean		.79924	.0017419
95% Confidence Interval for Mean	Lower Bound	.88121		95% Confidence Interval for Mean	Lower Bound	.79582	
	Upper Bound	.88854			Upper Bound	.80265	
5% Trimmed Mean		.88748		5% Trimmed Mean		.80293	
Median		.89153		Median		.80826	
Variance		.002		Variance		.004	
Std. Deviation		.042456		Std. Deviation		.06360	
Minimum		.60821		Minimum		.558636	
Maximum		1.000		Maximum		.930622	
Range		.39178		Range		.371986	
Interquartile Range		.04791		Interquartile Range		.082633	
Skewness		-1.380	.107	Skewness		-.882	.067
Kurtosis		4.439	214	Kurtosis		.985	.134

Descriptive statistics for MATTR show human poems have higher mean and median values, indicating greater lexical diversity. Human poems also exhibit less variability and a wider range, suggesting consistent complexity and broader lexical diversity. Both distributions are negatively skewed, with human poems being more so and human poems having a higher kurtosis, indicating a peak around the mean. These findings highlight the broader vocabulary and diversity in human poetry, with a tendency for clustering around higher MATTR values, while AI poetry shows greater variability and a flatter distribution.

TABLE 6. Hypergeometric Distribution D (HD-D)

Human				AI			
			Std. Error				Std. Error
Mean		.84818	.00225	Mean		.71208	.0018416
95% Confidence Interval for Mean	Lower Bound	.84374		95% Confidence Interval for Mean	Lower Bound	.70846	
	Upper Bound	.85262			Upper Bound	.71569	
5% Trimmed Mean		.85228		5% Trimmed Mean		.71359	
Median		.86063		Median		.71731	
Variance		.003		Variance		.005	
Std. Deviation		.051360		Std. Deviation		.06724	
Minimum		.61120		Minimum		.469835	
Maximum		1.000		Maximum		.90000	
Range		.38879		Range		.430173	
Interquartile Range		.058946		Interquartile Range		.092099	
Skewness		-1.412	.107	Skewness		-.346	.067
Kurtosis		3.062	.214	Kurtosis		.061	.134

For HD-D, the descriptive statistics show human poems have higher mean and median values, indicating greater rare word usage. Human poems also exhibit less variability and a narrower range and interquartile range, suggesting consistent complexity and narrower distribution. Both distributions are negatively skewed, with human poems being more so and human poems having a higher kurtosis, indicating a peak around the mean. These findings highlight the richer usage of rare words and diversity in human poetry, with a tendency for clustering around higher HD-D values, while AI poetry shows a flatter distribution.

TABLE 7. Hapax Legomenon Ratio

Human				AI			
			Std. Error				Std. Error
Mean		.49555	.00661	Mean		.27487	.00328
95% Confidence Interval for Mean	Lower Bound	.48256		95% Confidence Interval for Mean	Lower Bound	.26842	
	Upper Bound	.50855			Upper Bound	.28132	
5% Trimmed Mean		.49443		5% Trimmed Mean		.27312	
Median		.49659		Median		.27300	
Variance		.023		Variance		.014	
Std. Deviation		.15039		Std. Deviation		.120027	
Minimum		.16501		Minimum		.0038961	
Maximum		1.0000		Maximum		.60000	
Range		.83498		Range		.59610	
Interquartile Range		.20840		Interquartile Range		.176797	
Skewness		.036	.107	Skewness		.174	.067
Kurtosis		-.344	.214	Kurtosis		-.698	.214

Descriptive statistics for the Hapax Legomenon Ratio show human poems have higher mean and median values, indicating greater unique word usage. Human poems also exhibit greater variability, range, and interquartile range, suggesting a broader spread of unique word usage. Both distributions are positively skewed, with AI poems more so, and both have negative kurtosis, with AI poems more so, indicating a flatter distribution. These findings highlight the richer usage of unique words and diversity in human poetry, with a tendency for clustering around higher values, while AI poetry shows a flatter distribution.

TABLE 8. Lexical Density (LD)

Human				AI			
			Std. Error				Std. Error
Mean		.378134	.0019793	Mean		.362906	.00109971
95% Confidence Interval for Mean	Lower Bound	.374246		95% Confidence Interval for Mean	Lower Bound	.360748	
	Upper Bound	.382023			Upper Bound	.365063	
5% Trimmed Mean		.378010		5% Trimmed Mean		.36335	
Median		.378516		Median		.3633540	
Variance		.002		Variance		.002	
Std. Deviation		.0450055		Std. Deviation		.0401508	
Minimum		.23674		Minimum		.2345679	
Maximum		.53333		Maximum		.47706422	
Range		.29658		Range		.24249631	

Interquartile Range	.060855		Interquartile Range	.0580489	
Skewness	.077	.107	Skewness	-.128	.067
Kurtosis	.214	.214	Kurtosis	-.016	.134

The statistics for Lexical Density (LD) show human poems have a slightly higher mean and median values, indicating denser lexical item usage. Human poems also exhibit slightly greater variability but a similar range and interquartile range, suggesting comparable variability in both corpora. Both distributions are symmetrical with no extreme outliers. These findings highlight the slightly denser array of lexical items in human poetry, with similar levels of variation in lexical density in both human and AI poetry.

ADVANCED DATA ANALYSIS

As the exploration suggests statistical differences and the dataset is non-normal, the Mann-Whitney U test was then carried out together with Cliff's delta calculations, as shown in Table 9.

TABLE 9. Significance (Mann-Whitney U) and Effect Size (Cliff's delta) for Human Poems (n=517) and AI Poems (n=1333)

	Maas Index	MTLD	MATTR	HD-D	Hapax Ratio	Lexical Density
Mann-Whitney U	40385.5	63907.5	76433.5	37497.0	91721.0	276999.0
Asymp. Sig. (2-tailed)	<.001	<.001	<.001	<.001	<.001	<.001
Cliff's delta	0.8828	0.8145	0.7782	0.8912	0.7338	0.1961

Table 9 shows significant differences in all metrics between human and AI poems, with high effect sizes for Maas, MTLD, MATTR, HD-D, and Hapax Legomenon Ratio. Lexical Density also differs significantly, but less so. These metrics hint at what differentiates human and AI poetry. To further investigate, a Random Forest analysis, a versatile machine learning technique, was conducted using the *scikit-learn* Python library, with results in Table 10.

TABLE 10. Random Forest Feature Importance Ranking

Feature	Importance
HDD	0.308304564
Maas	0.247727514
Hapax Legomenon	0.140868338
MTLD	0.126902079
MATTR	0.111208443
Lexical Density	0.064989062

The results from the Random Forest analysis provide a ranking of lexical metrics based on their importance in distinguishing between human and AI-generated poetry. The 'Importance' score reflects how much each metric contributes to the accuracy of the classification. HDD leads the metrics, followed by Maas and Hapax Legomenon. The lower half of the ranking is made up of MTLD and MATTR, with Lexical Density coming in last. The first and the last ranking position by Random Forest is also reflected in the effect sizes in Table 9 for both HD-D and Lexical Density.

The next research question seeks to discover which of the three AI models is closest to human poetry based on the metrics. A Multivariate Analysis of Variance (MANOVA) procedure was carried out to establish the overall differences among the four groups (three AI models and human poetry). Then, post-hoc pairwise comparisons were carried out to see which AI model was most similar to the human group. Table 11 shows the results of the MANOVA, and Table 12 shows the results of Tukey's Honestly Significant Difference (HSD) test.

TABLE 11. MANOVA Results

Intercept	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.0019	6.0000	1841.0000	160897.3159	0.0000
Pillai's trace	0.9981	6.0000	1841.0000	160897.3159	0.0000
Hotelling-Lawley trace	524.3802	6.0000	1841.0000	160897.3159	0.0000
Roy's greatest root	524.3802	6.0000	1841.0000	160897.3159	0.0000
C(poetmodel)					
Wilks' lambda	0.1292	18.0000	5207.6196	307.2268	0.0000
Pillai's trace	1.4259	18.0000	5529.0000	278.2500	0.0000
Hotelling-Lawley trace	3.1827	18.0000	3676.0043	325.3476	0.0000
Roy's greatest root	1.9038	6.0000	1843.0000	584.7900	0.0000

TABLE 12. Tukey HSD results (Group 1 = Human, Groups 2-4 = AI1, AI2, AI3)

group1	group2	meandiff	p-adj	lower	upper	reject	Metric
1	2	0.0063	0	0.0054	0.0073	TRUE	maas
1	3	0.0171	0	0.0162	0.0181	TRUE	maas
1	4	0.0152	0	0.0143	0.0161	TRUE	maas
2	3	0.0108	0	0.0098	0.0118	TRUE	maas
2	4	0.0089	0	0.0079	0.0099	TRUE	maas
3	4	-0.0019	0	-0.0029	-0.0009	TRUE	maas
1	2	-54.3575	0	-59.4274	-49.2877	TRUE	mtld
1	3	-64.5883	0	-69.5703	-59.6064	TRUE	mtld
1	4	-73.666	0	-78.6075	-68.7246	TRUE	mtld
2	3	-10.2308	0	-15.4701	-4.9915	TRUE	mtld
2	4	-19.3085	0	-24.5094	-14.1077	TRUE	mtld
3	4	-9.0777	0	-14.1929	-3.9625	TRUE	mtld
1	2	-0.0682	0	-0.0773	-0.0592	TRUE	mattr
1	3	-0.0624	0	-0.0712	-0.0535	TRUE	mattr
1	4	-0.124	0	-0.1328	-0.1152	TRUE	mattr
2	3	0.0059	0.3664	-0.0034	0.0152	FALSE	mattr
2	4	-0.0558	0	-0.065	-0.0465	TRUE	mattr
3	4	-0.0617	0	-0.0707	-0.0526	TRUE	mattr
1	2	-0.1078	0	-0.1181	-0.0975	TRUE	hdd
1	3	-0.1506	0	-0.1608	-0.1405	TRUE	hdd
1	4	-0.1478	0	-0.1578	-0.1377	TRUE	hdd
2	3	-0.0428	0	-0.0535	-0.0322	TRUE	hdd
2	4	-0.04	0	-0.0506	-0.0294	TRUE	hdd
3	4	0.0029	0.8954	-0.0076	0.0133	FALSE	hdd
1	2	-0.1023	0	-0.1204	-0.0842	TRUE	hapax_ratio
1	3	-0.3089	0	-0.3267	-0.2911	TRUE	hapax_ratio
1	4	-0.2427	0	-0.2604	-0.2251	TRUE	hapax_ratio
2	3	-0.2066	0	-0.2253	-0.1879	TRUE	hapax_ratio
2	4	-0.1404	0	-0.159	-0.1218	TRUE	hapax_ratio
3	4	0.0662	0	0.0479	0.0844	TRUE	hapax_ratio

1	2	0.0111	0.0001	0.0047	0.0176	TRUE	lexicaldensity
1	3	-0.0179	0	-0.0242	-0.0116	TRUE	lexicaldensity
1	4	-0.0366	0	-0.0429	-0.0303	TRUE	lexicaldensity
2	3	-0.029	0	-0.0357	-0.0224	TRUE	lexicaldensity
2	4	-0.0477	0	-0.0543	-0.0411	TRUE	lexicaldensity
3	4	-0.0187	0	-0.0252	-0.0122	TRUE	lexicaldensity

The MANOVA results (Table 11) exhibit highly significant multivariate effects for the grouping variable, which consists of Human (1), AI1 (2), AI2 (3), and AI3 (4), across the lexical metrics. The low Wilks' lambda value (0.1292) and the highly significant F-values suggest that the mean vectors of the lexical metrics are substantially different among the four groups. This multivariate test establishes that the differences in lexical metrics are not due to random chance and warrants further investigation into pairwise differences.

The Tukey HSD test results reveal significant differences in lexical metrics between humans and AI models. For the Maas Index, an inverse metric, AI1's poetry is closer to human complexity than AI2's, while AI3 is less complex than AI2. For MTLD, all AI models show less lexical diversity than humans, with AI3 being the least similar. The same trend is observed for MATTR, with AI3 having the lowest similarity in lexical variation. For HDD, AI2 and AI3 are similar but differ significantly from humans and AI1, indicating lower word frequency diversity. In the Hapax Legomenon Ratio, AI1 is closest to humans in unique word usage, while AI2 deviates the most. Lastly, all AI models differ significantly from humans in Lexical Density, with AI3 being the least similar.

In order to provide a triangulation for the findings, another Random Forest classification test was carried out. In this case, the confusion matrix from Random Forest was utilised to provide a classification test to see if Random Forest could predict the 'authorship' of the poems accurately. The results are shown in Table 13.

TABLE 13. Random Forest (Confusion Matrix)

	Predicted 1	Predicted 2	Predicted 3	Predicted 4
Actual 1(Human)	99	5	0	10
Actual 2(AI1)	7	70	0	5
Actual 3(AI2)	0	6	72	13
Actual 4(AI3)	1	7	8	67

The results show that for Actual 1 (Human), 99 out of 114 poems (approximately 86.8%) were correctly classified as human-written. 5 were misclassified as AI1, ten as AI3, and zero for AI2. For Actual 2 (AI1), 70 out of 82 poems (approximately 85.4%) were correctly classified as AI1. There are small numbers of misclassifications across other AI models and human classes. Interestingly, AI1 is also misclassified as human poetry the most among the AI models. For Actual 3 (AI2), 72 out of 91 poems (approximately 79.1%) were correctly classified as AI2. Misclassifications are spread out, with a noticeable number being classified as AI3 (13). Lastly, for Actual 4 (AI3), 67 out of 83 poems (approximately 80.7%) were correctly classified as AI3. Misclassifications are mainly AI2 and AI1.

The Random Forest model accurately classifies human and AI poems based on selected metrics. Cross-classification is observed in AI2 and AI3, possibly due to similar generation methods or overlapping metrics. Despite some misclassifications, the model effectively differentiates between human and AI poems. AI3 is notably challenging to classify, suggesting shared characteristics with both human and AI poems or less distinct lexical traits.

The final step of the triangulation for RQ3 is the PCA (Principal Component Analysis) to reduce the high-dimensional lexical data into two or three dimensions. A scatter plot of this reduced data, where each point represents a poem coloured by its group (Human, AI1(GPT-4), AI2(GPT-3), AI3(PaLM2), would visually demonstrate the clustering of similar poems. This could be useful to show how closely AI poems are grouped with human poems.

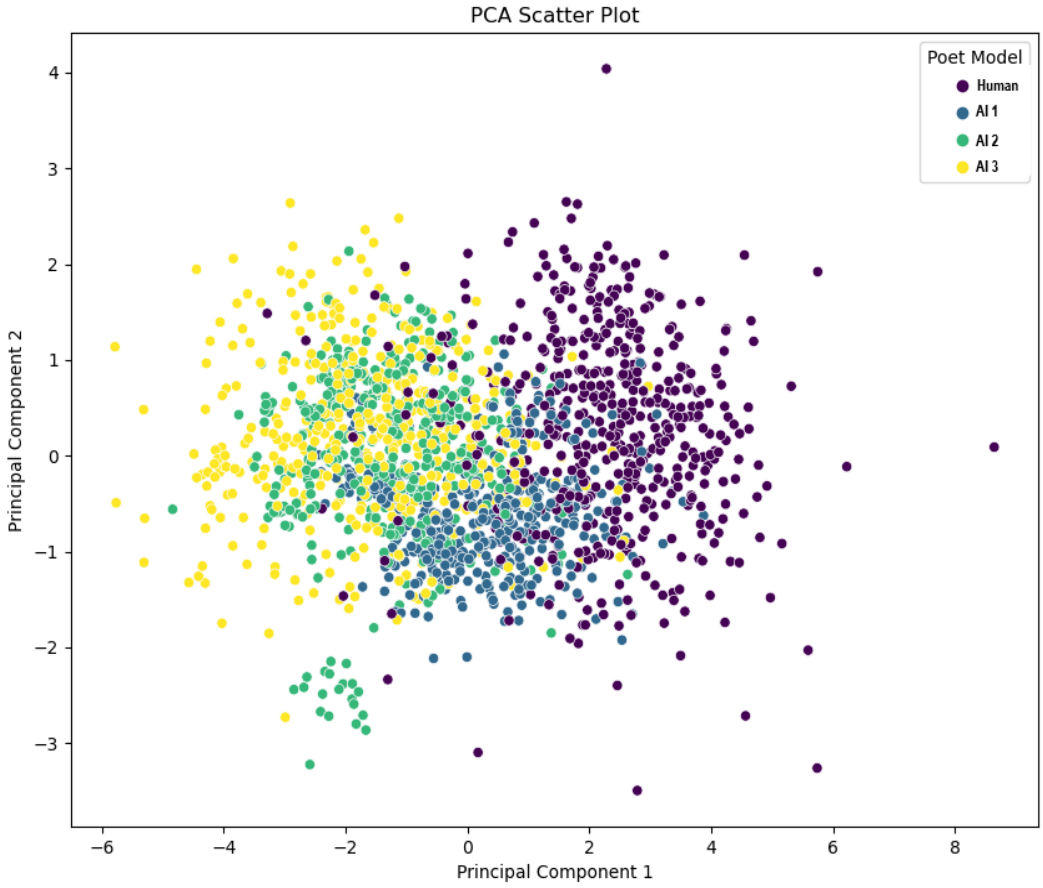

FIGURE 1. PCA Scatter Plot

TABLE 14. PCA Loadings

	PC1	PC2	PC3	PC4	PC5	PC6
maas	-0.461252	-0.030667	0.344934	0.091172	-0.428503	0.689492
mtld	0.421380	0.193947	0.326567	0.810905	-0.130012	-0.060880
mattr	0.417548	0.162980	0.509664	-0.547656	-0.455381	-0.178986
hdd	0.463665	0.216155	0.058531	-0.182215	0.533240	0.646002
hapax_ratio	0.407110	-0.111643	-0.673828	0.018620	-0.546371	0.262460
lexicaldensity	0.233701	-0.935786	0.239054	0.025385	0.096142	0.051520
Explained variance by component	0.69702368	0.14045899	0.09375915	0.04141857	0.02144186	0.00589775

The PCA loadings table (Table 14) provides insight into how each of the original variables (lexical metrics) contributes to the principal components. In conjunction with the scatter plot, this can help us understand the underlying patterns of lexical richness and diversity among the poems from humans and AI models.

Principal Component 1(PC1) accounts for 69.70% of the variance (the most substantial part), has high positive loadings for 'mtld', 'mattr', 'hdd', and 'hapax_ratio', and a high negative loading for 'maas'. Since 'maas' is an inverse metric (where lower values indicate greater complexity), its negative loading on PC1 means that higher scores on this component are associated with greater lexical simplicity. In other words, poems that score higher on PC1 tend to be less complex. PC2 explains 14.04% of the variance and has a very strong negative loading for 'lexicaldensity'. This suggests that poems with lower scores on PC2 are using a denser, more varied vocabulary.

The human poems are centralised in the scatter plot, indicating moderate complexity and lexical density. Since PC1 is associated with simplicity (due to the negative loading of 'maas'), human poems are likely to show balanced complexity. The AI1 poems overlap with human poems, suggesting that AI1 models can mimic human-like complexity in their poetry to some extent. Given the PC1's association with simplicity, the overlap implies that AI1 and human poems share similar levels of lexical simplicity or complexity. AI2 poems are lower on the PC2 axis, which could indicate higher lexical density (given the negative loading of 'lexicaldensity' on PC2), but possibly less complexity as they are also higher on PC1. AI3 poems, which appear higher on the plot (higher PC2 values), may be using a less dense vocabulary. They also spread towards the right, indicating varying degrees of lexical simplicity.

The scatter plot patterns suggest that AI1 is closest to human poems based on the lexical metrics. It can also be noted that AI2's poems are less complex but possibly denser than AI1's, and AI3's poems are the most lexically simple and least dense.

LIMITATIONS

This study acknowledges key methodological limitations. Firstly, the selection of poets and their works significantly influences the results; different choices or focusing on a single poetic era, with its unique linguistic styles, could alter the findings. Secondly, the chosen linguistic metrics, despite being well-justified, may not encompass all relevant aspects. Alternate metrics like Herdan's C or VM might offer additional insights, particularly for poems of similar length, suggesting an area for future research.

The study's approach to prompt engineering, intentionally simplistic to mimic non-expert interactions with AI chatbots, also poses a limitation. A more advanced prompt engineering could elicit different linguistic features from Large Language Models (LLMs), impacting the AI-generated poetry.

Finally, the reliance on publicly available LLMs is a limitation. Using specialised, finely-tuned LLMs for poetry generation could significantly alter the results. This highlights the potential effect of advanced fine-tuning on LLMs' generative abilities, presenting another avenue for future research.

DISCUSSION

This research examined the lexical characteristics of AI-generated versus human poetry to understand AI's ability to emulate human creative expression. It focused on three key questions: comparing the lexical richness and diversity of AI and human poetry, identifying the most discriminative lexical metrics, and determining which AI model most closely resembles human poetry.

The initial analysis showed human poetry surpassing AI in lexical richness and diversity, with significant differences in metrics like the Maas Index, HD-D, MTLD, MATTR and Hapax Legomenon Ratio, indicating greater lexical sophistication in human poetry. The exception was Lexical Density, where human and AI poetry were similar, suggesting AI's capability to match humans in using content-rich words but not in the breadth of vocabulary—a key element of poetic depth. While Sunico (2021) argues that poetry's emotive power does not rely on rich vocabulary, Obermeier et al. (2013) highlight the importance of lexical elements in poetry's aesthetic and emotional impact. Furthermore, if we look at poetry as a means to convey emotions with minimal words (Parsons & Pinkerton, 2022), those minimal words would need to be rich enough to convey the emotions.

When examining the metrics that most effectively differentiate AI-generated poetry from human-authored works, the Mann-Whitney U test and Cliff's delta identified HD-D as the most discerning, followed by the Maas Index and MTLD. Conversely, the Random Forest classification also identified HD-D as the most predictive, but it highlighted the Hapax Legomenon Ratio over MTLD. This discrepancy in ranking illustrates the complexity of evaluating poetry across various dimensions; significance and predictive power are not always aligned. These findings underline the multifaceted nature of poetic language and suggest that AI's capacity to imitate human, poetic language depends on the particular lexical metric in question. HD-D, which measures the distribution of words across different frequency bands, emerges as a crucial factor in differentiating AI from human poetry, perhaps indicating that AI models may not yet fully capture the subtleties of word frequency distributions characteristic of human poets.

The comparative analysis of AI models unveiled that AI1(GPT-4) most closely aligns with human poetry, a finding corroborated by PCA scatter plots which visually clustered AI1 with human poetry, distinct from AI2(GPT-3) and AI3(PaLM). This suggests that GPT-4 may have been trained on a corpus that closely mirrors the lexical patterns found in human poetry, or it may employ algorithms more adept at capturing the nuances of human linguistic expression with access to better computing resources to execute those algorithms. It is well known that GPT-4 is much more advanced than GPT-3 (Ayinde et al., 2023), and it comes only as a paid service. However, the scatter plots also revealed overlapping clusters, indicating that while GPT-4 is the closest to human poetry, the boundaries are not always clear-cut. This speaks to the growing competence of AI in mastering certain aspects of the poetic lexicon, yet it also highlights that the more subtle elements of poetic language, such as the interplay between words and their connotative layers, remain a distinctly human forte.

The PCA loadings provided further insights into the variables contributing to the distinction between human and AI poetry. For instance, the negative loading of the Maas Index on the first principal component (PC1) indicates that a higher Maas Index (denoting lower lexical sophistication) is associated with AI poetry. Given that the Maas Index is an inverse metric, this supports the notion that AI poetry tends to be less lexically rich. Conversely, the positive loadings for the other metrics on PC1 suggest that these are more characteristic of human poetry.

CONCLUSION

In 2016, the judges at an AI arts competition noted that "Robots would be starving artists if they attempt to write poetry" (Cramer, 2016). A lot has obviously changed since then, and while this study could not attest to the economic prospects of AI poets, it has empirically demonstrated that despite significant advancements, AI-generated poetry has not yet reached the lexical richness and diversity emblematic of human verse. This lexicometric analysis not only provides a measurable assessment of AI's current linguistic capabilities but also lays a methodological groundwork for future research. As AI technology continues to refine its grasp on human language, it opens up new avenues for inquiry into the nature of creativity and the potential for machines to partake in what was once considered a quintessentially human endeavour.

ACKNOWLEDGEMENTS

This work was supported by the Research Fund SK-2023-011 provided by the Faculty of Social Sciences and Humanities, Universiti Kebangsaan Malaysia.

REFERENCES

Abu-Rabiah, E. (2023). Evaluating L2 Vocabulary Development Features Using Lexical Density And Lexical Diversity Measures. *LLT Journal: Journal on Language and Language Teaching*, *26*(1), 168–182. https://doi.org/10.24071/llt.v26i1.5841

Alowedi, N. A., & Al-Ahdal, A. A. M. H. (2023). Artificial Intelligence based Arabic-to-English machine versus human translation of poetry: An analytical study of outcomes. *Journal of Namibian Studies : History Politics Culture*, *33*. https://doi.org/10.59670/jns.v33i.800

Atkinson, P., & Barker, R. (2023). AI and the social construction of creativity. *Convergence: The International Journal of Research into New Media Technologies*, *29*(4), 1054–1069. https://doi.org/10.1177/13548565231187730

Attak, E. H. (2023). *Tax Policy and Entrepreneurship* (pp. 421–442). https://doi.org/10.4018/978-1-6684-8781-5.ch019

Ayinde, L., Wibowo, M. P., Ravuri, B., & Bin Emdad, F. (2023). ChatGPT as an important tool in organisational management: A review of the literature. *Business Information Review*, *40*(3), 137–149. https://doi.org/10.1177/02663821231187991

Belfi, A. M., Vessel, E. A., & Starr, G. G. (2018). Individual Ratings of Vividness Predict Aesthetic Appeal in Poetry. *Psychology of Aesthetics, Creativity, and the Arts*, *12*(3), 341–350. https://doi.org/10.1037/aca0000153

Benoit, K. (2020). *The SAGE Handbook of Research Methods in Political Science and International Relations* (L. Curini & R. Franzese, Eds.; Vol. 2). SAGE Publications Ltd. https://doi.org/10.4135/9781526486387

Bestgen, Y. (n.d.). *Measuring Lexical Diversity in Texts: The Twofold Length Problem*. https://doi.org/https://doi.org/10.48550/arXiv.2307.04626

Brglez, M., & Vintar, Š. (2022). Lexical Diversity in Statistical and Neural Machine Translation. *Information*. https://doi.org/10.3390/info13020093

Chen, R., & Liu, H. (2014). Quantitative Aspects of *Journal of Quantitative Linguistics*. *Journal of Quantitative Linguistics*, *21*(4), 299–340. https://doi.org/10.1080/09296174.2014.944327

Collentine, J. (2004). The Effects of Learning Contexts on Morphosyntactic and Lexical Development. *Studies in Second Language Acquisition*. https://doi.org/10.1017/s0272263104262040

Cramer, J. (2016, May 16). *Can Robot Artists Create Human-Quality Work? Not Yet*. https://home.dartmouth.edu/news/2016/05/can-robot-artists-create-human-quality-work-not-yet

Fergadiotis, G., Wright, H. H., & West, T. M. (2013). Measuring lexical diversity in narrative discourse of people with aphasia. *American Journal of Speech-Language Pathology*, *22*(2). https://doi.org/10.1044/1058-0360(2013/12-0083)

Hakami, A., Alqarni, R., Almutairi, M., & Alhothali, A. (2021). Arabic Poems Generation using LSTM, Markov-LSTM and Pre-Trained GPT-2 Models. *Advances in Machine Learning*, 139–147. https://doi.org/10.5121/csit.2021.111512

Heng, R., Pu, L., & Liu, X. (2023). The Effects of Genre on the Lexical Richness of Argumentative and Expository Writing by Chinese EFL Learners. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2022.1082228

Hong, J.-W., & Curran, N. M. (2019). Artificial Intelligence, Artists, and Art. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *15*(2s), 1–16. https://doi.org/10.1145/3326337

Hutson, J., & Schnellmann, A. (2023). The Poetry of Prompts: The Collaborative Role of Generative Artificial Intelligence in the Creation of Poetry and the Anxiety of Machine Influence. *Faculty Scholarship*, *462*. https://digitalcommons.lindenwood.edu/faculty-research-papers/462/

Jarvis, S. (2013). Capturing the Diversity in Lexical Diversity. *Language Learning*, *63*(SUPPL. 1), 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x

Kantosalo, A., & Riihiaho, S. (2019). Quantifying co-creative writing experiences. *Digital Creativity*, *30*(1), 23–38. https://doi.org/10.1080/14626268.2019.1575243

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior*, *114*. https://doi.org/10.1016/j.chb.2020.106553

Kubi, B. (2018). Bemoaning of Love: An Aspect of Ga Women's Discourse on Love in Adaawe Song- Texts. *International Journal of Comparative Literature and Translation Studies*, *6*(2), 43. https://doi.org/10.7575/aiac.ijclts.v.6n.2p.43

Labaca-Castro, R. (2023). Generative Adversarial Nets. In *Machine Learning under Malware Attack* (pp. 73–76). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-40442-0_9

Lee, H.-K. (2022). Rethinking creativity: creative industries, AI and everyday creativity. *Media, Culture & Society*, *44*(3), 601–612. https://doi.org/10.1177/01634437221077009

Lo, K.-L., Ariss, R., & Kurz, P. (2022). *GPoeT-2: A GPT-2 Based Poem Generator*. http://arxiv.org/abs/2205.08847

Mandják, T., Lavissière, A., Hofmann, J., Bouchery, Y., Lavissière, M. C., Faury, O., & Sohier, R. (2019). Port marketing from a multi-disciplinary perspective: A systematic literature review and lexicometric analysis. *Transport Policy*, *84*, 50–72. https://doi.org/10.1016/j.tranpol.2018.11.011

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

Meng, Q. (2021). The Pedagogy of Corpus-Aided English-Chinese Translation From a Critical &Amp; Creative Perspective. *Theory and Practice in Language Studies*. https://doi.org/10.17507/tpls.1101.04

Obermeier, C., Menninghaus, W., von Koppenfels, M., Raettig, T., Schmidt-Kassow, M., Otterbein, S., & Kotz, S. A. (2013). Aesthetic and Emotional Effects of Meter and Rhyme in Poetry. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2013.00010

Parsons, L. T., & Pinkerton, L. (2022). Poetry and Prose as Methodology: A Synergy of Knowing. *Methodological Innovations*. https://doi.org/10.1177/20597991221087150

Pulvirenti, G., & Gambino, R. (2022). Einbildungskraft (Imagination). *Goethe-Lexicon of Philosophical Concepts*, *2*(1). https://doi.org/10.5195/glpc.2022.59

Rahmeh, H. (2023). Digital Verses Versus Inked Poetry: Exploring Readers' Response to AI-Generated and Human-Authored Sonnets. *Scholars International Journal of Linguistics and Literature*, *6*(09), 372–382. https://doi.org/10.36348/sijll.2023.v06i09.002

Rezwana, J., & Maher, M. L. (2023). Designing Creative AI Partners with COFI: A Framework for Modeling Interaction in Human-AI Co-Creative Systems. *ACM Transactions on Computer-Human Interaction*, *30*(5), 1–28. https://doi.org/10.1145/3519026

Richter, A., Ng, K., & Fallah, B. (2019). Bibliometric and text mining approaches to evaluate landfill design standards. *Scientometrics*, *118*. https://doi.org/10.1007/s11192-019-03011-4

Rockmore, D. (2020, January 7). What Happens When Machines Learn to Write Poetry. *The New Yorker*. https://www.newyorker.com/culture/annals-of-inquiry/the-mechanical-muse

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725. https://doi.org/10.18653/v1/P16-1162

Sugunan, D. (2022). Multifarious nature in Bharathy's Lyrical Literature. *International Research Journal of Tamil*, *4*(SPL 2), 1–7. https://doi.org/10.34256/irjt22s21

Sujatna, E. T. S., Heriyanto, H., & Andri, S. (2021). Lexical Density and Variation in Indonesian Folklores in English Student Textbooks: An SFL Study. *Leksika Jurnal Bahasa Sastra Dan Pengajarannya*. https://doi.org/10.30595/lks.v15i2.11102

Sunico, R. C. (2021). The Poetry of Simple Words. *Perspectives in the Arts and Humanities Asia*. https://doi.org/10.13185/paha2020.10208

Swett, B. A., Hahn, E. N., & Llorens, A. J. (2021). Designing Robots for the Battlefield: State of the Art. In *Robotics, AI, and Humanity* (pp. 131–146). Springer International Publishing. https://doi.org/10.1007/978-3-030-54173-6_11

Thorndike, E. L. (1921). *The Teacher's Word Book*. Teacher's College, Columbia University. https://pure.mpg.de/rest/items/item_2395369_2/component/file_2395368/content

Uccelli, P., Galloway, E. P., Barr, C. D., Meneses, A., & Dobbs, C. L. (2015). Beyond Vocabulary: Exploring Cross-Disciplinary Academic-Language Proficiency and Its Association With Reading Comprehension. *Reading Research Quarterly*. https://doi.org/10.1002/rrq.104

Van de Cruys, T. (2020). Automatic Poetry Generation from Prosaic Text. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2471–2480. https://doi.org/10.18653/v1/2020.acl-main.223

Vitta, J. P., Nicklin, C., & Albright, S. W. (2023). Academic Word Difficulty and Multidimensional Lexical Sophistication: An English-for-academic-purposes-focused Conceptual Replication of Hashimoto and Egbert (2019). *Modern Language Journal*. https://doi.org/10.1111/modl.12835

Wassiliwizky, E., Koelsch, S., Wagner, V., Jacobsen, T., & Menninghaus, W. (2017). The emotional power of poetry: neural circuitry, psychophysiology and compositional principles. *Social Cognitive and Affective Neuroscience*, *12*(8), 1229–1240. https://doi.org/10.1093/scan/nsx069

Yi, X., Li, R., & Sun, M. (2018). Chinese Poetry Generation with a Salient-Clue Mechanism. *Proceedings of the 22nd Conference on Computational Natural Language Learning*, 241–250. https://doi.org/10.18653/v1/K18-1024

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, *47*. https://doi.org/10.1016/j.asw.2020.100505

Zhang, Y., & Wu, W. (2021). How effective are lexical richness measures for differentiations of vocabulary proficiency? A comprehensive examination with clustering analysis. *Language Testing in Asia*, *11*(1). https://doi.org/10.1186/s40468-021-00133-6