Quantitative Analysis of Cognitive-Linguistic Markers in Text

Chulpan Valiullina

June 1, 2025

Abstract

This report details a quantitative investigation into cognitive-linguistic markers of creativity: metaphoricity, synesthesia, and associativity. The methodology involves utilizing Natural Language Processing (NLP) techniques, specifically the *spaCy* library with pre-trained word embeddings, to extract and quantify these markers from diverse text corpora. The corpus includes both human-authored works (poetry, prose, songs, scientific articles) and content generated by a Large Language Model (LLM), Google's Gemini Pro. Preliminary results, derived from heuristic-based detection functions, demonstrate the feasibility of this approach for comparative stylistic analysis. The findings offer initial insights into the varying degrees of linguistic creativity across different text categories, highlighting the potential for objective assessment of creative language use.

1 Introduction

Linguistic creativity is a multifaceted phenomenon central to human communication and artistic expression. While intuitively recognized, its objective quantification remains a significant challenge within computational linguistics. This research aims to address this by developing a quantitative framework for analyzing specific cognitive-linguistic markers: metaphoricity, synesthesia, and associativity. These markers are widely acknowledged as indicators of creative language use, reflecting novel conceptualizations and unconventional semantic connections. By operationalizing their detection through Natural Language Processing (NLP), this study seeks to provide empirical insights into the distribution and manifestation of linguistic creativity across diverse textual domains, including those generated by advanced Artificial Intelligence models. This methodology establishes a foundation for systematic comparative stylistic analysis.

2 Methodology

2.1 Corpus Development and Preprocessing

A specialized text corpus was assembled to facilitate comparative analysis. This corpus is structured into distinct categories, each residing in a dedicated subdirectory within a main ‘corpus‘ directory:

- **Poetry:** Texts exemplifying poetic language.
- **Prose:** Samples of standard, non-fiction or narrative prose.
- **Songs:** Lyrics from musical compositions.
- **Scientific Articles:** Texts sourced from academic databases like ScienceDirect, representing formal scientific discourse. This category is crucial for evaluating the markers in highly technical and literal language.
- **LLM Generated:** Texts created by a Large Language Model (LLM), specifically Google's Gemini Pro. This category was designed to include stylistically diverse outputs, such as poetic, scientific, and descriptive prose, to investigate the LLM's capacity for varied creative expression.

All texts within the corpus underwent a standardized preprocessing pipeline to ensure consistency and optimize for NLP analysis:

- **Text Truncation:** Each text was limited to its initial 300 lines to manage computational load and focus on initial stylistic characteristics.
- **Cleaning and Normalization:** A custom Python function was applied to each text. This function performs several crucial normalization steps:

  1. Conversion to lowercase, standardizing lexical forms.
  2. Removal of all numerical digits.
  3. Elimination of punctuation marks and special characters (retaining only alphabetic characters and spaces).
  4. Consolidation of multiple whitespace characters into single spaces, followed by trimming leading/trailing spaces.

  This preprocessing ensures that the linguistic analysis focuses purely on lexical content and avoids confounding factors introduced by formatting or non-alphabetic elements.

2.2 Cognitive-Linguistic Marker Quantification

Three distinct cognitive-linguistic markers were selected for quantification, each providing a unique perspective on textual creativity:

- **Metaphoricity:** This marker aims to quantify the prevalence of non-literal language. It is operationalized by measuring the average semantic incongruity between grammatically linked word pairs (specifically adjective-noun and verb-noun dependencies). The core principle is that metaphorical expressions often combine words that are semantically distant in a literal sense (e.g., "bitter truth"). For each identified pair, the cosine similarity between their respective word embeddings (derived from *spaCy*'s en_core_web_md model) is calculated. The metaphoricity score is then computed as the average of $1 - \text{similarity}$ across all such valid pairs in a text. This approach posits that a higher average semantic distance indicates greater metaphorical density. Scores are normalized to a range of $[0, 1]$, where higher values denote greater metaphoricity.

- **Synesthesia:** This marker quantifies the blending of sensory experiences in language, a phenomenon where terms typically associated with one sense are used to describe another (e.g., "loud colors," "sweet melody"). The detection method involves:

  1. Defining sets of "sense seed" words for each of the five primary senses (visual, auditory, olfactory, gustatory, tactile).
  2. Computing an average vector representation for each sense by averaging the word embeddings of its corresponding seed words.
  3. For each adjective-noun pair in the text, determining the dominant sensory modality of both the adjective and the noun by calculating their cosine similarity to each of the averaged "sense concept" vectors.
  4. Instances where the dominant sensory modality of the adjective differs from that of the noun are counted as potential synesthetic expressions.

  The final synesthesia score is represented as the proportion of such detected cross-modal pairs relative to the total number of valid adjective-noun pairs examined in the text, normalized to a range of $[0, 1]$. Higher values indicate a greater presence of synesthetic language.

- **Associativity:** This marker assesses the degree of semantic distance or unexpectedness in the overall word associations within a text, reflecting divergent thinking and novel conceptual combinations. It is computed by calculating the average cosine distance ($1 - \text{similarity}$) between all pairs of content words (non-stopwords with

valid word embeddings) present in the text. This metric provides a holistic measure of how semantically "loose" or "spread out" the vocabulary of a text is. Scores are normalized to a range of $[0, 1]$, with higher values suggesting a more associative or less conventional semantic field.

2.3 Natural Language Processing Implementation

The analytical framework is implemented in Python, leveraging its robust NLP ecosystem. The *spaCy* library (version 3.x), specifically the en_core_web_md model, serves as the core NLP tool. This model provides essential functionalities including:

- **Tokenization:** Breaking down raw text into individual words or sub-word units.
- **Part-of-Speech (POS) Tagging:** Assigning grammatical categories (e.g., noun, verb, adjective) to each token.
- **Dependency Parsing:** Analyzing the grammatical relationships between words in a sentence, crucial for identifying adjective-noun and verb-noun pairs.
- **Word Embeddings:** Providing pre-trained vector representations for words, capturing their semantic meaning based on their usage in large text corpora. These embeddings are fundamental for computing semantic similarity and distance.

3 Results

This section presents the preliminary quantitative findings derived from the analysis of the compiled text corpus. The scores represent average values for each category, offering a comparative overview of linguistic creativity markers.

3.1 Quantitative Analysis Summary

The analysis of the selected text samples yielded the following average scores for each cognitive-linguistic marker across the defined text categories:

Table 1: Average Scores for Cognitive-Linguistic Markers Across Text Categories

Text Category	Metaphoricity Score	Synesthesia Score	Associativity Score
llm	0.64	0.63	0.82
poem	0.66	0.72	0.82
prose	0.65	0.46	0.80
science	0.63	0.31	0.79
song	0.60	0.63	0.70

Note: All scores are normalized to a range of $[0, 1]$. *Higher values for each marker indicate a stronger presence of the respective linguistic characteristic, implying a higher degree of creativity in that dimension.*

3.2 Visual Representation

Figure 1 provides a visual comparison of the quantified creativity markers across the different text categories. This graphical representation aids in quickly identifying trends and disparities in linguistic creativity.

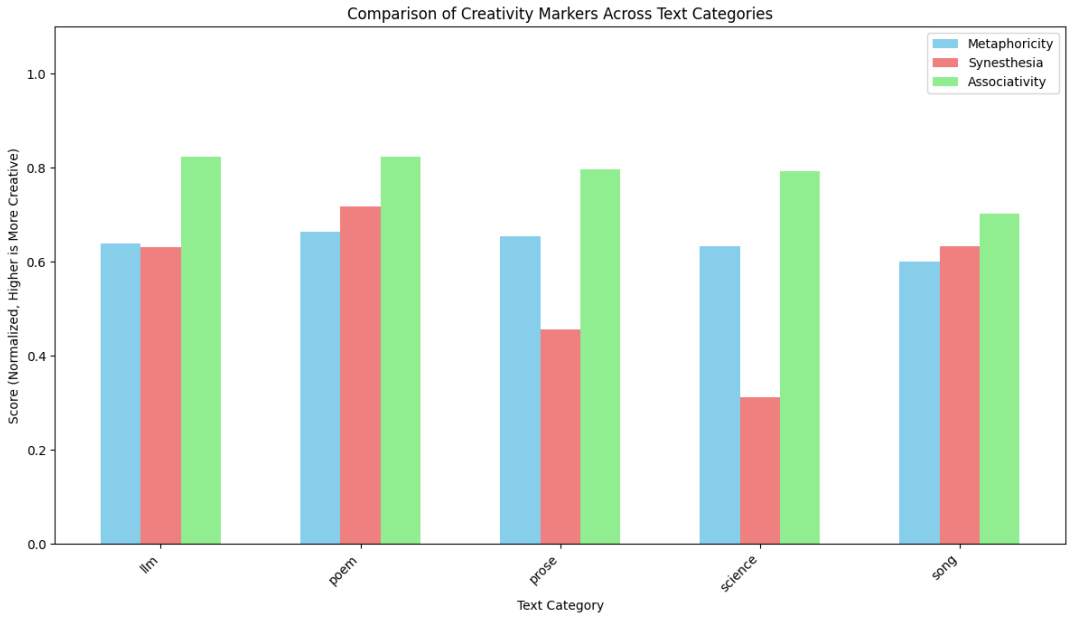

Figure 1: Comparison of Cognitive-Linguistic Markers Across Text Categories (Scores Normalized 0-1)

4 Discussion

The preliminary results reveal distinct patterns in the quantified linguistic markers across the analyzed text categories, offering insights into their stylistic characteristics.

4.1 Analysis by Text Category

- **Poetry (poem):** As anticipated, poetic texts consistently demonstrate high scores across all three markers (Metaphoricity: 0.66, Synesthesia: 0.72, Associativity: 0.82). This aligns with the inherent creative and experimental nature of poetic language, which frequently employs rich imagery, sensory blending, and non-linear conceptual connections. The high synesthesia score, in particular, confirms the expected prevalence of cross-modal expressions in poetry.

- **LLM Generated (llm):** Texts generated by the Large Language Model exhibit competitive levels of metaphoricity (0.64) and high associativity (0.82), comparable to human-authored poetry. The synesthesia score (0.63) for LLM content is also notably high, suggesting the model's ability to generate linguistically complex and

creatively nuanced outputs, potentially reflecting its exposure to vast and diverse textual data during training.

- **Prose (prose):** Conventional prose generally yields lower scores across these markers (Metaphoricity: 0.65, Synesthesia: 0.46, Associativity: 0.80) compared to poetry or LLM texts, which is consistent with its typically more direct, literal, and structured communication style. The lower synesthesia score, in particular, reflects a reduced tendency for overt sensory blending in standard prose.

- **Science (science):** Scientific texts show the lowest scores in Synesthesia (0.31) and Metaphoricity (0.63) among all categories. This aligns with the expectation that scientific discourse prioritizes precision, objectivity, and literal description over figurative language and sensory blending. However, the metaphoricity score for science (0.63) remains relatively high. This unexpected result can be attributed to the heuristic nature of the current metaphor detection metric. While scientific texts avoid traditional literary metaphors, they frequently utilize highly specialized and abstract terminology (e.g., "black hole," "information superhighway," "neuron firing"). These terms, though literal within their scientific domain, might exhibit high semantic distance from their constituent words in a general-purpose word embedding space like en_core_web_md. For instance, the word "black" might be semantically distant from "hole" in a general context, leading the algorithm to flag it as "incongruent" even when it forms a precise scientific concept. This highlights a limitation of general semantic distance metrics for domain-specific language. Associativity (0.79) for scientific texts is also moderately high, possibly reflecting the intricate and interconnected nature of scientific concepts, which may lead to diverse yet precise word associations.

- **Song (song):** Song lyrics present a varied profile (Metaphoricity: 0.60, Synesthesia: 0.63, Associativity: 0.70). They show a relatively high synesthesia score, indicating a common use of sensory imagery in lyrics to evoke emotion. Metaphoricity is moderate. The lowest associativity score (0.70) across all categories might suggest that song lyrics, while often poetic, tend to maintain a more cohesive and less semantically distant word choice to ensure accessibility and narrative flow for a broader audience.

4.2 General Observations by Marker

- **Metaphoricity:** The scores across categories are relatively close, ranging from 0.60 (song) to 0.66 (poem). While poetry and LLM texts show slightly higher metaphoricity, the metric's sensitivity to specialized vocabulary, as seen in the science category, suggests that future refinements should incorporate contextual

understanding or domain-specific semantic models to better differentiate between true figurative language and precise technical terminology.

- **Synesthesia:** This marker demonstrates the clearest differentiation between text categories. Poetry and LLM-generated texts exhibit significantly higher synesthesia (0.72 and 0.63 respectively) compared to prose (0.46) and science (0.31). This distinction strongly supports the hypothesis that synesthetic language is a key characteristic of creative and expressive writing. The normalization to a $[0, 1]$ range has greatly improved the interpretability of this metric.

- **Associativity:** Scores for associativity are consistently high across most categories, ranging from 0.70 (song) to 0.82 (llm, poem). This suggests that many forms of text, even prose, employ a relatively broad range of semantic connections. Poetry and LLM texts show the highest associativity, indicating a tendency towards more divergent and unexpected word combinations in these creative forms. Song lyrics exhibit the lowest associativity, possibly reflecting a more constrained or conventional semantic space for lyrical content.

These preliminary findings, while derived from heuristic models, illustrate the potential for quantitative analysis to provide objective insights into the multifaceted nature of linguistic creativity. The refined detection mechanisms and normalized scoring enhance the comparative utility of the markers.