What Makes Literary Quotes Stand Out? A Multidimensional Analysis of Literary Creativity

Berke Yazan, Almila Akdağ, Davide Dell'Anna
Department of Information and Computing Sciences
Utrecht University
{b.yazan, a.a.akdag, d.dellanna}@uu.nl

Abstract

Current approaches in computational linguistics often rely on statistical deviation as the primary proxy for measuring creativity. This fails to capture the full extent of the semantic, linguistic, and affective dimensions that distinguish creative from ordinary language. To explore this multidimensionality, we analyse literary quotes from Goodreads, curated by readers for their perceived value. We construct ALOFT, a dataset that pairs these quotes with six comparative corpora: length-matched random snippets from the same source texts, gathered through a novel human-in-the-loop workflow; popular quotes, public-domain quotes with surrounding context, and a non-literary baseline. Our analysis measures features such as complexity, lexical novelty, semantic movement, and affective tone across these corpora using explainable methods to reveal feature interactions. We demonstrate that widely shared quotes are distinguishable by a combination of accessible vocabulary, unexpected semantic arrangements, and emotional charge. Furthermore, we show that traditional information-theoretic metrics struggle with short text lengths and that sentence-level contextual embeddings significantly outperform static word vectors at identifying the signature of highly valued language. This work offers a new, explainable perspective on perceived value and linguistic creativity, along with a new dataset and a data-gathering workflow for future research. [1]

Introduction

There are nearly infinite letter combinations possible in human language, but only a vanishingly small fraction holds meaning or value. Moreover, among all grammatically correct and semantically coherent sentences, only a rare few achieve the exceptional value found in richer expressions, such as creative, artistic, and figurative artifacts. Understanding what exactly distinguishes such expressions from everyday language is currently an ongoing challenge.

Developing computational tools that can recognise and preserve creative uses of language would let us study them systematically. This can help enrich the creative capabilities of both humans and machines while also counteracting the growing risks of algorithmic monoculture caused by the statistically collectivist outputs of Large Language Models (LLMs) (Priyanshu and Vijay 2024; Wu, Black, and Chandrasekaran 2024).

Influentially, Boden (2004) argues that the creativity of an artifact arises from a combination of novelty and value (Boden 2004). Traditionally, Natural Language Processing (NLP) approaches have used novelty as the primary proxy for measuring creativity, considering metrics such as surprisal (Bunescu and Uduehi 2022; He, Peng, and Liang 2019) or semantic distance (Beaty et al. 2022; Orwig et al. 2021; Dumas, Organisciak, and Doherty 2021). Similarly, automatic creativity scoring systems have historically adopted this novelty-centric approach (Beaty and Johnson 2021; Organisciak et al. 2023).

Measuring novelty alone provides an incomplete picture of Boden's definition. A variety of studies have therefore attempted to operationalise the broader value aspect of creativity through a combination of measurable dimensions, such as linguistic memorability (Danescu-Niculescu-Mizil et al. 2012), structural and syntactic composition (Ashok, Feng, and Choi 2013; van Cranenburgh and Bod 2017), affective resonance (Jacobs 2018; Hipson and Mohammad 2020), and stylistic uniqueness (Potthast et al. 2018). More recently, LLMs have demonstrated the ability to evaluate creativity in written domains (Ismayilzada, Stevenson, and van der Plas 2024) at levels comparable to those of human judges in certain contexts. This suggests that such models have learned to approximate how humans communicate and interpret artistic or creative language.

Despite these advancements, we still lack transparent frameworks for understanding how or why certain linguistic expressions are considered creative, and existing creativity evaluation models still struggle to capture the nuanced and diverse nature of artistic expressions (Ismayilzada et al. 2024). Such an asymmetry between current generative capabilities, creativity metrics, and understanding of linguistic creativity becomes apparent when studying richer expressions, such as figurative text (Tsvetkov, Mukomel, and Gershman 2014; Mao, Lin, and Guerin 2019) and literary artifacts, especially when subtle deviations and cultural context are involved (Ismayilzada et al. 2024; Kuznetsova, Chen, and Choi 2013).

Another shortcoming of computational literary research is

[1] Generative AI tools (Grammarly, Sonnet 4.5, Gemini 3.0) were used for grammar checking, sentence refinement, LaTeX/code assistance, and adapting python code snippets. All research design, theoretical contributions, and scientific interpretations are solely the authors' own.

the reliance on public-domain resources such as the works of Shakespeare or the Gutenberg Corpus (Gerlach and Font-Clos 2018), partly due to the manual difficulty of extracting contemporary text under fair academic use. This inevitably leads to an imbalanced representation of human literature and introduces temporal and cultural bias. These create a need to provide the community with easier approaches to collect a more diverse dataset that includes contemporary artifacts.

In this paper, we address these shortcomings by first compiling a new dataset called ALOFT (A Lot Of Four Thousands)[2], which enables controlled comparisons across literary and non-literary corpora. ALOFT is created using a novel human-in-the-loop data-extraction pipeline described in this paper, sourced by both public-domain and contemporary text. Second, we introduce a multidimensional analysis of features that distinguish creative from ordinary language. Leveraging ALOFT, we compute a variety of creativity metrics, including language accessibility/complexity, information-theoretic measures, semantic embeddings and trajectories, and affective tone. Finally, we analyse how these dimensions interact using explainable methods, aiming to reveal the subtle mechanisms that contribute to linguistic creativity.

Our results demonstrate that linguistic creativity relies on an interplay of semantic trajectories, accessible vocabulary, and affective tone, moving beyond mere statistical novelty. Specifically, our analysis reveals that widely shared literary quotes are distinguished by the presence of familiar words in unexpected semantic arrangements that carry emotional weight. By evaluating contextual models against static word vectors, we show that sentence-level contextual embeddings are essential for capturing the subtle signature of creative language. Ultimately, by providing a transparent, multidimensional perspective on linguistic creativity, along with a novel data-gathering workflow and a novel dataset, we establish a robust foundation for future research in both computational literary studies and the development of more deeply informed generative models.

Related Work

Defining and Measuring Literary Creativity

Quantifying literary creativity has long been a focal point at the intersection of computational creativity and digital humanities. Van Cranenburgh and Bod (2017) developed a data-oriented model to predict literary ratings from textual features, using a corpus of Dutch novels evaluated by a large survey (van Cranenburgh and Bod 2017). They demonstrated that sentence length, vocabulary richness, and syntactic diversity are significantly correlated with perceived literariness, with an R^2 of 0.76. Extending this analysis, Van Cranenburgh et al. (2019) integrated machine learning alongside human judges to further explore the perceived literariness of contemporary Dutch novels (van Cranenburgh, van Dalen-Oskam, and van Zundert 2019). They found that literariness cannot be reduced exclusively to semantic deviations, but, promisingly, can still be identified without human oversight or sociological context.

In parallel to structural analysis, the Neurocognitive Poetics Model (Jacobs, 2015) links aesthetic impact to cognitive and affective processes and finds that textual features can systematically influence readers' emotional engagement and aesthetic appreciation (Jacobs 2015). Subsequently, Jacobs (2018) introduced the Gutenberg English Poetry Corpus (GEPC), a manually curated collection of over three million lines of poetry (Jacobs 2018). Applying methods such as latent semantic analysis, topic modelling, and sentiment estimation, they found that poems with higher surprisal, lexical diversity, and sonority are associated with greater aesthetic or emotional resonance.

From Static Features to Semantic Trajectories

Extending the scope from poetry to 2,722 mixed domain literary texts, Jacobs and Kinder (2021) treat literariness, creativity, and beauty as distinct, quantifiable axes and demonstrate that semantic complexity measures and affective tone can effectively capture them (Jacobs and Kinder 2021). They operationalise literariness through Stepwise Distance (SWD) (van Cranenburgh, van Dalen-Oskam, and van Zundert 2019), which measures the semantic divergence between adjacent text segments. Using this, they found that plays rank as the most literary category, followed by poems and novels. To quantify creativity, they utilised Forward Flow (FF) (Gray et al. 2019), calculating the average semantic distance between a word and all preceding words in a sequence. They identified poems and plays as the most creative categories in the corpus, whereas novels ranked lowest.

Quotability and Popularity

Danescu-Niculescu-Mizil et al. (2012) analysed thousands of movie quotes from IMDb (Internet Movie Database) (Internet Movie Database (IMDb)), comparing quoted and non-quoted snippets from the same movies (Danescu-Niculescu-Mizil et al. 2012). Their analysis revealed that movie quotes tend to adhere to conventional syntactic structures but employ less common word choices. Similarly, Tekir et al. (2023) formalised the literary quote detection task using a benchmark dataset (T50) derived from Goodreads and Project Gutenberg (Tekir et al. 2023). Their approach shows that quotes exhibit higher lexical distinctiveness than their surrounding context. From the popularity and likability perspective, using Goodreads ratings as the ground truth, Maharjan et al. (2017) proposed a multi-task neural architecture to predict book likability from their textual content alone (Maharjan et al. 2017). Their findings underscore the value of sentiment, dialogue structure, and narration style.

Non-Interpretability of Human and LLM Evaluation

Evaluating creative and artistic language remains a significant challenge, even for humans. Interpretations often

[2]Pipeline, analysis code, and dataset (excluding the Google Books-derived columns, which can be regenerated via the pipeline): `https://github.com/BerkeYazan/aloft`.

Column Name (Abbreviation)	Description
Goodreads Sample Quote (GSQ)	Random samples of user-curated quotes
Goodreads Popular Quote (GPQ)	Most highly-liked quotes from Goodreads corpus
Google Books Length-Matched (GBLM)	Same-length passages from same books of Goodreads samples
T50 Quote (T50Q)	User-curated quotes from the T50 study
T50 Quote-Free Context Length Matched (T50LM)	T50 Quote-Free Context Matched to original T50 quote length
Non-Literary Baseline (NLB)	Length-matched informational text passages (Wikipedia + Brown)
Google Books Page Text	Complete OCR-extracted page content
T50 Full Context	Complete 21-sentence context from Project Gutenberg
T50 Quote-Free Context	Context with quote sentences removed

Table 1: Columns of the final ALOFT dataset and their abbreviations. Bold entries are used in analysis. All columns contain $N = 4{,}569$ entries.

vary (Iser 1978; Barthes 1967; Sontag 1966), and explanations may involve post hoc rationalisations rather than objective criteria (Summers 2017). Recent studies show that human evaluation of creativity can be replicated by modern Large Language Models (LLMs) (Ismayilzada, Stevenson, and van der Plas 2024; Porter and Machery 2024; Marco, Rello, and Gonzalo 2024). However, because these models largely mirror the non-interpretable nature of human ratings, they inherit the same transparency problem, where neither human nor LLM judgements expose which textual features drove the rating, so the evaluation cannot be decomposed into its contributing factors (Ismayilzada et al. 2024).

The Dynamic Between Accuracy and Explainability

Recent work suggests that models can predict what is considered creative at levels approaching human judgement without providing a framework for why those linguistic arrangements hold value (Ismayilzada, Stevenson, and van der Plas 2024). For instance, automatic creativity scoring systems have started to transition from simple novelty-centric metrics to LLM-based models fine-tuned on human feedback (Beaty and Johnson 2021; Organisciak et al. 2023). While this significantly improved predictive performance, it has simultaneously limited our understanding of the evaluations. This dynamic between accuracy and interpretability highlights the need for multidimensional, explainable approaches that can both accurately evaluate and decompose the underlying features and mechanisms of creative language.

Dataset Creation

In this section, we introduce the ALOFT dataset and illustrate the human-in-the-loop data-extraction pipeline we implemented to create it. Table 1 outlines the various data sources integrated into ALOFT. We aim to construct a dataset in which quoted literary text, non-quoted literary text, and non-literary text can be represented in a diverse and comparable manner.

Data Sources

Goodreads Sample Quotes (GSQ) and Popular Quotes (GPQ). Our primary signal for extraordinary language comes from two Goodreads quote datasets on Kaggle (faellielupe 2020; Verma 2021), including nearly half a million user-submitted quotes combined. To ensure active community validation, we discard quotes with no likes, leaving 225,262 quotes. To serve as a separate popularity-filtered column, we extract a subset of the most-liked quotes.

Google Books Length-Matched Context (GBLM). Google Books page previews serve as one of our sources for the non-quoted literary contexts from the same works as our GSQ sample. This serves as a control for authorial style and a source for non-quoted literary language. This data is gathered using the custom extraction pipeline detailed in the Data Extension section.

T50 Quotes (T50Q). Tekir et al. (Tekir et al. 2023) provide a dataset with nearly five-thousand pairs from public-domain literary quotes and their immediate surroundings. These are sourced from Project Gutenberg (dominantly published before 1928), and are shown to be lexically distinct from their surrounding text (ibid). Enriching their data with our modern Goodreads corpus (mean publication year 1990) enables a diachronic analysis of literary preferences.

Non-Literary Baseline (NLB). To establish a non-literary contrast to our literary texts, we sample a control baseline using a diverse set of non-fiction texts from the Brown (Francis and Kučera 1982) (40%) and Wikipedia (Wikimedia Foundation 2023) (60%) corpora.

Data Extension

Goodreads to Google Books Pipeline. To create the GBLM partition, we introduce a terms of service (ToS) compliant pipeline that uses manual browser actions and human-in-the-loop cleaning, as illustrated in Figure 1. We first enrich our metadata for future research by using the Wikimedia API to add the publication year and the original language. We then draw a sample using occurrence and like-weighted stratified sampling to ensure that the subset reflects the platform's actual distribution of reader endorsements, which serves as our proxy for literary value. To extract sampled book pages from the same source works, we manually capture and review 12,000 randomly selected candidate preview pages for over 50 hours, yielding 5,333 valid full-page screenshots. We then process these validated screenshots using Google Cloud Vision OCR. OCR imperfections were detected algo-

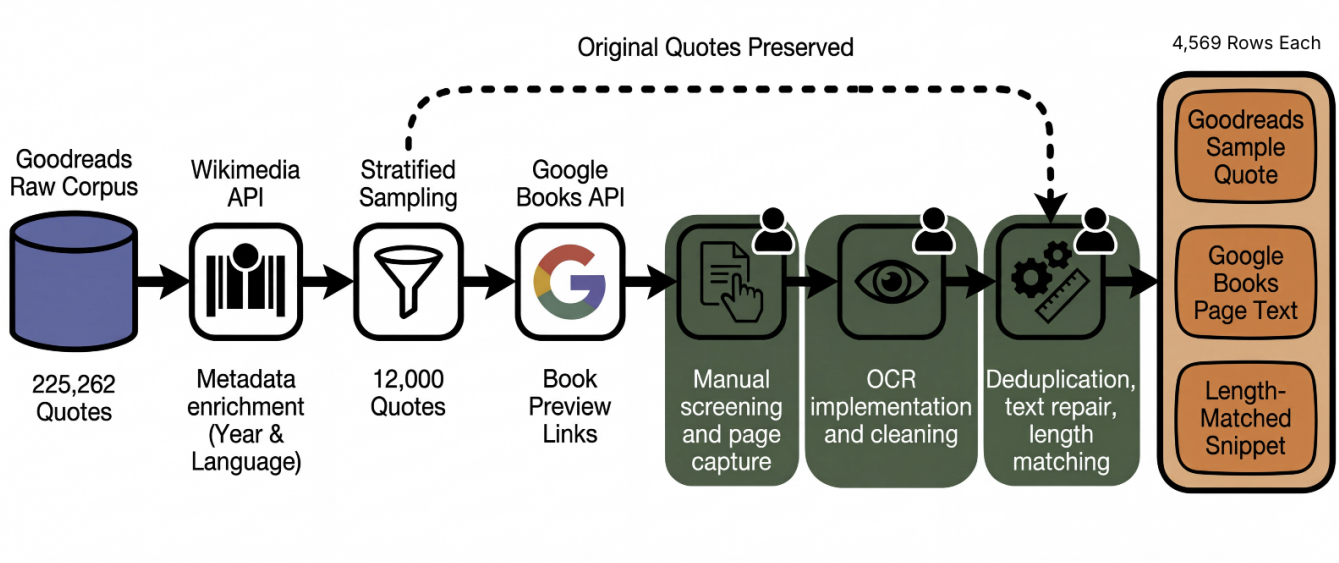


Figure 1: Goodreads to Google Books pipeline combining hybrid extraction with human-in-the-loop validation.

rithmically and corrected manually via a custom interface.

Data Preparation

Length Matching and Finalising the ALOFT Dataset. Because computational metrics scale with text size, we apply a length-matching protocol (within a 70% to 130% token-count tolerance) tailored to each corpus. This involves sampling passages from Google Books pages, cropping the surrounding context from the T50 dataset, and extracting appropriately sized text from our non-literary sources. To prevent data leakage, we apply a fuzzy matching algorithm to identify and remove quotes that also appeared in the sampled page text. We employ an additional algorithmic-detection/manual-correction approach for each data source. We create separate custom interfaces to highlight and correct spelling and punctuation errors that were already present in the data or produced during previous steps. For potential duplicates, we compute the semantic distance between data points using SBERT and manually remove overlapping records, yielding a dataset of 4,569 parallel entries. Designing specialised user interfaces, we enable human-in-the-loop involvement at each step, ensuring human verification of algorithmic processing.

Methodology

To systematically identify the features that distinguish creative language from ordinary language, we propose a framework. This section outlines our approach in three distinct parts. First, we introduce the computational metrics used to characterise the multidimensional features of creativity. Second, we define the comparative axes across which these texts are evaluated. Finally, we detail the statistical and predictive methods used to compute the relationships and significance of the extracted features.

Multidimensional Analysis

Our framework is designed to capture multiple dimensions of short creative text. We operationalise our analysis by mapping established computational methods across five complementary dimensions: (1) accessibility and complexity, (2) affective tone, (3) novelty, (4) semantic representations, and (5) semantic movement. To quantify these dimensions, we consider the following metrics:

Accessibility and Complexity We measure structural readability using the Flesch Reading Ease (Flesch 1948) and the Coleman-Liau Index (Coleman and Liau 1975). These established psycholinguistic metrics allow us to test whether creativity requires syntactic complexity or if it relies on more accessible phrasing.

Affective Tone We compute Sentiment Polarity, as aesthetic impact is closely tied to emotional resonance (Jacobs and Kinder 2021). We utilise a RoBERTa-based sentiment classifier fine-tuned on social media text (Liu et al. 2019) to extract granular positive and negative probability scores. This metric allows us to assess whether extraordinary language relies on a concentrated emotional charge compared to baseline narrative prose.

Novelty We measure Shannon's Entropy (Shannon 1948), Measure of Textual Lexical Diversity (MTLD) (McCarthy and Jarvis 2010), GPT-2 Surprisal (Radford et al. 2019), and Pointwise Mutual Information (PMI) (Church and Hanks 1990). Measuring departures from expected frequency patterns, often described by Zipf's law (Zipf 1949), these metrics quantify how much a quote deviates from statistically predictable language.

Semantic Representations We aim to test both static, word-level approaches that lack context awareness and sentence-level approaches that represent meaning dynamically and are context-aware. To achieve this comparison, we

utilise word-level GloVe (Pennington, Socher, and Manning 2014), and sentence-level SBERT (Reimers and Gurevych 2019) models to generate spatial semantic representations of our data. Static embeddings (GloVe) are context-free, where each word is assigned a single fixed vector regardless of the surrounding context. In contrast, dynamic embeddings (SBERT) represent meaning at the sentence level, where a word's contribution depends on the surrounding text.

Semantic Movement We calculate Stepwise Distance (SWD) (van Cranenburgh, van Dalen-Oskam, and van Zundert 2019), originally proposed as an index of literariness, to trace the semantic divergence between adjacent sentences. On the word level, we utilise Forward Flow (FF) (Gray et al. 2019), which measures how much a text departs from its own established context by calculating the average semantic distance from all preceding words. These metrics allow us to quantify movement through a semantic space (Szemes and Nagy 2024), providing insights into whether extraordinary language relies on divergent exploration or convergent cohesion.

Experimental Setup and Comparison Axes

To systematically uncover the features of linguistic creativity, we apply our multidimensional metrics across four distinct axes of comparison:

- **Literary vs. Non-Literary:** We contrast our literary quotes (GSQ+GPQ+T50Q) against the non-literary baseline (NLB) to establish the general linguistic and semantic signatures that separate artistic expression from standard informational prose.
- **Within-Book Quoted vs. Non-Quoted Literary Text:** To isolate the specific features that make a sentence quoted, we control for authorial style and genre. We achieve this by comparing quotes (GSQ, GPQ, T50Q) against non-quoted context from the exact same source texts (GBLM, T50LM).
- **High vs. Sampled Popularity:** To understand the drivers of what makes a quote popular, we contrast the most widely shared popular quotes (GPQ) against our standard baseline of reader-curated quotes (GSQ).
- **Contemporary vs. Classic:** Finally, we compare contemporary quotes (GSQ) against classic quotes(T50Q) to observe potential diachronic differences.

Statistical Validation and Explainability

Given the non-normal distributions frequently observed in linguistic data, we use a non-parametric statistical framework to evaluate the significance of feature differences across our distinct text categories. We apply the Mann-Whitney U test (Mann and Whitney 1947) for pairwise comparisons, coupled with Cliff's Delta (Cliff 1993) to measure the effect sizes. Cliff's Delta ranges from -1 to 1: a value of 0 indicates complete overlap between the two groups, while values approaching 1 or -1 indicate that one group almost always scores higher than the other. In the tables, a positive delta indicates that the first (left-hand) group of each comparison scores higher, and a negative delta indicates that the second (right-hand, baseline) group scores higher. We interpret the magnitude using the thresholds of Romano et al. (Romano et al. 2006): $|d| < 0.147$ is negligible, $|d| < 0.33$ is small, $|d| < 0.474$ is medium, and larger values are large. To account for multiple comparisons across our varied metrics, all reported significance levels are adjusted using the Benjamini-Hochberg False Discovery Rate (FDR) procedure (Benjamini and Hochberg 1995).

Complementing this inferential approach, we train extreme gradient-boosted decision trees (XGBoost) (Chen and Guestrin 2016) to identify which specific features best distinguish quotes from their context. We calculate SHAP (SHapley Additive exPlanations) values (Lundberg and Lee 2017) to reveal the directionality and magnitude of individual feature contributions. This method ensures that our findings remain transparent and theoretically grounded.

Results and Discussion

Multidimensional Findings

Evaluating the ALOFT corpus across our five multidimensional metrics and four comparison axes reveals distinct patterns in linguistic creativity. The inferential statistics for these comparisons are detailed in Table 2 and Table 3. Breaking down this analysis across the five core dimensions yields the following insights:

Accessibility and Complexity The non-literary baseline (NLB) shows distinctly higher (harder-to-read) Coleman-Liau Index scores, centred at 11.7, compared to quoted literary text (GSQ, GPQ, T50Q), centred between 5.7 and 6.7, and non-quoted contexts (GBLM, T50LM), centred between 6.7 and 7.7. The same ranking emerges using the Flesch Reading Ease metric. These patterns show that literary language, especially quoted literary text (GSQ, GPQ, T50Q), exhibits a more focused, accessible vocabulary than the non-literary baseline (NLB). Goodreads Popular Quotes (GPQ) are the simplest (≈ 5.7), and the contrast between popular quotes (GPQ) and the non-literary baseline (NLB) produces the largest effect sizes for readability (Flesch Reading Ease $d = 0.74^{***}$, Coleman-Liau $d = -0.81^{***}$), indicating a correlation between popularity and simplicity.

Affective Tone As illustrated in Figure 2, the stacked bars show the proportion of positive, neutral, and negative sentiment across each text type, making the emotional profile of each corpus directly comparable. All quote categories (GSQ, GPQ, T50Q) contain a higher proportion of emotionally valenced content than their corresponding contexts (GBLM, T50LM). While non-literary baseline (NLB) is predominantly neutral, popular quotes (GPQ) exhibit the widest emotional range, with the highest positive sentiment among other quotes (GSQ, T50Q).

Novelty As reported in Table 2 and Table 3, quoted literary text (GSQ, GPQ, T50Q) exhibits lower overall lexical diversity and, notably, lower GPT-2 surprisal scores for contemporary quotes (GSQ, GPQ). This indicates that literary quotes can achieve their impact without relying on rare vocabulary or highly unpredictable sentence structures. Taken

Table 2: Multidimensional Comparison: Literary vs. Non-Literary

	GSQ+GPQ+T50Q vs. LB+NLB	GPQ vs. NLB	GSQ vs. NLB	T50Q vs. NLB
Accessibility, Novelty, and Affective Features (Cliff's Delta)				
Shannon Entropy	-0.08***	-0.23***	0.00	-0.03**
Flesch Reading Ease	0.15***	0.74***	0.51***	0.46***
Coleman–Liau Index	-0.25***	-0.81***	-0.62***	-0.69***
Lexical Diversity	-0.20***	-0.30***	-0.13***	-0.16***
GPT-2 Surprisal	-0.19***	-0.34***	-0.22***	0.03**
Sentiment Polarity	-0.09***	-0.10***	-0.23***	-0.14***
Semantic Representations: SBERT and (GloVe)				
Classifier AUC	0.898 (0.884)	0.995 (0.992)	0.980 (0.973)	0.993 (0.991)
Semantic Outlier (d)	-0.384 (-0.134)	-0.904 (-0.428)	-0.711 (-0.317)	-0.851 (-0.382)
Centroid Distance	0.1414 (0.0067)	0.6128 (0.0437)	0.4537 (0.0289)	0.4831 (0.0302)
Silhouette Score	0.024 (0.048)	0.073 (0.191)	0.051 (0.142)	0.081 (0.153)
Semantic Movement (Cliff's Delta)				
Forward Flow (FF)	-0.456***	-0.824***	-0.730***	-0.775***
Stepwise Dist. (SWD)	-0.164***	0.079***	0.108***	0.006

Note: Significance from Mann–Whitney U test (Benjamini–Hochberg corrected): * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Semantic Outlier (d) is Cliff's Delta from nearest-neighbour distance comparison. Centroid Distance is the cosine distance between class mean vectors. Silhouette Score is a cluster-separation measure. LB denotes the Literary Baseline, the combination of the non-quoted literary contexts GBLM and T50LM (see Table 1).

Table 3: Multidimensional Comparison: Quoted vs. Non-Quoted, High vs. Sampled Popularity, and Contemporary vs. Classic

	Within-Book		Popular	Contemporary
	GSQ vs. GBLM	T50Q vs. T50LM	GPQ vs. GSQ	GSQ vs. T50Q
Accessibility, Novelty, and Affective Features (Cliff's Delta)				
Shannon Entropy	-0.03**	0.00	-0.22***	0.03*
Flesch Reading Ease	-0.10***	-0.20***	0.28***	0.08***
Coleman–Liau Index	0.00	0.11***	-0.30***	0.05***
Lexical Diversity	-0.19***	-0.13***	-0.18***	0.03*
GPT-2 Surprisal	-0.12***	-0.11***	-0.15***	-0.26***
Sentiment Polarity	-0.10***	-0.06***	0.09***	-0.06***
Semantic Representations: SBERT and (GloVe)				
Classifier AUC	0.852 (0.831)	0.841 (0.840)	0.761 (0.745)	0.826 (0.845)
Semantic Outlier (d)	-0.292 (-0.052)	-0.208 (-0.048)	-0.213 (-0.034)	-0.031 (-0.040)
Centroid Distance	0.1355 (0.0045)	0.0709 (0.0038)	0.0470 (0.0025)	0.0370 (0.0015)
Silhouette Score	0.021 (0.036)	0.018 (0.029)	0.008 (0.021)	0.012 (0.013)
Semantic Movement (Cliff's Delta)				
Forward Flow (FF)	-0.244***	-0.216***	-0.235***	0.043***
Stepwise Dist. (SWD)	-0.211***	-0.322***	-0.037*	0.115***

Note: Significance from Mann–Whitney U test (Benjamini–Hochberg corrected): * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Semantic Outlier (d) is Cliff's Delta from nearest-neighbour distance comparison. Centroid Distance is the cosine distance between class mean vectors. Silhouette Score is a cluster-separation measure. Delta-PMI was excluded from inferential analysis due to circularity.

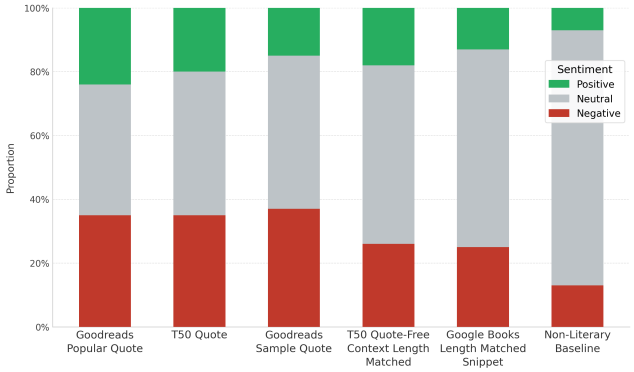


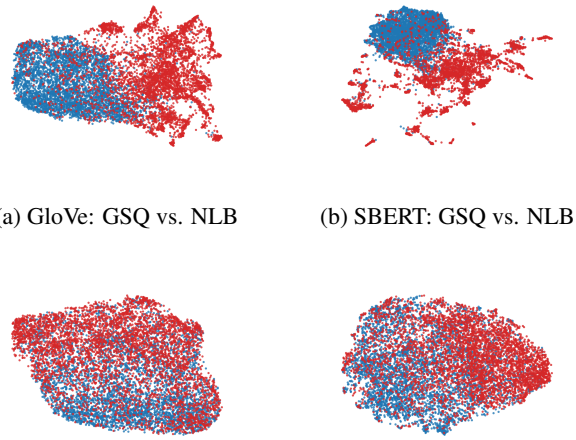
Figure 2: Proportions of sentiment labels across text types. Quotes (GSQ, GPQ, T50Q) show higher proportions of emotionally valenced content (positive and negative) compared to non-quoted contexts (GBLM, T50LM), which contain more neutral content. The non-literary baseline shows the highest proportion of neutral sentiment (≈80%)

together with our findings on accessibility and affective tone, these results demonstrate that literary quotes are not necessarily complex or surprising, but they exhibit accessible delivery with a concentrated emotional charge.

Semantic Representations To examine the semantic geometry of our data, we compare static (GloVe) and dynamic (SBERT) embeddings. Inferred from the classifier AUC results from Table 2, both SBERT and GloVe models perform accurately while separating quoted literary text (GSQ+GPQ+T50Q) from the non-literary baseline (NLB). The accuracy of both models decreases in quoted vs. non-quoted, high vs. sampled popularity, and contemporary vs. classic comparisons, shown in Table 3. In general, we observe that SBERT consistently outperforms GloVe, except when separating popular from sampled Goodreads quotes. SBERT successfully identifies quoted literary text as semantic outliers relative to the non-literary baseline, with a large effect size ($d = -0.904$ for popular quotes), outperforming the static GloVe model ($d = -0.428$).

Furthermore, as visualised in Figure 3, the context-aware SBERT model outperforms the non-contextual GloVe model in separating Goodreads Sample Quotes (GSQ) from Google Books length-matched (GBLM) snippets and non-literary baseline (NLB). These results indicate that the creativity of quoted literary texts can be captured more comprehensively when structural arrangement and context are taken into account.

Semantic Movement As detailed in Table 3, both contemporary (GSQ, GPQ) and classic (T50Q) quotes exhibit lower FF ($d = -0.244$ and $d = -0.216$, respectively) and lower SWD ($d = -0.211$ and $d = -0.322$, respectively) than their non-quoted literary contexts (GBLM, T50LM). This observation suggests that quoted passages are characterised by a more constrained and coherent semantic trajectory compared to the more exploratory nature of their non-quoted contexts. The comparison against baseline texts shows a



(a) GloVe: GSQ vs. NLB (b) SBERT: GSQ vs. NLB

(c) GloVe: GSQ vs. GBLM (d) SBERT: GSQ vs. GBLM

Figure 3: UMAP projections for Goodreads Sample Quotes (GSQ) compared to different corpora. The top row shows the Non-Literary baseline (NLB), where GloVe (a) exhibits some structure but SBERT (b) achieves a much clearer class separation. The bottom row shows the quotes versus their corresponding context (GBLM), where GloVe (c) fails to separate the classes, whereas SBERT (d) reveals a discernible, albeit overlapping, structure.

similar pattern. Each quote (GSQ, GPQ, T50Q) had substantially lower FF scores than the non-literary baseline (NLB), with effect sizes (d between -0.730 and -0.824) indicating a strong tendency toward conceptual focus in literary language.

The SWD results show a more nuanced pattern. While classic quotes (T50Q) show no apparent difference from the NLB, modern literary quotes (GPQ, GSQ) exhibit a slight but statistically robust increase in SWD (d = 0.079 and d = 0.108, respectively; $p < 0.001$). These reflect the trade-off between exploration and exploitation, which maps onto the distinction between divergent and convergent creativity. Our results suggest that quotes are products of convergent creativity, exploiting a specific theme through a focused semantic path. In contrast, the non-quoted literary contexts (GBLM, T50LM) explore a wider range of topics.

Classifier Results and Feature Importance

Combining the metrics we implemented during this work, we conducted a predictive modelling analysis to quantify the relative importance of the linguistic features.

Our XGBoost models exhibited varied yet informative performance across the five classification tasks, confirming that each comparison group has a distinct and measurable linguistic fingerprint. A summary of the key evaluation metrics for each model, including precision and recall for the positive (quote) class, is presented in Table 4. The results show a clear hierarchy of distinguishability. The model most effectively separated literary quotes from a non-literary baseline (GSQ vs. NLB), achieving an AUC of 0.93 with strong precision (0.86) and recall (0.89). The next most distinguishable comparison involved different types of literary text: Goodreads Sample Quotes versus T50 Quotes (AUC = 0.74), and quotes versus their surrounding narrative context (AUC = 0.70 for both GSQ and T50Q). The most nuanced comparison was between popular and regular sample quotes (GSQ vs. GPQ, AUC = 0.69), which also showed the most balanced trade-off between precision (0.60) and recall (0.59), suggesting that the features driving social resonance are the most subtle.

Beyond raw predictive performance, interpreting the specific features that distinguish literary creativity is essential. Extracting these feature contributions across our four comparative axes yields the following insights:

- **Literary vs. Non-Literary (GSQ vs. NLB):** The model relies heavily on semantic structure and emotion. Low FF and high positive/negative polarity indicate that literary texts are more cohesive and emotionally expressive than everyday prose.
- **Quoted vs. Non-Quoted (GSQ vs. GBLM):** Figure 4 presents a SHAP summary plot for the GSQ versus Context classification task, illustrating both the magnitude and directionality of individual feature contributions to the model's predictions. The most essential features driving the prediction are low sentiment neutrality, low SWD, and low FF. This indicates that, compared to the surrounding narrative, quotes are significantly more emotionally expressive and follow a more dynamic, yet coherent, semantic path.
- **High vs. Sampled Popularity (GSQ vs. GPQ):** Positive sentiment alongside lower Coleman-Liau scores (simpler language) predicts popularity. This suggests that widely shared quotes are more accessible and positive.
- **Contemporary vs. Classic (GSQ vs. T50Q):** The model overwhelmingly prioritises unexpectedness. Low GPT-2 surprisal strongly predicts the classic, curated T50 quotes, indicating less predictable word choices than modern Goodreads quotes. However, this can also be explained by the architecture of the GPT-2 model. Because it was predominantly trained on public-domain data, T50 quotes might have been flagged as less surprising than contemporary works that were not present in the training data.

Table 4: Summary of XGBoost Classifier Performance Across Experimental Tasks.

Experimental Task	AUC	Accuracy	F1-Score	Precision	Recall
GSQ vs. NLB	0.93	86.7%	0.87	0.86	0.89
GSQ vs. T50Q	0.74	67.9%	0.67	0.70	0.75
GSQ vs. Context	0.70	64.4%	0.64	0.63	0.55
T50Q vs. Context	0.70	63.9%	0.63	0.59	0.55
GSQ vs. GPQ	0.69	63.8%	0.63	0.60	0.59

Note: AUC is Area Under the ROC Curve. F1-Score is the macro average. Precision and Recall are for the positive quote class.

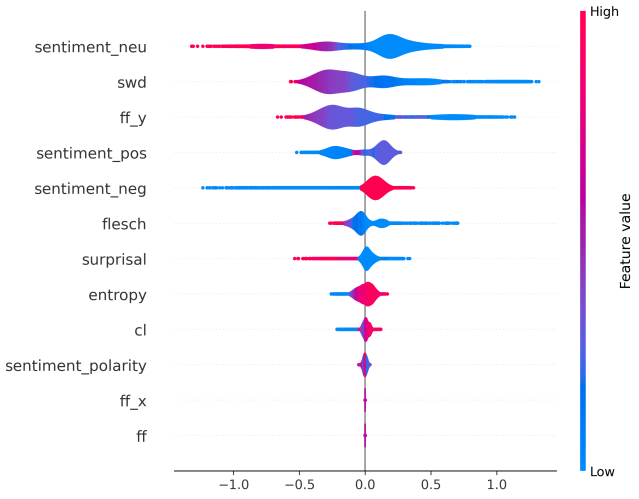


Figure 4: SHAP summary plot for the Goodreads Sample Quotes (GSQ) vs. Google Books Length-Matched (GBLM) Snippet classification task. Each point represents the SHAP value of a single text for a given feature. The colour indicates the feature's value (red is high, blue is low), and its position on the x-axis indicates its impact on the prediction. Key predictors are low sentiment neutrality, low SWD, and low FF.

Conclusions and Future Work

In this paper, we share ALOFT, a curated dataset that aligns literary quotes with length-matched context from the same books, public-domain parallels, and a non-literary baseline, providing the community with an extendable, controlled testbed for creativity research. ALOFT is constructed through a human-in-the-loop workflow in which automation (APIs, page capture, OCR, fuzzy matching) is continuously audited and corrected via manual checkpoints with custom interfaces, ensuring data quality without breaching platform ToS or academic fair use. We used ALOFT to conduct a multidimensional analysis of creativity, employing an integrated framework that combines established metrics to explore the features of creative artifacts and identify what distinguishes them from the ordinary. Coupling XGBoost classifiers with SHAP explanations enabled ranking the salience and directionality of our features, providing interpretable evidence on how they interact to create these creative artifacts.

While our study provides a comprehensive analysis of creative literary quotes, it is not without limitations. First, the analysis is constrained by the nature of the data itself. Despite our length-matching efforts, some quotes are short, which can make length-sensitive metrics such as surprisal more variable. Our dataset reflects the cultural and demographic biases of its source, the predominantly Western, English-speaking Goodreads community. In terms of temporal scope, the data are biased toward English-language prose published after 1800, so these findings may not apply to poetry, drama, or other cultures' literary traditions. Additionally, the popularity of certain authors can be a confounding factor, as it mixes stylistic signals with the author's reputation. Our evaluation also relies on Goodreads endorsements as a proxy for quality, without a ground truth based on expert literary-critical annotation.

Reader endorsement operationalises the value component of creativity rather than creativity as a whole. In our corpus, popularity correlates with accessible vocabulary and positive sentiment, signatures more characteristic of likability than of novelty. This is consistent with theoretical decompositions of creativity into novelty and value (Boden 2004), where value is partly constituted by recognition from an audience rather than by properties of the artifact alone. Goodreads endorsement is therefore a signal for the value dimension, while novelty is captured separately through surprisal, lexical diversity, and semantic movement.

Second, the metrics and models we employed to represent different dimensions of our corpora were not specifically trained on literary language and may not fully capture its nuances.

Our paper triggers numerous future research directions. One possible next step is to validate our results against human perception through controlled participant studies and expert annotation. While costly and time-consuming, human involvement would yield a higher-quality, curated dataset to serve as ground truth. Similarly, validating our results against LLM evaluation would yield novel insights and comparative studies.

Another exciting direction is to model intra-textual quotations, analysing all candidate quotes within a single work to understand how their prominence relates to narrative position and thematic structure. Linking the texts to cultural metadata, such as author demographics or book genre, could also reveal sociocultural and temporal factors that influence memorability.

Following the dynamic semantic distance approaches of FF and SWD, the same "journey" analogy can be extended to include emotional features, as the sentiment contribution of each word or sentence may also yield relevant insights. This approach can also be enriched by part-of-speech analysis to examine how word order and the distribution of grammatical roles influence the structure and flow of information. To capture more nuance, incorporating metrics of figurative language would also be valuable.

Performance-wise, fine-tuning transformer models with relevant data could improve their sensitivity to literary language, and developing length-normalised versions of our trajectory metrics would make them more robust for short texts. A similar path forward involves using our findings as a blueprint to train a generative model, then evaluating its output in human preference experiments to determine whether this "recipe" for quotability is sufficient to create genuinely memorable language.

Advancing our understanding of what distinguishes creative from ordinary language is an open challenge in computational creativity. Our initial analyses, which drew on both well-established metrics and more recent dynamic approaches, yielded promising insights into patterns in creative literary language.

Our work emphasises the importance of explainability for understanding the contributions of features across dimen-

sions to the building blocks of highly-valued literature. In doing so, we also show that information-theoretic and static metrics, more frequently set aside in favour of black-box approaches, still contribute to our understanding of linguistic creativity. We demonstrate that a fuller picture emerges only when these dimensions are combined, ranging from straightforward readability indices to dynamic semantic trajectories to surprisal estimates of contemporary language-models. Together, they reveal that creativity is not the product of any single feature but of an interplay across multiple levels and dimensions of language. This also supports the earlier point we drew from Boden that creativity is more than novelty (Boden 2004), since novelty alone was not enough to distinguish quotes from ordinary text in our results.

As such, this paper contributes to three intersecting fields: the computational assessment of creativity, data-driven analysis of literary artifacts, and the generation of creative and memorable language. To build on our insights, further research should focus on diverse explainable features that contribute to both novelty and value, incorporating richer dimensions such as figurative language, narrative structure, and affective trajectories, validated against human perception and expert annotation across a broader range of literary traditions and cultural contexts.

Acknowledgements

This research was funded by the project "Hybrid Intelligence: Augmenting Human Intellect", a 10-year Gravitation programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research (grant number 024.004.022). We thank the anonymous reviewers of ICCC 2026 for their careful reading and constructive feedback, which improved the clarity and framing of this paper.

References

Ashok, V. G.; Feng, S.; and Choi, Y. 2013. Success with style: Using writing style to predict the success of novels. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1753–1764.

Barthes, R. 1967. The death of the author. *Aspen* (5-6). Reprinted in Image-Music-Text, 1977, Fontana Press.

Beaty, R. E., and Johnson, D. R. 2021. Automating creativity assessment with semdis: An open platform for computing semantic distance. *Behavior Research Methods* 53(2):757–780.

Beaty, R. E.; Johnson, D. R.; Zeitlen, D. C.; and Forthmann, B. 2022. Semantic distance and the alternate uses task: Recommendations for reliable automated assessment of originality. *Creativity Research Journal* 34(3):245–260.

Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57(1):289–300.

Boden, M. A. 2004. *The creative mind: Myths and mechanisms*. Psychology Press.

Bunescu, R. C., and Uduehi, O. O. 2022. Distribution-based measures of surprise for creative language: Experiments with humor and metaphor. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, 68–78. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Chen, T., and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. ACM.

Church, K. W., and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1):22–29.

Cliff, N. 1993. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological bulletin* 114(3):494.

Coleman, M., and Liau, T. L. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology* 60(2):283–284.

Danescu-Niculescu-Mizil, C.; Cheng, J.; Kleinberg, J.; and Lee, L. 2012. You had me at hello: How phrasing affects memorability. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 892–901.

Dumas, D.; Organisciak, P.; and Doherty, M. 2021. Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts* 15(4):645–663.

faellielupe. 2020. Goodreads quotes dataset. Kaggle dataset. Accessed on 2025-05-05.

Flesch, R. 1948. A new readability yardstick. *Journal of Applied Psychology* 32(3):221–233.

Francis, W. N., and Kučera, H. 1982. *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin Company.

Gerlach, M., and Font-Clos, F. 2018. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics.

Gray, K.; Anderson, S.; De-Arteaga, M.; Holtzman, N. M.; and Pennebaker, J. W. 2019. Exploring the 'extraordinary': A computational analysis of creative language. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2486–2495.

He, H.; Peng, N.; and Liang, P. 2019. Pun generation with surprise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1734–1744. Minneapolis, Minnesota: Association for Computational Linguistics.

Hipson, W., and Mohammad, S. M. 2020. PoKi: A large dataset of poems by children. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*,

1578–1589. Marseille, France: European Language Resources Association.

Internet Movie Database (IMDb). Internet movie database (imdb).

Iser, W. 1978. *The Act of Reading: A Theory of Aesthetic Response*. Baltimore: Johns Hopkins University Press.

Ismayilzada, M.; Paul, D.; Bosselut, A.; and van der Plas, L. 2024. Creativity in ai: Progresses and challenges. Preprint.

Ismayilzada, M.; Stevenson, C.; and van der Plas, L. 2024. Evaluating creative short story generation in humans and large language models. *arXiv preprint arXiv:2411.02316*.

Jacobs, A. M., and Kinder, A. 2021. Computational analyses of the topics, sentiments, literariness, creativity and beauty of texts in a large corpus of english literature. *Language and Literature*. Preprint or in press version; Final publication venue may vary.

Jacobs, A. M. 2015. Neurocognitive poetics: Methods and models for investigating the neuronal and cognitive–affective bases of literature reception. *Frontiers in Human Neuroscience* 9:186.

Jacobs, A. M. 2018. The gutenberg english poetry corpus: Exemplary quantitative narrative analyses. *Frontiers in Digital Humanities* 5:5.

Kuznetsova, P.; Chen, J.; and Choi, Y. 2013. Understanding and quantifying creativity in lexical composition. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1246–1258. Seattle, Washington, USA: Association for Computational Linguistics.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lundberg, S. M., and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, 4765–4774.

Maharjan, S.; Arevalo, J.; González, F. A.; Montes-y Gómez, M.; and Solorio, T. 2017. A multi-task approach to predict likability of books. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 1217–1227. Valencia, Spain: Association for Computational Linguistics.

Mann, H. B., and Whitney, D. R. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* 18(1):50–60.

Mao, R.; Lin, C.; and Guerin, F. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3888–3898. Florence, Italy: Association for Computational Linguistics.

Marco, G.; Rello, L.; and Gonzalo, J. 2024. Small language models can outperform humans in short creative writing: A study comparing slms with humans and llms. *arXiv preprint arXiv:2409.11547*.

McCarthy, P. M., and Jarvis, S. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. In *Behavior Research Methods*, volume 42, 381–392. Springer.

Organisciak, P.; Acar, S.; Dumas, D.; and Berthiaume, K. 2023. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity* 49:101356.

Orwig, W. D.; Diez, I.; Faskowitz, J.; Medaglia, J. D.; Bassett, D. S.; and Beaty, R. E. 2021. Creative connections: Computational semantic distance captures individual creativity and resting-state functional connectivity. *NeuroImage* 227:117632.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Porter, B., and Machery, E. 2024. Ai-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports* 14(1):12345–12356.

Potthast, M.; Rangel, F.; Tschuggnall, M.; Stamatatos, E.; Rosso, P.; and Stein, B. 2018. Overview of the author identification task at pan 2018: Cross-domain authorship attribution. In *Working Notes of CLEF*.

Priyanshu, A., and Vijay, S. 2024. The silent curriculum: How does llm monoculture shape educational content and its accessibility?

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1(8):9.

Reimers, N., and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*.

Romano, J.; Kromrey, J. D.; Coraggio, J.; and Skowronek, J. 2006. Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen's d for evaluating group differences on the NSSE and other surveys? In *Annual Meeting of the Florida Association of Institutional Research*.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27(3):379–423.

Sontag, S. 1966. *Against Interpretation and Other Essays*. New York: Farrar, Straus and Giroux.

Summers, J. S. 2017. Post hoc ergo propter hoc: Some benefits of rationalization. *Philosophical Explorations* 20(sup1):66–79.

Szemes, B., and Nagy, M. 2024. Repetition and innovation in dramatic texts: An attempt to measure the degree of novelty in character's speech. *Journal of Computational Literary Studies* 3(1).

Tekir, S.; Güzel, A.; Tenekeci, S.; and Haman, B. U. 2023. Quote detection: A new task and dataset for nlp. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2023)*, 21–27. Dubrovnik, Croatia: Association for Computational Linguistics.

Tsvetkov, Y.; Mukomel, R.; and Gershman, A. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 248–258.

van Cranenburgh, A., and Bod, R. 2017. A data-oriented model of literary language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 1228–1238. Valencia, Spain: Association for Computational Linguistics.

van Cranenburgh, A.; van Dalen-Oskam, K.; and van Zundert, J. 2019. Vector space explorations of literary language. *Language Resources and Evaluation* 53(4):625–650.

Verma, A. 2021. Kaggle goodreads quotes dataset. Accessed on 2025-05-05.

Wikimedia Foundation. 2023. Wikipedia database dump. Available at `https://dumps.wikimedia.org/enwiki/20231101/`.

Wu, F.; Black, E.; and Chandrasekaran, V. 2024. Generative monoculture in large language models.

Zipf, G. K. 1949. *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley Press.