Quantitative Framework for Word-Color Association and Application to 20th Century Anglo-American Poetry

Sungpil Wang[1], Juyong Park[1,*]

[1] *Graduate School of Culture Technology (GSCT), Korea Advanced Institute of Science & Technology (KAIST), Daejeon, Republic of Korea*

Abstract

Color symbolism is considered a critical element in art and literature, yet determining the relationship between colors and words has remained largely subjective. This research presents a systematic methodology for quantifying the correlation between language and color. We utilize text-based image search, optical character recognition (OCR), and advanced image processing techniques to establish a connection between words and their corresponding color distributions in the CIELch color space. We generate a color dataset based on human cognition, and apply it for analysis of the literary works of poets associated with Imagism and Black Arts Movements. This helps uncover the characteristic color patterns and symbolic meanings of the movements with enhanced objectivity and reproducibility in literature research. Our work has the potential to provide a powerful instrument for a systematic, quantitative examination of literary symbolism, filling in the gaps in prior analyses and facilitating novel investigations of thematic aspects using color.

Keywords

Word Color Association, Digital Humanities, Distant Reading, Semantic Analysis, Lexical Relationship, Lexical Discovery

1. Introduction

Color symbolism in art, literature, and anthropology refers to the use of colors as symbols in various cultural contexts [13]. This effect is particularly pronounced in literature, where colors assume roles that extend beyond being mere background elements. Researchers have deeply investigated the application of colors in texts and their significance in literary works. For instance, Andreeva [1] examines Stephen King's "Rose Madder" to show how color depicts characters' emotional and psychological states, providing nuanced moral or philosophical implications. Mukhitdinovna [48] explores the linguistic aspects of color symbolism in novels, highlighting how color characterizes objects, social attitudes, and moral concepts. This indicates that colors convey complex meanings rather than simply serve as decorative elements. By choosing specific colors, authors can emphasize emotions or concepts related to the story's theme, making color an essential tool for literary expression. Additionally, color reflects cultural, historical, and practical meanings [32]. Polshchykova and Polshchykova [52]

CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark

*Corresponding author.

wangsp0317@kaist.ac.kr (S. Wang); juyongp@kaist.ac.kr (J. Park)

0000-0003-4571-0017 (J. Park)

demonstrates how color in works from various races and cultures illustrates the diversity and complexity of American culture, showing the colors' different meanings in different contexts. These studies explore the intricate circumstances surrounding the use of color, linking literary works to the external environment. Furthermore, Underwood [66] reviews the growing importance of color terminology in literature from the 18th to the early 20th centuries, highlighting how the trend has intensified over time. This evolution indicates a transition from theoretical descriptions to tangible depictions, with an emphasis on sensory details. It signifies the increasing focus on sensory portrayal in writings as a means to mirror shifts in society and culture. The heightened use of color underscores its ability to enhance narrative complexity and emotional depth, making this aspect increasingly significant in modern literary analysis. To summarize, color symbolism is an important literary technique that enriches the intricacy of literary works, enabling readers to engage more profoundly with the text. This technique has become more prominent in modern times; This amplifies the visual, emotional, cultural, and societal significance of literary works, both inside the text itself and in a wider context.

Color representation in literary works is generally executed in two forms: Directly describing color using adjectives associated with certain hues [56, 67] or indirectly evoking a particular color through noun imagery [65, 33]. Researchers encounter two obstacles when assigning color to specific phrases for analysis: First, term-image linkage is subjective. The researcher's common-sense perception, which depends on "objectivity and universality," as expected by the researcher, is frequently used to determine the associated colors of nouns [26, 25]. For instance, a researcher might define a tree as green, expecting readers to accept this without question. But there cannot be a one-to-one correlation between words and colors [6], and research findings could be biased by participant assignments, casting doubt on objectivity and cross-cultural generalization; Second, traditional humanities methods of the study of color symbolism have shown consistent limitations. Prior research, which can be called 'close reading' studies, frequently analyzes color based on a small number of terms and their subjectively assigned colors. Critics point out that this approach, which seeks to comprehend the entire structure through specific incidents, falls short of understanding the larger literary context [47]. Over-reliance on a limited number of noteworthy color examples may impede a thorough comprehension of the role of color in literature. Current color studies are restricted to counting direct color depictions [56, 67], even when employing 'remote reading' techniques. Researchers still have to categorize the images and colors that phrases represent, the subjectivity issue with matching terms and pictures notwithstanding. However, one-to-one bonds between pictures and terms are often deemed the most practical, though tagging all vocabularies is a daunting task due to time requirements and the possibility of human errors or omissions [10, 9].

Through an intuitive use of massive imagery data, this study makes concrete the hitherto ambiguous relationships between words and colors that were previously limited to the conceptual realm. These data show the word-color associations that reflect real-world human cognition and color perception, and the associations are quantified via generalized distributions in the color gamut. Words and pictures are correlated through a search platform, and these images are translated into a color system aligned with human visual perception. The color concept is then used to partition the color space and represent the picture as a distribution of familiar colors. These distributions are cross-examined on a general level to express how strongly each color corresponds to the target term, providing a quantitative relationship

between words and colors that reflects human ideas and perceptions. By assigning colors to words in a data-based objective manner, we provide a framework that identifies various colors associated with words within a general frame of perception and expresses the relevance quantitatively. This methodology is then applied to the complete works of Anglo-American Imagist poet and Harlem Renaissance poet. Since the entire oeuvre published by a poet constitutes their world of poetry, and an individual poem consists of the intentional arrangement of terms, with each possibly expressing multiple colors, colors can be said to form a complex relationship with poetry. Analyzing the multi-layered structure to discover patterns or groups of colors used by the poet can quantitatively characterize the poet's literary universe from the perspective of color symbolism. Via this process, we can hope to conduct a distant reading study on the colors in a large literary corpus and demonstrate its potential for new symbolism studies.

2. Related Works

Recent years have witnessed the advancement of quantitative frameworks for modeling and analyzing large-scale heterogeneous data, extending their scope to the field of liberal arts [46, 5, 17, 8]. Researchers are not only using traditional cultural study methodologies but also attempting the application of statistical tools to cultural data [63]. These approaches enable a comprehensive examination of extensive text and image data, along with intricate metadata, offering additional parameters that complement personal readings of humanities resources with objective and quantitative features [44, 36]. Technological progress has made it possible to objectively depict universal human concepts, allowing for new methods to explore abstract topics such as color symbolism. As a result, numerous efforts have been made to comprehend the correlation between words and colors. These can be classified into three main types: First, creating a word-color dataset by giving predetermined terms to participants in surveys and collecting their replies on the corresponding colors. This endeavor has been motivated by the limited availability of databases that match colors and phrases [22], occasionally utilizing gamification elements to stimulate data gathering [35]; Second, using natural language processing techniques to assign colors to words that are not directly color-related by analyzing the surrounding terms tied to color expressions. Contextual reflection, aided by semantic embedding, allows for the quantification of the relationship between text and color at a level that does not involve images, using surrounding words to recognize color illustrations [50, 22]; Third, although there are studies pairing terms and pictures using search engine results, they simply use the RGB color space, forcing researchers to manually assign colors to visuals[50]. While all these strategies use statistical procedures to allocate color compositions to words in the form of ratios (weights) of colors, they do not specifically depict the relationship between the term and the visuals it represents: they instead either use text or surveys, or directly assign images to terms, with only a few people manually and subjectively partitioning and naming colors within the color space. Some even restrict the composition of a word to an arbitrary number of colors.

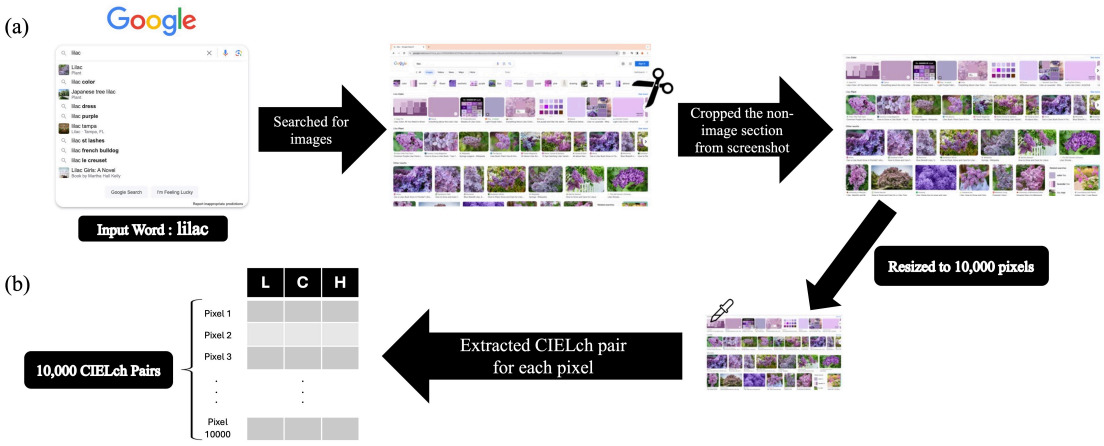


Figure 1: Word to Image to Color. The process of converting a word (example *lilac*) to color space coordinates using Google Image Search and resizing. (a) The returned screen is cropped, resized to 10,000 pixels, then (b) the CIELch color space coordinates are extracted.

3. Methodology

We aim to establish a quantitative and set workflow linking words and colors, and then apply it to textual data. A series of steps is necessary to assign color compositions to a word. We start by obtaining images representing the given word and extracting the color values from each pixel of the images. Next, the word can be said to be represented by the pixel values. Then we compare the composition with that of general words to determine which colors are statistically overrepresented that can then said to be characteristic of the given word. We later demonstrate this methodology on a corpus of Anglo-American poets as an example of 'objective' research in color and literature.

3.1. Word to CIELch List Conversion

Our goal is to represent a word using color values that reflect general human cognition more accurately. To achieve this, we use Google Image search to concretize a word into multiple images for enhanced statistical confidence and then convert the image pixels into the CIELch color space, known for reflecting true human color cognition better than the simple RGB color space. The procedure is illustrated in Figure 1.

3.1.1. Word to Image

We further combine the power of text-based image search with image analysis techniques. The main idea of this approach is that the colors inherent in a concept are statistically predominant in pictures indexed in relation to the word explaining that concept [50]. To utilize pictures reflecting human expectations for any given word, we use Google's image search. This modern lookup engine provides access to illustrations uploaded and tagged by numerous users. We designate one screen filled with images arranged in order of relevance to the word by the

algorithm as the representative image set for the word. The search, scrolling, counting images, and taking screenshots are performed on the Selenium web browser automation library in Python [62]. Subsequently, resizing is performed to represent a word as a set of 10,000 pixels. As a result, an arbitrary word (search term) is represented by a screenshot uniformly resized to 10,000 pixels.

Google Search Google's search algorithm, which connects terms and pictures, evolved from the PageRank system that evaluates webpage importance by analyzing link structure. If many users access an image when searching for a specific word, the image is considered highly relevant. Studies have used Google search to correlate language and universal illustrations, supporting the relevance of visuals returned for specific texts [34, 69, 71]. Google's search engine has improved by integrating image information, user preferences, feature extraction, and text annotation [68, 30], strengthening the relationship between input text and resulting images [49, 29]. Therefore, using Google's search engine is effective for linking terms and pictures. Empirically, we found that some words either have few images or mostly text images. If fewer than 30 pictures are found or more than three out of the first ten contain text, the word is assumed to lack associated images. Pytesseract, a Python library with OCR capabilities, is used to quickly recognize text within images [24].

Capturing Images When a word is probed, pictures typically related with that term are displayed in order of relevance on one screen. We use screenshots to capture the benefits of algorithms that apprehend human concepts while extending the scope of their universality. This allows us to construct representative illustrations for words while capturing multiple highly relevant images with minimal storage. All screenshots are taken after scrolling down 15% of the screen to exclude website logos, search bars, and toolbars.

Resizing Images We use Python's Pillow library to resize images [7]. Reducing the number of pixels to be processed allows for a quicker link between words and colors while maintaining color distribution. The resize method in the Pillow library uses High-LANCZOS and Antialiasing filters to effectively preserve the color distribution of the original image while minimizing jaggies [11]. The LANCZOS filter, a resampling technique, maintains color distribution and detail by handling high-frequency components, while the Antialiasing filter smooths the pixel boundaries in the reduced image to minimize visual distortion. These two filters effectively preserve the color distribution of the original image.

3.1.2. Images to CIELch Color List

The image's pixel colors are converted from the RGB into the CIELch color space that takes into account the features of human visual perception. In order to accurately measure the color composition of a term represented by the 10,000 pixels, it is necessary to know the precise location of each pixel inside the color space. Our image processing involves utilizing the scikit-image package to transform the RGB color values of every pixel into CIELch. The CIELch color space is a polar coordinate representation of the CIELAB color space, which was devised by the International Commission on Illumination (CIE) in 1976 with the intention of creating a color space

that is perceptually uniform [61, 57]. The color system defines colors in terms of chroma $c*$ and hue $h*$ while preserving the lightness $L*$ value of CIELAB. By transforming CIELAB color space into polar coordinates, it becomes possible to comprehend and control colors in a more instinctive manner. This approach highlights disparities in both hue and saturation, enabling precise color comparison and analysis that aligns with human perception. This color space represents the entire range of human's photopic (daylight) vision and provides a comprehensive framework for color representation that closely matches how colors are perceived in real life. Consequently, every pixel is located in the CIELch color space using a specific combination of (L, c, h) values. Thus a word, via an image, translates to 10,000 (L, c, h) color space coordinates.

3.2. Removal of Grids

A resized screenshot contains images associated with the target word separated by white spaces that need to be removed in principle, as they are not related to the visual representing the term. The white spaces are different in each screenshot, and the detection is computation intensive, making it impractical to perform it for every screenshot. For efficiency therefore we calculate the general total area of the white space in sample screenshots by choosing 100 random content words and counting the average number of pixels with $L = 100$ (the brightest white). We then discount this number of white pixels from the screenshots.

Table 1
Basic Colors and Their Synonyms

Basic Color	Synonyms
Red	*scarlet, vermilion, ruby, carmine*
Orange	*tangerine, marmalade, orangish, apricot*
Yellow	*yellowish, yellowy, gold, golden*
Green	*greenish, verdant, leafy, greenery*
Blue	*azure, cobalt, cerulean, ultramarine*
Purple	*violet, purply, purplish, amethyst*
Pink	*rosy, blushing, shellpink, rose*
Black	*pitchblack, pitchdark, jetblack, blackish*
Grey	*silver, slategray, smokegray, silvery*
White	*snowywhite, milkwhite, milkywhite, chalkwhite*

3.3. Establishing Color Standards

3.3.1. Selection of Basic Colors and Method of Color Composition

Basic Color Terms As the hue parameter of the colors of words, we use the 11 Fundamental Color Terminologies that are generally employed across various civilizations [2]. The color of a certain word is a composite blend of multiple hues, rather than a singular color. Thus, to faithfully depict colors, we take into account both the hue and luminance of the visual that represents the term, assigning chromatic and achromatic compositions to the word. Excluding Brown, which traverses both chromatic and achromatic [54, 37], we select Red, Yellow, Blue,

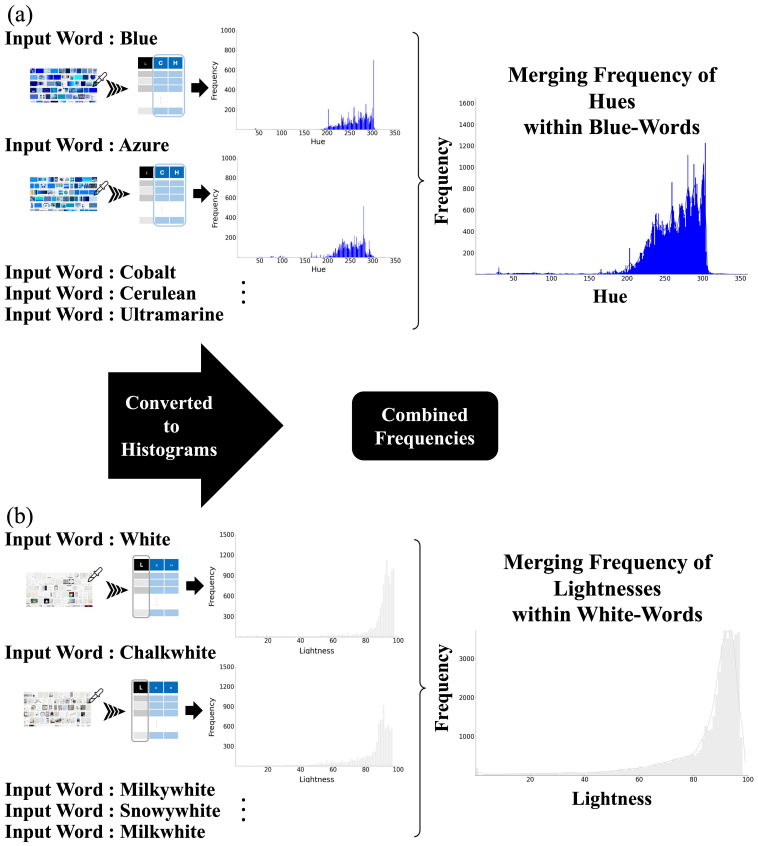


Figure 2: The diagrams illustrate the merging frequency process for color distributions using synonyms. (a) The combined frequencies show the hue distribution for blue-related words as an example of chromatic colors. (b) The combined frequencies show the lightness distribution for white-related words as an example of achromatic colors.

Orange, Green, Purple, Pink, White, Grey, and Black as representative colors. These representative colors include both Primary and Secondary Colors and have been typically recognized as major colors in art and design since the 17th century [58, 14, 15].

Use of Color Synonyms To determine a more accurate and reliable color profile of the named colors, we also consider their alternative names (synonyms) from the Oxford American Writer's Thesaurus (OAWT), creating a color standard that reflects broad cultural and linguistic contexts. This standard is established by using the color values of collected color terms, reflecting the universal expectations that representative colors hold as linguistic concepts. The OAWT is an extensive and globally recognized reference, regularly updated with the most recent evidence and research, making it a reliable benchmark [59, 55] with an easy accessibility as a free component of the Apple Operating System. Using different linguistic expressions for one color allows for a more realistic and precise setting of the universal range for that color. By utilizing linguistic variations of the basic colors, we robustly set the categories for the ba-

sic colors while clearly reflecting the general concepts surrounding them. Including various linguistic expressions for each color in the standard is a practical method for linking language and color. Thus, we search for four synonyms for each core color term through the dictionary and use them additionally to create specific color distinctions that correspond to people's perceptions. The basic color terms and their synonyms are listed in Table 1. The list of five words representing one color is each converted to a CIELch list and merged into one, allowing for duplicates. The combined basic color list is used to distinguish it from other colors as the color value representing that color. For example, the basic color list for red is created by merging the CIELch lists of "red," "scarlet," "vermilion," "ruby," and "carmine," and it is distinguished from the basic color lists of the other nine colors created in the same way. This process is illustrated in Figure 2.

Chromatic and Achromatic Colors Three achromatic and seven chromatic hues make up the ten fundamental color names. Our objective is to concurrently capture multiple different hues that a word evokes, not merely a single hue. It makes it possible to express a phrase from both chromatic and achromatic perspectives using the CIELch color space, which includes both hue and lightness. Hue information links words to the seven chromatic colors, while lightness information links to the three achromatic colors. We use seven basic color lists to divide the hue into seven parts and three arrays to divide the lightness into three parts.

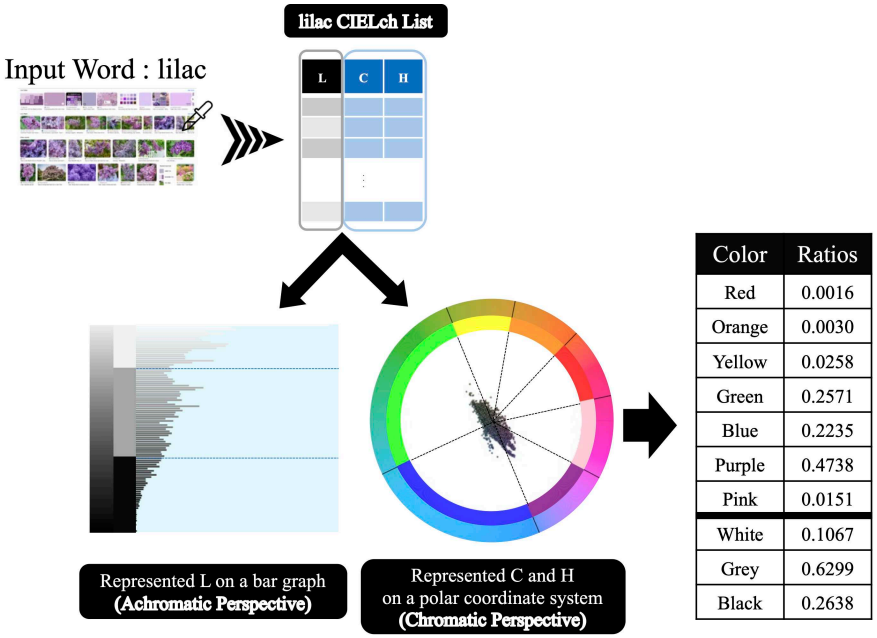
Chromatic Colors Hue identifies a color's exact location on the color wheel, represented numerically from 0.0° to 360.0°. We partition this plane into seven sections, assigning each core color term to a section. The basic color list, merged from five arrays, is expressed as a histogram for hue values. Each chromatic color is depicted as a distribution, showing its range within the hue at a general perception level.

Achromatic Colors Lightness indicates how bright or dark a color is, expressed as a value between 0.0 and 100.0 in the CIELch space. We divide this range into three sections for the achromatic colors. Using histograms, we determine the range for each achromatic color within lightness at a general perception level.

We apply a smoothing algorithm to the histograms to represent them as curved distributions. The rolling mean helps mitigate noise and emphasize trends and continuity to ensure the distribution represents the population. We use the rolling method from the Python library pandas, calculating the average of ten data points at a time [64]. This process transforms the histograms into smooth distributions. If multiple colors share a specific hue or lightness range, the predominant color takes precedence in practice.

3.4. Color Allocation

Once the intervals for the basic colors are determined, we can represent a word with a CIELch list and its color composition. As depicted in Figure 3, we use the h values from the CIELch list to express the word in terms of the proportions of the seven chromatic colors and the L values to express the proportions of the three achromatic colors. Each term can thus be described by a



Color	Ratios
Red	0.0016
Orange	0.0030
Yellow	0.0258
Green	0.2571
Blue	0.2235
Purple	0.4738
Pink	0.0151
White	0.1067
Grey	0.6299
Black	0.2638

Figure 3: Word to Color Ratio. The input word is converted into a CIELCh list and then transformed into color ratios based on the assigned intervals from both chromatic and achromatic perspectives. The resulting color ratios for *lilac* are listed in the table.

pair of chromatic and achromatic proportions, reflecting the visual diversity of the target term. Additionally, by comparing these color compositions to those of general words, we can use the Z-score to express which colors are emphasized in the target word, enhancing statistical confidence in the word-color association.

3.4.1. Wildtype

After describing a word with colors and their proportions, it is crucial to determine how much it differs from a typical, average word. This is similar to the biological notion of "wildtype," which describes the mean genotype or phenotype [3]. This functions as the baseline state in statistical analysis. Using the Z-score, we can analyze the tie between a word and its colors by comparing the word's color composition to the wildtype.

To create a reference wildtype dataset, we employ the WordNet interface from the NLTK toolkit [4]. WordNet is a comprehensive database classifying semantic relationships between words in English [45]. Using this approach, we randomly obtain a collection of 100 content phrases with their accompanying photos. Each word is converted into CIELch lists and allocated a proportionate composition for the 10 colors according to the agreed-upon color standards. And, we compute the Z-scores for the color composition ratios of each word, evaluating how much a word highlights colors compared to the wildtype. However, discrepancies in the distribution of colors can compromise the reliability of statistical analysis. To address this, By utilizing the Yeo-Johnson transformation to standardize color composition ratios [70], we enhance the precision of the Z-score calculation, resulting in a more reliable study of the re-

lationships between words and colors. This transformation can be performed using the 'PowerTransformer' module from the 'sklearn' package [51]. This method decreases variation in color composition and improves the accuracy of statistical analysis. By establishing a connection between a random word and its visual representation, expressed as color values, and building overarching color standards using color terminology and visual representations, we examine the distribution of colors in a word and compare it to a reference group. This helps ascertain whether it emphasizes specific colors more than average terms, providing a quantitative demonstration of the associations between specific words and colors in universal human perception.

3.5. Application

As an application of our method of word-color association, we have chosen to study Imagism and the Harlem Renaissance, two prominent movements in Anglo-American literature. Imagism, an early 20th-century literary movement, is known to emphasize the utilization of vibrant and precise imagery and concise language, giving significance to visual and intuitive modes of communication. In contrast, Harlem Renaissance utilizes written literature as a method to emphasize African American identity and cultural confidence, effectively conveying social and cultural messsages. This study seeks to evaluate the significance of color usages in the literary realms of these movements and poets, showcasing a quantitative analysis of literary works from a color-centric viewpoint. By employing this methodology, we hope to explore the possibility of new directions in the study of color use.

3.5.1. Data Collection and Preprocessing

Table 2
Data Sources for Selected Authors

Movement	Author	Data Source	Title
Imagism	Amy Lowell	Google Digital Copy (2024)	*Legends (1921)*
		Internet Archive (2024)	*Can Grande's Castle (1918)*
			Pictures of the Floating World (1919)
		Project Gutenberg (2024)	*A Dome of Many-Coloured Glass (1912)*
			Sword Blades and Poppy Seed (1921)
			Men, Women and Ghosts (1916)
Harlem Renaissance	Georgia Douglas Johnson	Internet Archive (2024)	*The Selected Works of Georgia Douglas Johnson (1997)*

Our choice of Imagist poet is Amy Lowell, while our selected Harlem Renaissance poets is Georgia Douglas Johnson. The works were compiled from their whole poetic world, or in instances where the total collected works were not accessible, specific poetry collections were combined. The collections were compiled from Project Gutenberg and freely distributed PDFs, as summarized in the Table 2. We employed OCR (Optical Character Recognition) technology [60], namely the image-to-text function of the Tesseract package [24], to extract tex from the PDFs. Subsequently, we conducted data preparation on the text utilizing the pandas and NLTK tools in Python [4]. The initial steps were eliminating stopwords, special characters, and nu-

merical values, and converting all uppercase letters to lowercase in the gathered text corpus for each poet, followed by lemmatization to convert words to their base or root forms.

3.6. Network Construction and Analysis

The oeuvre of a poet is their creative universe, with each poem comprising words that evoke various chromatic images. To understand the significance of colors in poetry, it can be helpful to elucidate the connections among artistic aspects, language, and colors as a network structure. By examining this structure and identifying patterns or clusters of colors used by the poet, we gain a quantitative comprehension of their work in terms of color symbolism. This examination reveals concealed significances and patterns in the poet's use of colors.

Building Process We establish a tripartite network connecting the works, words, and colors of each poet as separate node groups. A work is linked to every word it contains, and each word is linked to corresponding colors based on the Z-scores. The relationships between works and words are determined by frequencies, while ties between vocabularies and visuals follow the aforementioned guidelines for assigning color compositions. This forms the basis of our method for analyzing a poet's works via color symbolism.

Analysis The principles for ascertaining color compositions were applied to every phrase in the complete works of each poet. To identify the primary colors in the poet's literary universe, we made the projection to reduce the tripartite network into a bipartite network connecting works and colors. Our objective was to assess the importance of colors and identify the most significant ones in each poet's oeuvre. By transforming the network, we characterize the poet's works from a color-oriented viewpoint and compute the centrality – network-based measure of importance – of each hue.

In a tripartite network, we define artworks (titles of artworks) as P, words as W, and colors as C. The edges between P and W are determined by frequency, and the edges between W and C are weighted by Z-score. The process of projecting the artwork-word-color tripartite network into an artwork-color bipartite network is as follows:

$$w_{ij} = \sum_{k \in W} f_{ik} \cdot z_{kj}$$

where:

- w_{ij} is the weight between artwork i and color j.
- f_{ik} is the frequency between artwork i and word k.
- z_{kj} is the Z-score between word k and color j.
- $k \in W$ denotes the sum over all words k included in artwork i.

We apply the Birank random-walk-based algorithm to the resulting four works-color bipartite networks to calculate centrality [23]. This algorithm ranks nodes by considering the network topology of both classes, minimizing information loss during successive projection processes. Differences in color centrality between poets can be used to compare the influence of colors chosen by poets in constructing their poetic worlds.

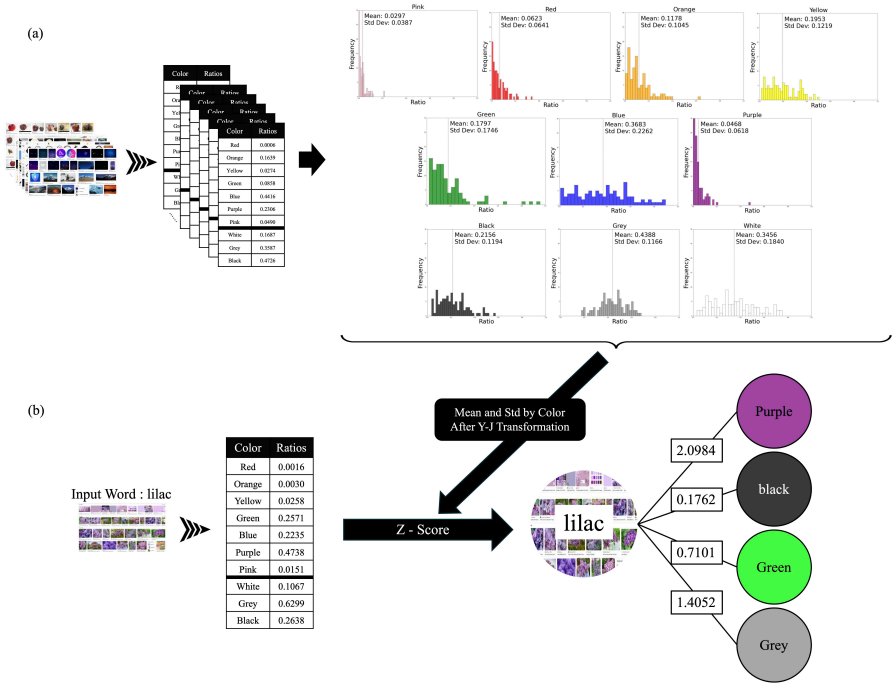


Figure 4: (a) **Reference Group Color Distribution.** The reference group (wild type) shows the mean and standard deviation for each color ratio. Histograms display the frequency distribution of color ratios for basic colors before applying the Yeo-Johnson transformation. (b) **Z-Score Calculation for the word *lilac*.** Using the mean and standard deviation from the reference group (after Yeo-Johnson transformation), the color ratios for the word *lilac* are transformed into Z-scores. Only colors with Z-scores of 0.0 or higher are shown.

4. Results and Discussion

4.1. Color Standards and Intervals

As shown in Figure 5, the distribution of each of the seven chromatic colors is represented in terms of hue, and the distribution of each of the three achromatic colors is represented in terms of lightness. The area of each color distribution is smoothed based on the histograms of the hue values of 50 000 pixels representing five words for each color, so the areas are equal. If different distributions meet at a point, that point becomes the boundary separating the colors. The finalized intervals are detailed in Table 3.

4.2. Wildtype

We selected 100 random content words and applied two procedures. The first is gap removal. All images of the words are screenshots of small pictures representing the word, with small gaps between them. We remove these gaps to emphasize the association between words and colors. The wildtype is then used to determine which colors are overrepresented by a word based on its color composition. We calculate the mean and standard deviation of the propor-

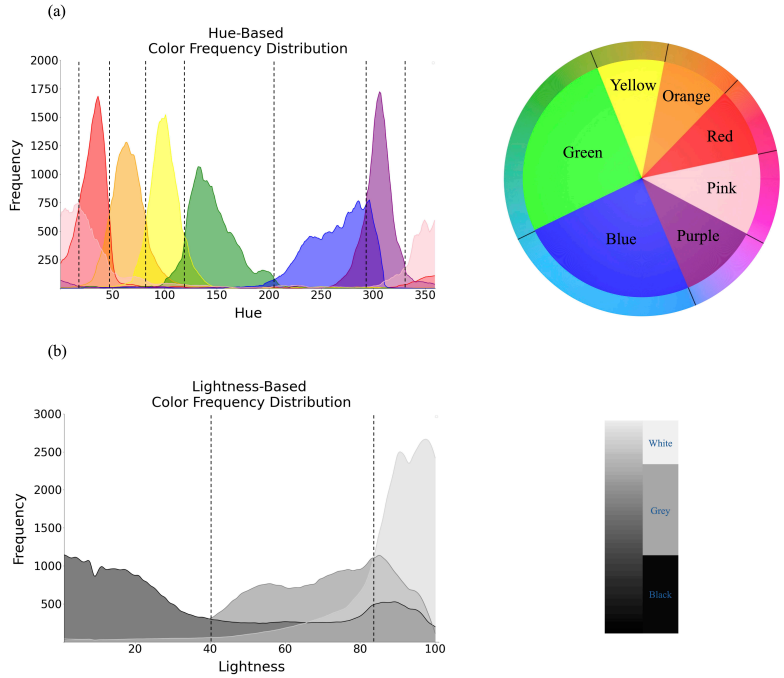


Figure 5: (a) **Hue-Based Color Frequency Distribution.** Frequency distribution of hue values for seven chromatic colors, smoothed to show range and frequency. The color wheel summarizes the hue ranges. (b) **Lightness-Based Color Frequency Distribution.** Frequency distribution of lightness values for three achromatic colors, smoothed to reflect general perception. The bar chart represents the ranges for white, grey, and black.

Table 3
Detailed Color Range Specifications

Basic Color	Ranges
Red	*Hue (degree) : (19.3270, 45.6767)*
Orange	*Hue (degree) : (45.6767, 80.9054)*
Yellow	*Hue (degree) : (80.9054, 120.1811)*
Green	*Hue (degree) : (120.1811, 203.2800)*
Blue	*Hue (degree) : (203.2800, 293.8690)*
Purple	*Hue (degree) : (293.8690, 331.2452)*
Pink	*Hue (degree) : (331.2452, 19.3270)*
Black	*Lightness (value) : (0.0000, 41.6539)*
Grey	*Lightness (value) : (41.6539, 82.8975)*
White	*Lightness (value) : (82.8975, 100.0000)*

tions of each color in 100 screenshots. Z-scores can be computed based on these values. Subsequently, the Yeo-Johnson transformation is applied, and Z-scores can be computed based on the transformed data. The detailed results of these transformation parameters are presented in the Table 4. These values serve as a baseline to help determine which colors are more prevalent

in a given word compared to others.

Table 4
Wildtype Statistics with Yeo-Johnson Transformation

Color	Y-J Mean	Y-J Std	Lambda
Red	0.0351	0.0237	−11.3730
Orange	0.0685	0.0382	−6.1276
Yellow	0.1421	0.0721	−1.8217
Green	0.0925	0.0447	−5.1438
Blue	0.2963	0.1620	−0.0635
Purple	0.0243	0.0163	−16.7988
Pink	0.0154	0.0118	−24.7169
Black	0.1485	0.0609	−2.2434
Grey	0.5328	0.1621	+1.9217
White	0.2554	0.1101	−0.6978

White Grids The proportion of pixels with $L = 100.0$ is determined for each random content word, and the mean is calculated. The proportion of pure white pixels is 0.1272. If the proportion of pixels with $L = 100.0$ in a specific word is less than this mean, all such pixels are removed. Otherwise, only the equivalent proportion is removed.

Distribution and Statistics for Each Color Each random content word is converted into a list of CIELch pairs. These lists are then expressed as proportions of colors according to the established color standards. Thus, each color is represented by 100 such values of proportion. As shown in Figure 4, the distributions and statistics differ by color, so we normalize the distribution differences between color proportions and then calculate the mean and standard deviation for each color. The transformed statistics are used to express the color composition of the target word as a Z-score.

Application to Literary Works We examine the intricate connections of artworks, language, and hues, and delineate associations between artworks and colors. This enables us to identify the prominent hues associated with each poet, facilitating our understanding. To evaluate the importance of colors for each poet, we utilize a threshold on the Z-scores, which indicate the strength of the correlations between words and colors. Raising the threshold entails implementing a more stringent criterion for the correlation between words and colors. As the threshold rises, only colors with strong associations to words are retained in the analysis, while those with weaker links are omitted. This method emphasizes the predominant hues in a poet's work, providing clearer insights into their creative decisions.

In Table 5, the prevalence of specific hues, especially black, intensifies as the threshold escalates for Georgia Douglas Johnson. At the initial threshold (0.000), where many colors are evident, black has a prominent place. As the threshold escalates, black becomes progressively central, and upon reaching a threshold of 2.576, black has a predominant birank value. This

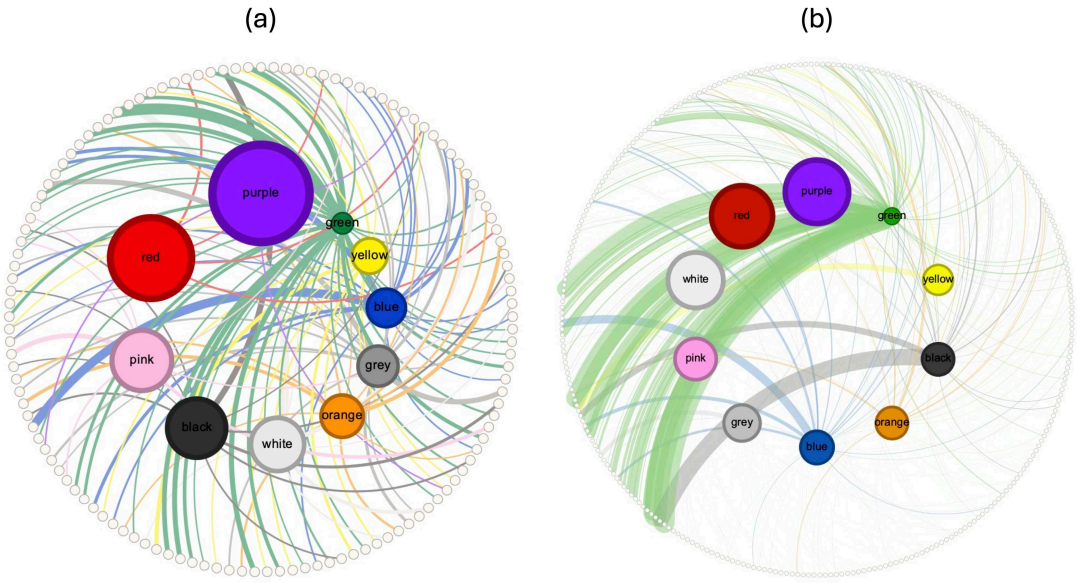


Figure 6: The bipartite networks illustrate the connections between the works (outer circle) and colors (inner circle) in the poetic works of Georgia Douglas Johnson and Amy Lowell. Nodes represent colors and are sized by their birank-centrality, with a Z-score threshold of 1.282 indicating significant associations. And, The edges are represented as an MST (Minimum Spanning Tree). Both poets emphasized red, purple and pink. (a) **Georgia Douglas Johnson's Poetic Networks.** black is relatively prominent. (b) **Amy Loewll's Poetic Networks.** white is comparably emphasized, and the variation between the colors is less pronounced.

use of color corresponds with the literary framework of the Black Arts Movement, noted for its focus on Black identity and culture [27, 53, 16]. Amy Lowell exhibits a notably broad and uniform application of color at the initial threshold (0.000). This equitable allocation of color application corresponds with the tenets of the Imagism movement, which prioritizes tangible and varied imagery, encompassing both natural vistas and urban environments, to elicit sensory experiences [21, 12, 20]. As the threshold rises, Lowell increasingly employs white, underscoring a distinct pattern of color focus in contrast to Johnson.

5. Limitations and Future Works

The separation between chromatic and achromatic hues leads to the omission of brown, potentially limiting detailed color analysis. Maintaining the CIELch space in a three-dimensional form and assigning intervals could allow more differentiated colors.

Nevertheless, our color palette has the potential to accommodate the inclusion of new colors. Each color is associated with the word and its synonyms, meaning new colors can be expressed as distributions over the color range and occupy specific intervals. This flexibility allows researchers to choose a variety of colors of interest to incorporate to the text being analyzed.

Currently, the network construction recognizes the significance of colors based on their asso-

ciations with artworks. Incorporating additional metrics like community detection algorithms or topic modeling techniques can offer new insights into the influence of colors within text networks and their contribution to thematic elements. Identifying clusters of subjects, phrases, and colors can provide a comprehensive understanding of color symbolism in texts.

Acknowledgments

This work is supported by the KAIST Post-Al Research Grant, BK 21 FOUR Program, the National Research Foundation of Korea (NRF-RS-2023-00245361, NRF-2020S1A5C2A03093177), and the Culture, Sports, and Tourism R&D Program through the Korea Creative Content Agency, funded by the Ministry of Culture, Sports, and Tourism in 2024 (KOCCA:RS-2023-00270043, Contribution Rate: 50%)

References

[1] K. Andreeva. "The Impact of Colour Foregrounding on the Text of the Novel". In: *International Journal of Linguistics* 11.4 (2019), pp. 82–94.

[2] B. Berlin and P. Kay. *Basic color terms: Their universality and evolution*. Univ of California Press, 1991.

[3] Biology Online. *Wild Type - Definition from Biology Online Dictionary*. https://www.biologyonline.com/dictionary/wild-type. 2024.

[4] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., 2009.

[5] J. Byszuk. "The Voices of Doctor Who–How Stylometry Can be Useful in Revealing New Information About TV Series." In: *DHQ: Digital Humanities Quarterly* 14.4 (2020).

[6] Y. Chen, J. Yang, Q. Pan, M. Vazirian, and S. Westland. "A method for exploring word-colour associations". In: *Color Research & Application* 45.1 (2020), pp. 85–94.

[7] A. Clark and Contributors. *Pillow: The friendly PIL fork*. Pillow. 2024.

[8] H. Craig and A. F. Kinney. *Shakespeare, computers, and the mystery of authorship*. Cambridge University Press, 2009.

[9] S. K. Dandapat et al. "Tagging multi-label categories to points of interest from check-in data". In: *IEEE Transactions on Emerging Topics in Computational Intelligence* (2023).

[10] M. Diligenti, M. Gori, and M. Maggini. "Learning to tag text from rules and examples". In: *AI*IA 2011: Artificial Intelligence Around Man and Beyond: XIIth International Conference of the Italian Association for Artificial Intelligence, Palermo, Italy, September 15-17, 2011. Proceedings 12*. Springer. 2011, pp. 45–56.

[11] C. E. Duchon. "Lanczos filtering in one and two dimensions". In: *Journal of Applied Meteorology and Climatology* 18.8 (1979), pp. 1016–1022.

[12] I. Fariha. "Poetic Image as Central Element in Early Imagist Poetry". In: *Journal of English and Education (JEE)* (2009), pp. 37–52.

[13] E. Feisner and R. Reed. “Color symbolism”. In: 2014, pp. 184–205. DOI: 10.5040/9781501303364.ch-014.

[14] J. Gage. *Color and meaning: Art, science, and symbolism*. Univ of California Press, 1999.

[15] J. Gage. *Color and culture: Practice and meaning from antiquity to abstraction*. Univ of California Press, 1999.

[16] A. Gayle. “The Harlem renaissance: Towards a black aesthetic”. In: *Midcontinent American Studies Journal* 11.2 (1970), pp. 78–87.

[17] T. Gessey-Jones, C. Connaughton, R. Dunbar, R. Kenna, P. MacCarron, C. O’Conchobhair, and J. Yose. “Narrative structure of A Song of Ice and Fire creates a fictional world with realistic measures of social complexity”. In: *Proceedings of the National Academy of Sciences* 117.46 (2020), pp. 28582–28588.

[18] *Google Digital Copy*. https://books.google.com. Google, 2024.

[19] *Project Gutenberg*. https://www.gutenberg.org. Project Gutenberg, 2024.

[20] B. S. Hama. “Imagism and Imagery in the Selected Poems of Major Imagist Poets”. In: *Koya University Journal of Humanities and Social Sciences* 3.1 (2020), pp. 88–93.

[21] C. A. Hamilton. “Toward a cognitive rhetoric of imagism”. In: *Style* 38.4 (2004), pp. 468–490.

[22] J. Harashima. “Japanese Word__Color Associations with and without Contexts”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*. 2016, pp. 2119–2123.

[23] X. He, M. Gao, M.-Y. Kan, and D. Wang. “Birank: Towards ranking on bipartite graphs”. In: *IEEE Transactions on Knowledge and Data Engineering* 29.1 (2016), pp. 57–71.

[24] S. Hoffstaetter and Contributors. *Pytesseract: A Python wrapper for Google Tesseract*. Pytesseract. 2024.

[25] D. Hunt. “Colour Symbolism in the Folk Literature of the Caucasus: TOPICS, NOTES AND COMMENTS”. In: *Folklore* 117.3 (2006), pp. 329–338.

[26] J. Hutchings. “Folklore and symbolism of green”. In: *Folklore* 108.1-2 (1997), pp. 55–63.

[27] G. Hutchinson. *The Harlem Renaissance in black and white*. Harvard University Press, 1995.

[28] *Internet Archive*. https://archive.org. Internet Archive, 2024.

[29] T. Jiang and A.-H. Tan. “Discovering image-text associations for cross-media web information fusion”. In: *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer. 2006, pp. 561–568.

[30] Y. Jing and S. Baluja. “Pagerank for product image search”. In: *Proceedings of the 17th international conference on World Wide Web*. 2008, pp. 307–316.

[31] G. D. Johnson. *The selected works of Georgia Douglas Johnson*. G.K. Hall, 1997.

[32] F. I. Kartashkova and L. E. Belyaeva. “Colour Meaning in English Literary Pieces”. In: *RUDN Journal of Language Studies, Semiotics and Semantics* 13.1 (2022), pp. 201–212.

[33] M.-O. Kim. “A study on the color images of Jeong-Ju Seo Midang’s poetry - focusing on the variation of desires”. Master’s thesis. Seoul: Seoul National University of Science, Technology, Graduate School of Industry, and Engineering, 2012, pp. iv, 83.

[34] J. Kiros, W. Chan, and G. Hinton. “Illustrative language understanding: Large-scale visual grounding with image search”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018, pp. 922–933.

[35] M. Lafourcade, N. Le Brun, and V. Zampa. “Crowdsourcing word-color associations”. In: *Natural Language Processing and Information Systems: 19th International Conference on Applications of Natural Language to Information Systems, NLDB 2014, Montpellier, France, June 18-20, 2014. Proceedings 19*. Springer. 2014, pp. 39–44.

[36] A. Lewis. “Modeling the humanities: Data lessons from the world of education”. In: *International Journal of Humanities and Arts Computing* 10.1 (2016), pp. 51–62.

[37] D. T. Lindsey and A. M. Brown. “Sunlight and ”blue”: The prevalence of poor lexical color discrimination within the ”grue” range”. In: *Psychological Science* 15.4 (2004), pp. 291–294.

[38] A. Lowell. *A Dome of Many-Coloured Glass*. ReadHowYouWant. com, 1912.

[39] A. Lowell. *Can Grande’s castle*. Macmillan, 1918.

[40] A. Lowell. *Legends*. Houghton Mifflin, 1921.

[41] A. Lowell. *Men, women and ghosts*. IndyPublish. com, 1916.

[42] A. Lowell. *Pictures of the floating world*. Macmillan, 1919.

[43] A. Lowell. *Sword Blades and Poppy Seed*. ReadHowYouWant. com, 1921.

[44] P. Martín-Rodilla, J. I. Panach, C. González-Pérez, and O. Pastor. “An experiment on accuracy, efficiency, productivity and researchers’ satisfaction in digital humanities data analysis: dataset appendix”. In: (2016).

[45] G. A. Miller. “WordNet: a lexical database for English”. In: *Communications of the ACM* 38.11 (1995), pp. 39–41. DOI: 10.1145/219717.219748.

[46] S. Min and J. Park. “Narrative as a Complex Network: A Study of Victor Hugo’s Les Misérables”. In: *Proceedings of HCI Korea*. 2016, pp. 100–107.

[47] F. Moretti. *Distant reading*. Verso Books, 2013.

[48] K. I. Mukhitdinovna. “The Problem Of Symbolic Of Colours In Different Language Cultures”. In: *American Journal of Interdisciplinary Research and Development* 4 (2022), pp. 123–129.

[49] P. Nieuwenhuysen. “Information Discovery and Images A Case Study of Google Photos”. In: *2018 5th International Symposium on Emerging Trends and Technologies in Libraries and Information Services (ETTLIS)*. Ieee. 2018, pp. 16–21.

[50] G. Özbal, C. Strapparava, R. Mihalcea, and D. Pighin. “A comparison of unsupervised methods to associate colors with words”. In: *International Conference on Affective Computing and Intelligent Interaction*. Springer. 2011, pp. 42–51.

[51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[52] O. Polshchykova and A. Polshchykova. "Colour signs-symbols in the artistic discourse as components of the cultural semiosphere (on the material of the ethnic literature of the 20th-century USA)". In: *Filologičeskie nauki. Voprosy teorii i praktiki* 16 (2023), pp. 1194–1200. DOI: 10.30853/phil20230202.

[53] R. J. Powell and D. A. Bailey. *Rhapsodies in black: Art of the Harlem renaissance*. Univ of California Press, 1997.

[54] P. C. Quinn, J. L. Rosano, and B. R. Wooten. "Evidence that brown is not an elemental color". In: *Perception & Psychophysics* 43.2 (1988), pp. 156–164.

[55] I. G. Roberts, ed. *The Oxford handbook of universal grammar*. Oxford University Press, 2017.

[56] N. Romanyshyn. "Corpus Based Analysis of Color Names Function in Poetic Discourse". In: *2022 IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT)*. Ieee. 2022, pp. 111–114.

[57] J. Schanda. *Colorimetry: understanding the CIE system*. John Wiley & Sons, 2007.

[58] A. E. Shapiro. "Artists' colors and Newton's colors". In: *Isis* 85.4 (1994), pp. 600–630.

[59] J. A. Simpson and E. S. C. Weiner, eds. *Oxford English Dictionary*. Vol. 3. Oxford University Press, 1989.

[60] R. Smith. "An overview of the Tesseract OCR engine". In: *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*. Vol. 2. Ieee, 2007, pp. 629–633.

[61] C. Standard et al. "Colorimetry-part 4: CIE 1976 L*a*b*colour space". In: *International Standard* (2007), pp. 2019–06.

[62] S. Stewart and Contributors. *Selenium WebDriver*. Selenium. 2024.

[63] N. Tahmasebi, N. Hagen, D. Brodén, and M. Malm. "A Convergence of Methodologies: Notes on Data-Intensive Humanities Research." In: *Dhn*. 2019, pp. 437–449.

[64] T. pandas development team. *pandas: Powerful Python data analysis toolkit*. pandas. 2024.

[65] A. N. Toxirovna. "Colors and their artistic image creation features". In: *International Journal on Integrated Education* 3.10 (2020), pp. 251–254.

[66] T. Underwood. *Distant horizons: digital evidence and literary change*. University of Chicago Press, 2019.

[67] F. B. Wadsworth, J. Vasseur, and D. E. Damby. "Evolution of vocabulary in the poetry of Sylvia Plath". In: *Digital Scholarship in the Humanities* 32.3 (2017), pp. 660–671.

[68] Y. Wan, X. Liu, Y. Chen, et al. "Online image classifier learning for Google image search improvement". In: *2011 IEEE International Conference on Information and Automation*. Ieee. 2011, pp. 103–110.

[69] Y. Yao, J. Zhang, F. Shen, X. Hua, J. Xu, and Z. Tang. “Exploiting web images for dataset construction: A domain robust approach”. In: *IEEE Transactions on Multimedia* 19.8 (2017), pp. 1771–1784.

[70] I.-K. Yeo and R. A. Johnson. “A new family of power transformations to improve normality or symmetry”. In: *Biometrika* 87.4 (2000), pp. 954–959.

[71] Y. Zhou, K. Chen, and X. Yang. “Google image search refinement: Finding text in images using local features”. In: *2012 IEEE International Conference on Computer Science and Automation Engineering (CSAE)*. Vol. 1. Ieee. 2012, pp. 98–101.

Table 5
Color Centrality Ranking by Poet Against Different Thresholds

Imagism		Harlem Renaissance	
Amy Lowell		Georgia Douglas Johnson	
threshold : 0.000			
color	birank	color	birank
purple	0.136256	purple	0.156427
red	0.134740	red	0.134771
white	0.124037	pink	0.105000
blue	0.102301	white	0.104932
pink	0.099765	blue	0.102740
grey	0.098624	black	0.101768
orange	0.086580	grey	0.084472
yellow	0.080566	orange	0.084240
black	0.079782	yellow	0.071469
green	0.057347	green	0.054181
threshold : 1.282			
color	birank	color	birank
purple	0.159047	purple	0.192556
red	0.153338	red	0.159229
white	0.138381	pink	0.114978
pink	0.103135	black	0.112283
grey	0.090997	white	0.101010
blue	0.081868	orange	0.078249
orange	0.079721	grey	0.074453
black	0.079462	blue	0.069734
yellow	0.073915	yellow	0.063331
green	0.040136	green	0.034178
threshold : 1.645			
color	birank	color	birank
red	0.172395	red	0.186979
purple	0.161989	purple	0.183828
white	0.145447	black	0.124643
grey	0.094579	pink	0.116571
pink	0.092988	whit	0.104163
black	0.083757	orange	0.079978
orange	0.074665	grey	0.069749
yellow	0.070590	yellow	0.058651
blue	0.069440	blue	0.052452
green	0.034150	green	0.022987
threshold : 1.960			
color	birank	color	birank
red	0.202281	red	0.276942
white	0.160345	black	0.160740
purple	0.138880	pink	0.113437
grey	0.107884	purple	0.112160
black	0.090139	white	0.109497
yellow	0.087211	orange	0.078618
orange	0.074131	yellow	0.059716
pink	0.072172	grey	0.055331
blue	0.044089	blue	0.022949
green	0.022868	green	0.010608
threshold : 2.576			
color	birank	color	birank
white	0.316941	black	0.664419
yellow	0.238126	white	0.131132
black	0.226836	grey	0.116518
grey	0.217889	yellow	0.087931
orange	0.000207	-	-
-	-	-	-
-	-	-	-
-	-	-	-
-	-	-	-
-	-	-	-