The Abdus Salam
International Centre
for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

# Quantitative analysis of the frequency of letters on the alphabet as an explanatory factor of letter-color synesthetes

Dissertation Thesis of the Diploma Program, Quantitative Life Science Section

Author: Carlos Enmanuel Soto López *(carlos.soto362@gmail.com)*
Supervisor: Matteo Marsili *(marsili@ictp.it)*

August 19, 2022

# Contents

# 1 Introduction

Using a big sample of letter-color associations in English speakers synesthetes [20], we test the hypothesis that the frequency of letters in the literature used during the process of learning is an important explanatory factor for the associations.

Letter-color synesthesia is a phenomenon where the synesthetes perceived an elicited color when they see a letter, these associations are persistent between many repetitions over time, but inconsistent between people, nevertheless, these associations tend to be less consistent when the synesthetes are children than when they are adults [20]. This makes us think that the synesthetes are learning their associations, for which we speculate that the frequency of letters in the literature that they encounter is important in the moment of learning this associations.

To test the hypothesis, we used the frequencies of a sample analyzed in [4], and the colors that each synesthete associated with each letter in the English alphabet in the L$^*$a$^*$b$^*$ color space, in cartesian and cylindrical coordinates.

First, in 2 we described in more detail the phenomenon of synesthesia and letter-color synesthesia, as well as the concept of an explanatory factor. Then, in 3 we described the data that was used in the analysis, the reason to use the L$^*$a$^*$b$^*$ color space instead of the RGB color space or other coordinate system for colors 3.1, and the four different methods that we used to test the hypothesis 3.3 which are:

A linear regression and the computation of the $R^2$ and Kendall Tau statistics between the mean value of each coordinate in the cylindrical representation of the L$^*$a$^*$b$^*$ color space and the frequency of letters in picture books 3.3.1, the computation of the correlation between the principal component of the color associations, taking each letter as a component, and the frequency of letters in picture books, this based on principal component analysis 3.3.2, clustering of the data 3.3.3 and the comparison between the data, and the probabilistic representation learned by a Restricted Boltzmann Machine learning the distribution of a sample of the EMNIST [3] data 3.3.4.

In section 4, we described the results obtained, and finally, in 5, we discussed the results. All the codes used in this work can be found in [7].

## 2 Letter-color synesthesia

Synesthesia could be described as a pair "inducer-concurrent", where the inducer is a stimulus that elicits an experience called concurrent [18]. There is still debate on what can be identified as an inducer or a concurrent. Some authors refer to synesthesia as a mixture of senses, like mirror-touch synesthesia, where the synesthete feels the sensation of touch when seeing other people being touched, or sound-color synesthesia when a specific sound elicits a color in the visual field. But there are cases when the inducer or the concurrent don't come from the senses, for example, for some synesthetes, an idea, like language, could elicit the concurrent [19].

Synesthesia is not a pathological condition and can exist as a developmental condition, where the synesthesia is developed throughout the life of the synesthete without a precipitation event, and as an acquired condition, which is the result of a precipitation event, like the loss of a sense or the gesture of some hallucinogenic drugs. As for the causes, is still not clear if developmental and acquired synesthesia have the same neurological causes. All candidate neural mechanisms have in common that they reflect differences in connectivity relative to the neurotypical brain, but is not clear if the differences are structural or functional. Because synesthesia is a consequence of these differences, studying it means studying cognition in a more general way [18].

Some characteristics that distinguish synesthesia from other conditions are automaticity and consistency, where the synesthete can't control when to experience the elicit experience, and the same concurrent is always elicited from the same inducer [18].

Grapheme-color Synesthesia is one of the most studied types of synesthesia [19][13], in specific, letter-color synesthesia, which has been used not only to study the synesthesia on its own but also to study relationships between visual, acoustic and semantic aspects of language [13]. In letter-color synesthetes, a letter will elicit the perception of color, each color perceived by different graphemes is very consistent over periods of time if tested for the same person, but very inconsistent between people. Nevertheless, is far from been random [13][12].

The inducer-concurrent relationship in letter-color synesthetes is influenced by many factors, which Root. Nicholas, et al [13] called Explanatory Factors, which can influence the relationship in a direct or indirect manner. For example, semantic properties like the color names appear to influence the inducer-concurrent relations in no synesthetes [13]. The question is, what Explanatory Factors are special for synesthetes? And what explanatory Factors are more important in the inducer-concurrent relationships?

There are many studies analyzing one or many different explanatory factors for the letter-color synesthetes with several methods [2][12][13][15][20]. Usually, two groups of people are used, one group of synesthetes and the other of control participants. Is shown that both, the synesthetes and control participants, make associations between letters and colors in a non-random way. For example, [15] used 70 English speaker synesthetes, and 317 control participants, for which, both groups showed more associations between green and the letter 'f' than what is expected by chance, this means that, for $n$ participants, the probability that $m$ of them choose a given color $c$ for a given letter $l$, $p_m(c, l)$, if each possible combination is chosen equally likely, is less than a given value $\epsilon$. Usually an $\epsilon = 0.05$ is said to be highly significant, that is, $p_m(c, l) < 0.05$.

Simner et al [15] Showed for their two groups that, there were common explanatory factors between synesthetes and non-synesthetes, but also some unique ones for the synesthetes. For example, the grapheme frequency is an explanatory factor that only appears to be significant for synesthetes.

# 3 Method

Winawer et al [20] used the data of 6588 synesthetes that provided their data to a battery online (http://www.synesthete.org/), where a consistency test was performed to ensure that the participants were indeed synesthetes. The participants selected the color that they associated with a given letter from an RGB color map. They used the data to test the hypothesis that synesthetes learned their associations from a learning toy, that consisted of colored letters. They found that at least 6% of the participants had higher associations with this toy than what was expected by chance [20]. Using the same data set, we test the hypothesis that the frequency of letters in the literature during the process of learning is a relevant explanatory factor for synesthesia.

For developmental synesthetes, there is a strong consistency between the letter-colors associations over time, but this consistency seems to be less strong for kids than it is for adults [18]. We speculate that the frequency of the letters used while learning to read and write has a strong influence on the associations of letters and colors learned by the synesthetes. In order to test this hypothesis, many different methods were used, mostly analyzing correlation matrices formed by representing the colors in color space and using the letters as labels. In section 3.1, we discuss the color space used for this analysis, then in section 3.2 we discuss the letter frequencies that were used to test the hypothesis, and lastly, in 3.3 we discuss the different methods used to test it.

## 3.1 Color space

Because the hypothesis tested concerned the relation between the internal representation of the colors in the human brain, and the frequency of the letters, is important to use a representation of colors that is akin to what the human brain captures, in opposite to a representation that is good for reproducing the colors in a different medium.

Color is the perceptual result of electromagnetic waves between 400 and 700 *nm* of wavelength, captured by three types of receptors in the retina. Each of them responds to the incoming radiation with different spectral response curves. The Commission Internationale de L'Éclairage (CIE) has defined many systems to compute a color using three numbers, as coordinates in a color space, each of them fitting different needs [11].

All the systems are based on the "CIE XYZ" system. The "X" and "Z" components match the $\bar{\mathrm{x}}$ and $\bar{\mathrm{z}}$ components of the $\bar{\mathrm{x}}$, $\bar{\mathrm{y}}$ and $\bar{\mathrm{z}}$ color matching functions, spectral weight functions defined by the CIE based on an experiment in the 1920s to characterize the relationship between the spectral power distributions and color perception. The "Y" value has a spectral sensitivity that corresponds to the lightness sensitivity of human vision. Is equal to the integral of the spectral color distribution of the incoming light weighted by the $\bar{\mathrm{y}}$ color matching function [11].

In the 1940s an experiment was performed in order to test the "CIE XYZ" system, known as the MacAdam's experiment. In this experiment, MacAdam identifies the regions over the xy-diagram, where two colors were perceived as identical. The result was ellipsoids that differ in size and direction depending on the region of the xy-diagram (see figure 1). In order to get a space that return more uniform results, CIE defined the "CIE 1996 ($\mathrm{L^*a^*b^*}$)" [14], which is the color space used in this work, transforming the colors that originally where selected in a RGB color space, into the $\mathrm{L^*a^*b^*}$ space.

The $\mathrm{L^*a^*b^*}$ system would be the Cartesian coordinates of an approximately uniform color space, while the "CIE 1996 $\mathrm{L^*}$", Luminosity, "CIE 1996 a,b chroma" $\mathrm{C^*_{a,b}} = \left(\mathrm{a^{*2}} + \mathrm{b^{*2}}\right)^{1/2}$, and 'CIE 1996 a,b hue angle" $\mathrm{h_{a,b}} = \arctan\left(\mathrm{a^*/b^*}\right)$, would be the cylindrical coordinates [14]. The Chroma and Hue are normally defined in such a way that differences in Hue values meant different colors, while differences in Chroma meant differences in saturation of a given color.

## 3.2 Frequency of letters

As a point of comparison, we used the frequency of letters in English speaker children's picture books that Nicolas E. Fears et al found analyzing 100 different picture books [4].

The lower case letters had a bigger variation in frequency than the upper case letters, for which we choose to use the frequency of lower case letters for this work. Is important to note that, using Pearson's product-moment and

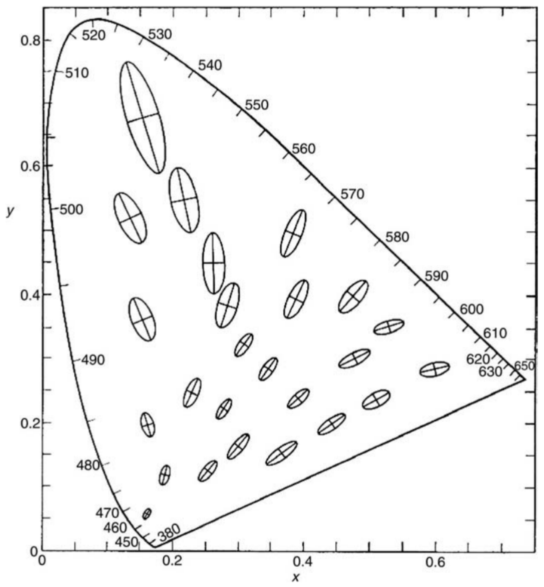

Figure 1: “MacAdam ellipses 10-times enlarged (from Judd DB, Wyszecki G (1963) Color in Business, Science, and Industry, Wiley, New York)” [11].

Spearman’s rank correlation, both, the lower case letters, and the upper case letters presented high correlation coefficients between the frequency of child books and the frequency of adult books [4].

## 3.3 Analysis of correlation between letter-color associations and letter frequency

### 3.3.1 Linear Regression

One way in which the correlation between colors and letters can be analyzed is to use one of the coordinates on the color space and see if the value of this coordinate is correlated with the frequency of a letter. In this way, if most of the synesthetes tend to choose high values of a coordinate for letters that are more frequent, there would be a positive correlation between that coordinate and the frequency of the letters. That is what Beeli et al [2] did use the Luminance and Saturation as the coordinates in the color space for 11 letter-color and digit-color synesthetes. What they found was a strong correlation between the frequency of digits and Luminance [2], but because the most frequent digits they report were bigger in value and the least frequent digits were the smallest in value, this correlation could be explained by a different explanatory factor, like the magnitude of the digit. In regards of the letter-color synesthetes, they found a small correlation between the frequency and the saturation [2].

In the present work, a similar procedure was performed, where the $L^*a^*b^*$

coordinates were transformed into his cylindrical counterparts, Luminosity, Hue, and Chroma, and then performed a linear regression taking the frequency of letters as an independent variable. To measure the correlation, the $R^2$ and the Kendall Tau statistic were computed.

### 3.3.2 Principal Component Analysis

Using the ideas of Principal Component Analysis (PCA), taking each letter as component, the idea was to find the linear combination of letters that best described the data, and then see if this principal component was correlated with the frequency of letters.

**PCA:**
In PCA, $N$ observations of $M$ dimensional data are compressed in a $N \times M$ matrix $\mathbf{X}$, the idea is to compute new linear combinations of the original data called principal components, such that the direction of the new components presents the maximum variance. The principal components are obtained by the singular value decomposition $\mathbf{X} = \mathbf{P}\Delta\mathbf{Q}$, where $\mathbf{P}$ and $\mathbf{Q}$ are the matrices of the left and right eigenvectors of the matrix $\mathbf{X}^T\mathbf{X}$, and $\Delta$ is the diagonal matrix with entries equal to the eigenvalues of $\mathbf{X}^T\mathbf{X}$ [1].

The matrix $\mathbf{F} = \mathbf{P}\Delta$ is the matrix of the factor scores, the $N \times L$ matrix, where $L$ is the number of non-zero eigenvalues, such that each row contains the value of each observation in the direction of each principal component [1].

The matrix $\mathbf{Q}$ gives the coefficients of the linear combinations of the new components. Is the projection matrix such that $\mathbf{F} = \mathbf{X}\mathbf{Q}$ [1]. If the eigenvalues of $\mathbf{X}^T\mathbf{X}$ are $\lambda_1 > \lambda_2 > \cdots > \lambda_L$, then his corresponding eigenvectors are the principal components, with the eigenvector with the largest eigenvalue being the one that describes the most amount of the variance of the data.

**Analysis using PCA:**
First, the matrix $\mathbf{X}$ was computed using only the synesthetes that associate a color to all the letters, a total of 4269 samples. Each observation is a 26-dimensional vector, where each entry is the value of the color that the synesthete associated to each letter. This was made for each of the three components in the cylindrical representation of the L$^*$a$^*$b$^*$ space, Luminosity, Hue and Chroma, and with the Cartesian representation, where each component of the observations were a 3-dimensional vector $(\mathrm{L}^*, \mathrm{a}^*, \mathrm{b}^*)$. All of them were re-scaled in such a way that the mean was zero and the standard deviation was one, meaning the vector zero and $X_i^2 = X_i^T X_i$ for the matrix in Cartesian coordinates.

After this, for each of these matrices, the correlation matrices were computed, $\mathbf{C} = \mathbf{X}^T\mathbf{X}$. These correlation matrices were compared with the resulting correlation matrices of the original data shuffling the colors that each synesthete associated to each letter, by applying a Kolmogorov Smirnov test [9] with the null hypothesis that the values of both correlation matrices belonged to the same distribution.

Finally, it was computed the eigenvalues and eigenvectors of each of the correlation matrices, comparing to the eigenvalues of the original data and the eigenvalues of the shuffle data. The biggest eigenvalues are the principal components, which were compared with the frequency of letters to see if they presented a correlation.

### 3.3.3 Clustering

One aspect that could affect the analysis, is that not all the synesthetes learn in the same way or with the same books. For this reason, clustering was implemented on the samples, and then the linear regression and the PCA analysis were performed with each of the resulting clusters. We used an implementation in Scipy, a library from python that uses a hierarchical agglomerative algorithm called linkage, where each data point starts as its own cluster, and then, they start to get united with the closest clusters, with the distance between points being the euclidean distance, and for the distance between clusters the Ward variance minimization algorithm is used [17].

### 3.3.4 Restricted Boltzmann Machine

A Restricted Boltzmann Machine (RBM) is a parameterized generative model that represents a probability distribution [5], given a set of data composed of different pictures with handwritten lowercase letters, with each letter having different frequencies, the aim is to test if the internal representation of the RBM after learning the parameters that best describe the set of pictures, has a joint probability distribution over letters $P_{RBM}(l, l')$ with entries correlated with the correlation matrices described in section 3.3.2.

**RBM:**

The RBM is a model with a form in analogy to the Boltzmann distribution

$$P_{RBM}(x, s) = \frac{1}{Z} e^{-E(x,s)}$$

$$E(x, s) = -\sum_{i=1}^{I} a_i x^i - \sum_{j=1}^{J} b_j s^j - \sum_{i,j=1}^{I,J} x_j W_i^j s^i = -a^T x - b^T s - x^T W s$$

where $x$ is the visible layer of dimension $I$, and $s$ is the hidden layer of dimension $J$. The vectors $a$ and $b$ of dimensions $I$ and $J$ respectively are the biases and the matrix $W$ of size $I \times J$ is the weights [6]. Given a set of data of $N$ observations $X^T = (x_1, x_2, \ldots, x_N)$, finding the biases and weights that best fit the probability distribution of the data is interpreted as learning, in the unsupervised learning regime. The aim is to maximize the likelihood of obtaining the set of data $X$ given the set of parameters $\theta = (a, b, W)$. Maximizing the Likelihood is equal to minimizing minus the log-likelihood, that is,

$$\hat{\theta} = \text{argmin}_\theta \left( -\log \Pi_i^N p(x_i|\theta) \right) = \text{argmin}_\theta \mathcal{L}$$

which can be approximated using vanilla gradient descent [5]. With concavity assumptions, and a small learning rate $\eta$, then the update rule

$$\theta^{(t+1)} = \theta^{(t)} - \eta \Delta_\theta \mathcal{L}$$

for sufficiently large $t$, approximates the value of $\hat{\theta}$. Using this simple update rule, the updates for the different parameters are,

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial W_{ij}} &= \left(< x_i s^j > - < x_i s^j | x >\right) \\
\frac{\partial \mathcal{L}}{\partial a_i} &= (< x_i > - < x_i | x >) \\
\frac{\partial \mathcal{L}}{\partial b_j} &= \left(< s^j > - < s^j | x >\right)
\end{aligned}$$

where $x_i$ is the $i$th component of one observation $x$, $s_j$ is the $j$th component of one state of the hidden layer $s$, $< \cdot >$ stands for expected value over the distribution $p(x, s)$, and $< \cdot | x >$ stands for expected value over the distribution $p(s|x)$ [6]. This expected value is approximated by sampling over subgroups of the total data that is been used to fit the model, the Training Data. The idea is to divide the Training Data into random groups, called mini-batches, and approximate the gradient given these mini-batches. Each time that the entire set of Training Data is used is called an Epoch.

There are many methods to approximate these expected values given the mini-batches. For this work, we used an implementation in Sklearn, a library from python [10] that uses Persistent Contrastive Divergence [16].

**Analysis using the RBM:**
The Data was chosen from the EMNIST data [3], where only lowercase letters were chosen. Each image is composed of $28 \times 28$ pixels, each pixel having a value between 0 to 255. Before training, the Training Data was chosen in such a way that the total amount of images used to train was around 50,000, with the restriction that the fraction of images for each letter matches the fraction of letters in the picture books described in section 3.2. Because of not having enough pictures for the letters “i”, “o” and “s”, these letters didn’t match the frequency of the picture books, and the total amount of images used to train was 47,984. After choosing the set of data, it was transformed into vectors of length 784, with binary entries.

Using this Training Data, first, we proceed to tune the learning rate and the number of hidden layers. After choosing suitable parameters, we trained the model using 100 epochs.

Once the RBM was trained, learning a set of parameters that approximates thous that maximizes the Likelihood, we estimated the joint probability distribution over letters. Using the property of the RBM that the probability distribution of two different configurations of the visible layers given an internal representation are conditionally independent, and using the Bayes rule,

$$\begin{aligned} P_{RBM}(l,l') &= \sum_s P_{RBM}(l|s)P_{RBM}(l'|s)P_{RBM}(s) \\ &\approx \sum_{TestData} \frac{\hat{P}(s|l)\hat{P}(s|l')\hat{P}(l)\hat{P}(l')}{\hat{P}(s)} \end{aligned}$$

where $P_{RBM}(\cdot)$ stands for probability of the RBM model, $\sum_s \cdot$ stands for the sum over all possible hidden layers $s$, and $\hat{P}(\cdot)$ stands for an estimate of the probability using a set of test data not used to train the model, the sample probabilities of the Test Data. Using the same amount of images for each letter in the Test Data, the sample probability of each image is

$$\hat{P}(l) = 1/26.$$

The sample probability of each hidden layer is

$$\hat{P}(s = s') = \frac{1}{N_{TestData}} \sum_{i=1}^{N_{TestData}} \delta(s_i = s').$$

where $\delta(\cdot)$ is 1 if the condition is true, and 0 otherwise. $s_i$ is the hidden layer associated with each image, is the value of $s_i$ that maximises the probability of the image given the internal representation.

$$\begin{aligned} s_i &= \mathrm{argmax}_s \left( P_{RBM}(x_i|s,\theta) \right) \\ &= \mathrm{argmax}_s \left( \sum_{j=1}^{J} b_j s^j + \sum_{k,j=1}^{I,J} x_j W_k^j s^k \right) \\ &= \{\delta(b_j + \sum_{k=1}^{I} x_j W_k^j s^k > 0)\}_{j=1}^{J} \end{aligned}$$

where the last line represents the hidden layer that has zeros if the condition inside the parenthesis is false, and ones if it is true. Finally, the sample probability of each hidden layer given a letter is

$$\hat{P}(s = s', l = l') = \frac{1}{N_{TestData}} \sum_{i=1}^{N_{TestData}} \delta(s_i = s', l_i = l').$$

For a sufficiently big sample of Test Data, this would approximate the real joint probability distribution. For time-consuming reasons, only 50 images per letter were used in the Test Data.

Once the joint probability distribution has been computed, we compared the values of the upper triangle of each of the correlation matrices described in section 3.3 with the upper triangle of the joint probability, making a linear regression of each of the values of the correlation matrices as independent variables and the values of the joint probability as the dependent ones. With this, the $R^2$ statistic was computed as well as the Kendall Tau statistic.

# 4 Results

## 4.1 Linear Regression

| letters | frequencies |
|---|---|
| a | 7.94 |
| b | 1.54 |
| c | 2.07 |
| d | 4.47 |
| e | 11.48 |
| f | 1.52 |
| g | 2.41 |
| h | 5.92 |
| i | 5.65 |
| j | 0.14 |
| k | 1.33 |
| l | 4.35 |
| m | 2.28 |
| n | 6.14 |
| o | 7.97 |
| p | 1.57 |
| q | 0.07 |
| r | 5.15 |
| s | 5.54 |
| t | 8.06 |
| u | 3.01 |
| v | 0.77 |
| w | 2.24 |
| x | 0.12 |
| y | 2.28 |
| z | 0.13 |

Table 1: Mean weighted frequencies of English picture books [4].

The frequency for letters that was mostly used was the mean weighted letter frequency of 100 picture books computed in [4] (see table 1).

A linear regression between this frequency as an independent variable and the mean value of luminosity, hue, and chroma over the 4269 synesthetes that associated a color to each letter were computed.

In figure 2 can be seen the linear regression of the three coordinates, luminosity, hue, and chroma, as well as his slope, $R^2$ statistic, Kendall Tau statistic, and the p-value of the Kendall Tau statistic.

For all of the regressions, the $R^2$ statistic showed a value close to 0, as well as the Kendall Tau statistic. All of them having a p-value bigger than 0.05 says that the null hypothesis that the two sets of data are independent was not rejected. The smallest value for the p-value was the one between the mean luminosity and the frequency of the letters, with a value of $p_{value} = 0.085$.

This result was done using only the mean value of each of these coordinates. Is important to use a method that uses more of the structure of the data than just the mean to test the original hypothesis.

## 4.2 Principal Component Analysis

Figure 3 shows the correlation matrix $\mathbf{C} = \mathbf{X}^T\mathbf{X}$ for the colors computed with the three components in the Cartesian representation, where the entry $\mathbf{C}_{ij}$ is the correlation between the $i$th letter and the $j$th letter of the alphabet, computed as described in section 3.3, and the correlation matrix where the colors that each synesthete associated to each letter was shuffled between the synesthetes.

Comparing the entries of the real data and douse of the shuffle data shows

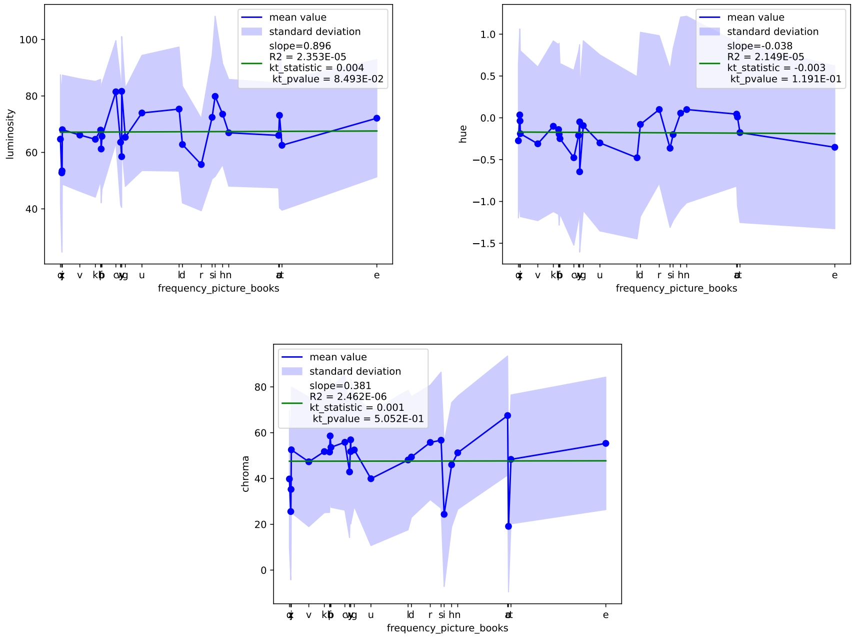

Figure 2: Linear Regression between the mean value of luminosity (top left), hue (top right) and chroma (bottom), and the frequency of picture books [4], as well as showing the $R^2$ statistic, the Kendall Tau statistic, and the p-value of the Kendall Tau statistic, for the three regressions.

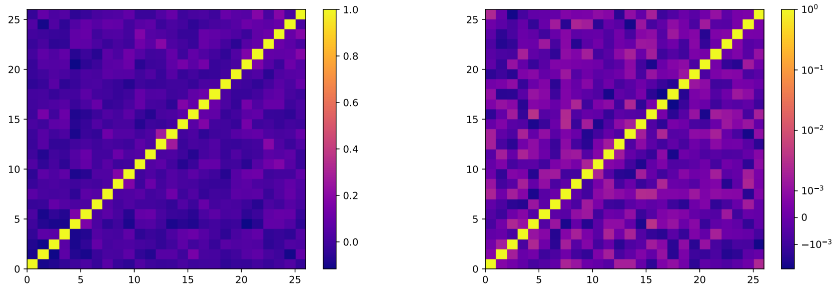

Figure 3: correlation matrix $\mathbf{C} = \mathbf{X}^T\mathbf{X}$ computed with the three components in the Cartesian representation, where the entry $\mathbf{C}_{ij}$ is the correlation between the $i$th letter and the $j$th letter of the alphabet, computed as described in section 3.3 (left), and the correlation matrix of the shuffle data, where the color that each synesthete associate to each letter where shuffle between all the synesthetes (right).

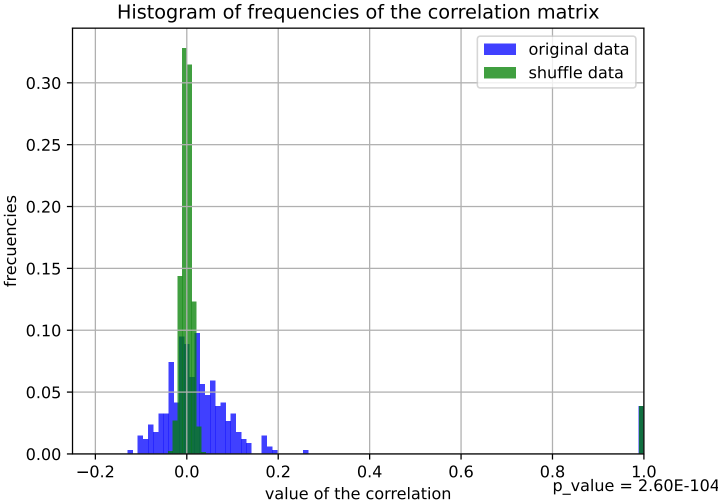

Figure 4: Histogram of values in the correlation matrix calculated with the three components in the Cartesian representation (blue) and the shuffle data (green). At the right bottom of each histogram, the p-value of the Kolmogorov-Smirnov Test.

that each entry in the real data tent to have bigger values than dose with the shuffle data. The letters for the real data are more correlated with each other. This can be seen in figure 4, where the histogram of the values of the real correlation matrix and douse of the shuffle data are plotted. A Kolmogorov-Smirnov Test [9] was computed, obtaining a high significant p-value, where the null hypothesis that the two samples came from the same distribution was rejected. This result is another confirmation that for the sample used in this work, the synesthetes have associations between letters and colors with a reach structure.

The same analysis was performed for the correlation matrices computed using only the Luminosity, the hue, or the chroma, obtaining the same result for each of these components (see figure 5).

After this, for each of these correlation matrices, we compute the eigenvalues and eigenvectors and compute the correlation between the eigenvector corresponding with the biggest eigenvalue, the one that points in the direction that has the biggest variance, and the frequency of letters in picture books.

The only correlation matrix that presented a significant correlation between the eigenvector corresponding with the biggest eigenvalue, and the frequency of letters in picture books, was the correlation matrix computed with the Chroma coordinate, having a value of $\text{corr} = -0.408$. The correlation was computed by taking the dot product between the unitary eigenvector and a normalized vector in the same direction as the frequency of letters in picture books shown in the

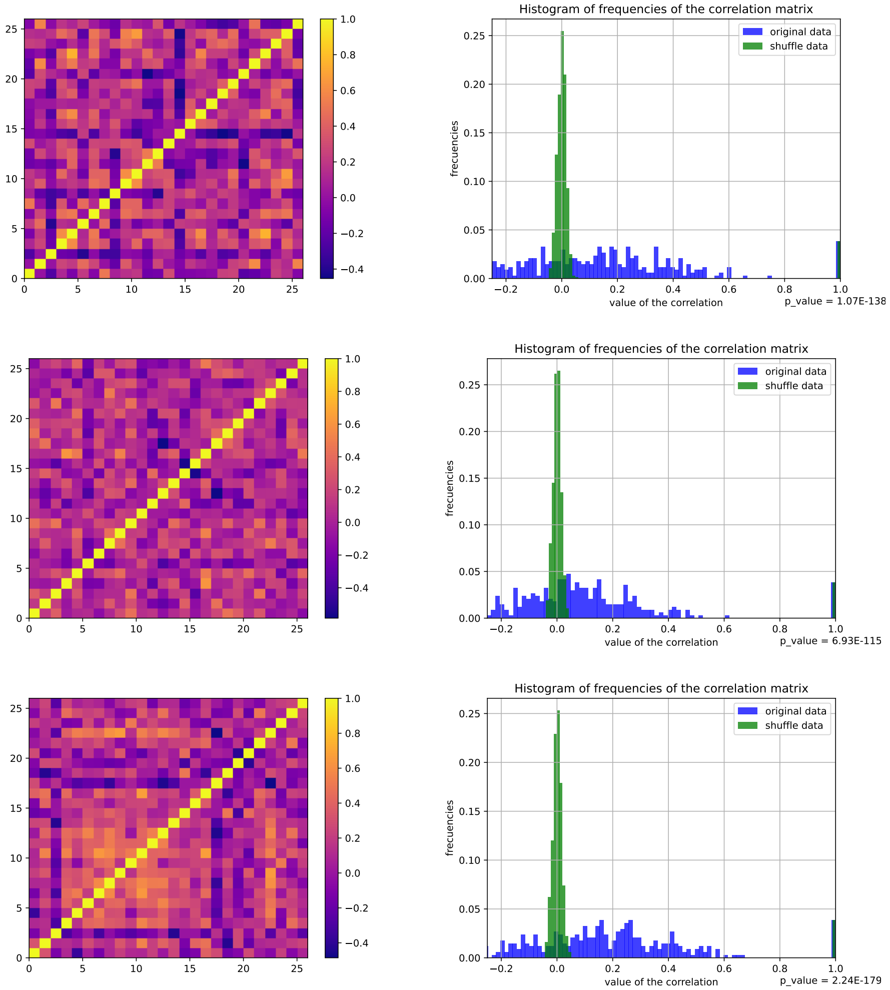

Figure 5: correlation matrix $\mathbf{C} = \mathbf{X}^T\mathbf{X}$ computed with the Luminosity (top), hue (middle), and chroma (bottom) representations, where the entry $\mathbf{C}_{ij}$ is the correlation between the $i$th letter and the $j$th letter of the alphabet, computed as described in section 3.3 (left), and the histogram of values with the real values and the shuffle ones, in blue and green respectively, and at the right bottom, the p-value of the Kolmogorov-Smirnov Test (right).

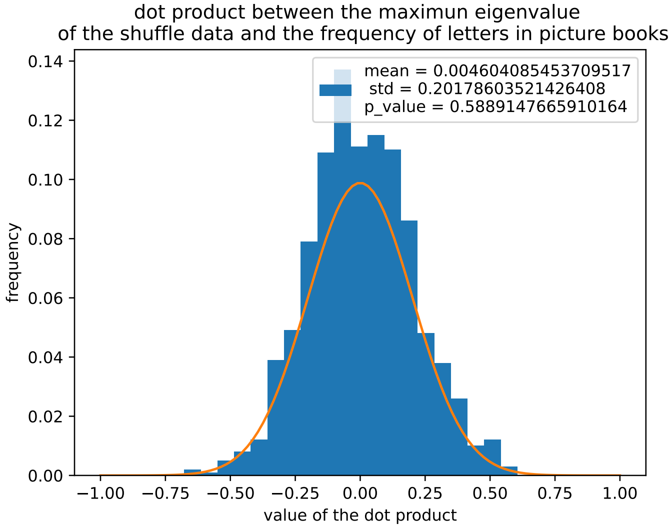

Figure 6: Histogram of the dot product between the frequency of letters in picture books and the eigenvector corresponding with the biggest eigenvalue of 1000 correlation matrices of shuffle sets of data.

table 1.

In order to see if a given value is significant, we compute this dot product between the frequency of letters in picture books and the eigenvector corresponding with the biggest eigenvalue of 1000 correlation matrices of shuffle sets of data. The histogram of the results is plotted in figure 6. Using a Kolmogorov-Smirnov test between the data and a Normal distribution with the same mean and variance result in a $p_{\text{value}} = 0.6$, which means that the null hypothesis that this data coming from a normal distribution is not rejected. Then, assuming that they do belong to a normal distribution, a value of corr $= -0.408$ is in the tails of this distribution, with a probability of occurring $p < 0.05$, which according to our criterion, is highly significant.

This means that the eigenvector corresponding to the biggest eigenvalue of the correlation matrix, computed with the Chroma coordinate, has a highly significant negative correlation with the frequency of letters in English picture books.

## 4.3 Clustering

Nine clusters were computed, using the implementation mentioned in section 3.3.3. As a mode of confirmation for what Winawer et al [20] did, we computed the correlation matrix $Y^TY$, with $Y$ the chroma, hue, and luminosity of the color that the colored toy had, and compute the eigenvector corresponding with the biggest eigenvalue. We found that one cluster with 17.6% of the sample

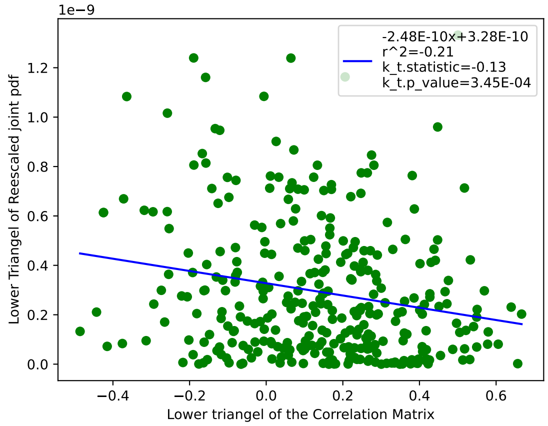

Figure 7: Linear regression between the values of the lower triangle of the correlation matrix computed using the chroma coordinate as an independent variable, and the values of the joint probability distribution of the RBM.

used had a highly significant correlation between the eigenvector of the chroma correlation matrix of the colored toy and the chroma correlation matrix of the synesthetes data, corr $= 0.496$.

Regarding the frequency of letters, no cluster presented a significant correlation.

## 4.4 Restricted Boltzmann Machine

To train the RBM, first, the learning rate was tuned, using 10 epochs and 256 hidden layers (see figure 8). The value with the best score was $\eta = 0.003$.

Then, the number of hidden layers was tuned, as can be seen in figure 9.

Because the upgrade between 30 to 256 hidden layers is not very strong, and using fewer hidden layers means needing a smaller amount of sample data to approximate the joint probability distribution, we used 30 hidden layers.

After tuning the hyperparameters, we train the model using 100 epochs. Once the biases and weights were learned, we estimated the joint probability distribution $\hat{P}_{RBM}(l, l')$ as described in section 3.3.4. The resulting probability distribution can be seen in figure 10.

Finally, we made a linear regression between the values of the upper triangle of the correlation matrices and the upper triangle of the joint probability distribution and computed the $R^2$ statistic as well as the Kendall Tau statistic.

The only correlation matrix that showed a highly significant correlation was the one computed using the chroma coordinate, with a Kendall Tau statistic $K_\tau = -0.13$ and a p-value $k_\tau \mathrm{p} - \mathrm{value} = 0.0003$, which means that the null hypothesis that the two sets of data are not correlated is rejected. According

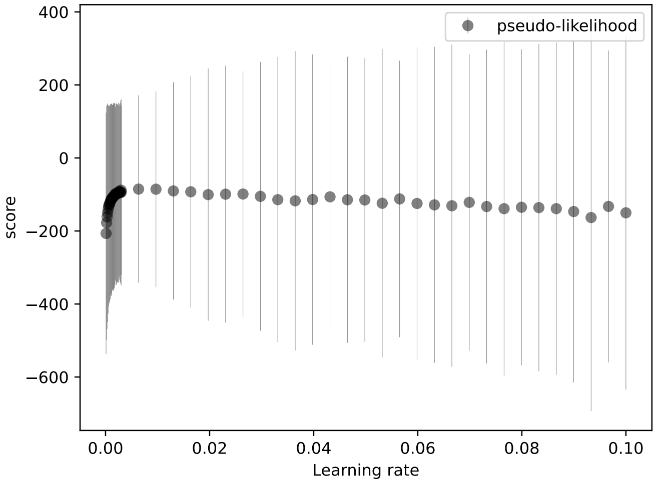

Figure 8: Tuning of the learning rate for the RBM, using 10 epochs and 256 hidden layers. The points represent the mean value and the bars, the standard deviation of the pseudo-likelihood as described in the method score_samples of the BernoulliRBM function from Sklearn [10].

to our criterion, both sets of data have a highly significant probability of being negatively correlated.

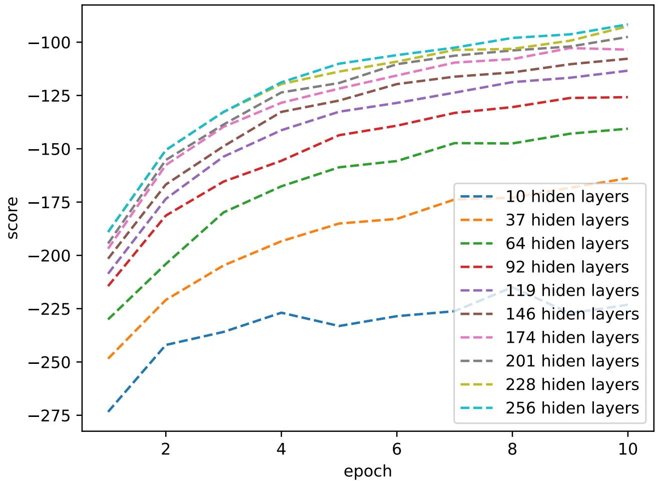

Figure 9: Tuning of the number of hidden layers, using 10 epochs and a learning rate of $\eta = 0.003$.

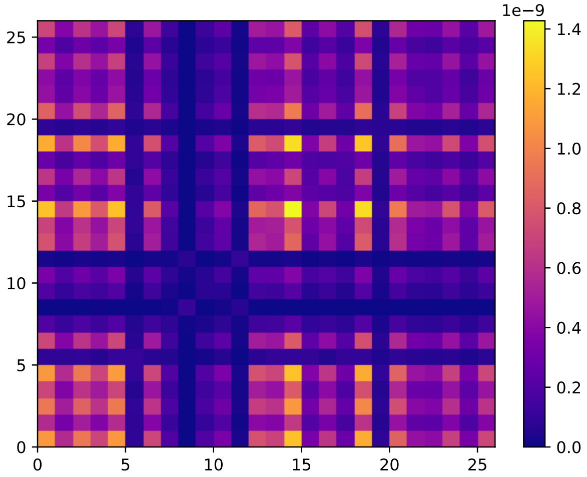

Figure 10: Estimation of the joint probability distribution, where $P_{ij}$ represents the probability distribution of the $i$th letter and the $j$th letter in the alphabet.

# 5 Discussion

In Principal Component Analysis, the principal component is the one that explains most of the variance, taking each letter as coordinates, and only using the values of chroma that the synesthetes associated with each letter, this direction have a strong probability to be inversely correlated with the frequency of letters in English picture books. This means that, if we make random associations between letters and values of chroma and we compare it with the associations made by synesthetes, by ordering the letters by those that diverge the most of the random ones, this ordering would be inversely correlated with the frequency of letters in English picture books.

This correlation was perceived by using the principal component analysis, as well as comparing the associations made by a learning machine (the RBM) and the synesthetes, trying to learn the letters, given the same frequency of letters in the process of learning, but it was not seen by computing the correlation directly with the mean value of chroma. This is not necessarily a problem, one characteristic of the picture books is that there was a very big variance in the frequency of letters across the different books [4], also each person is influenced by many random factors in the moment of learning. As chowed in [20], there were a group of people that learn a big part of their associations from a magnet-colored toy, and in the same way, each person could learn their associations based on completely different factors, making them seem random when they are compared with a specific ordering, like the frequency of letters in picture books. But by using a big sample, and focusing on a method that relies on more than just the mean values, the correlation was made visible.

Chroma is understood by the intensity or saturation of a color, for example, the color perceived by light on a very narrow range of the frequency of blue would have a big value of chroma, while light with the same dominant frequency, but in a more wide range would have a smaller value of chroma. This negative correlation says that letters that are encountered more frequently tend to be associated with less intense colors, while less frequent letters tend to be associated with more intense colors.

This could be explained in many ways, like hypothesizing that the most frequent colors perceived in childhood are less intense, so their brains learn to associate more frequent colors with more frequent letters. One possible explanation that would be interesting to test is by arguing that colors with the biggest value of chroma are harder to code than those with smaller values of chroma.

An optimal coding algorithm will use small codes to codify frequent data and bigger codes for less frequent data, minimizing the coding length, but for the process of learning, it makes sense that is equally important the maximization of the information transmission, as well as minimization of the coding cost [8]. Nevertheless, in general, the most frequent data would be codified with smaller codes. It would be interesting to test if the human brain needs longer internal representations for colors with bigger values of chroma, and in this way, small chroma would mean associations with more frequent words by means of coding optimization.

# References

[1] Hervé Abdi and Lynne J Williams. “Principal component analysis”. In: *Wiley interdisciplinary reviews: computational statistics* 2.4 (2010), pp. 433–459.

[2] Gian Beeli, Michaela Esslen, and Lutz Jäncke. “Frequency correlates in grapheme-color synaesthesia”. In: *Psychological Science* 18.9 (2007), pp. 788–792.

[3] Gregory Cohen et al. “EMNIST: Extending MNIST to handwritten letters”. In: *2017 international joint conference on neural networks (IJCNN)*. IEEE. 2017, pp. 2921–2926.

[4] Nicholas E Fears and Jeffrey J Lockman. “Case-and form-sensitive letter frequencies in children’s picture books”. In: *Early Childhood Research Quarterly* 53 (2020), pp. 370–378.

[5] Asja Fischer and Christian Igel. “An introduction to restricted Boltzmann machines”. In: *Iberoamerican congress on pattern recognition*. Springer. 2012, pp. 14–36.

[6] Geoffrey E Hinton. “A practical guide to training restricted Boltzmann machines”. In: *Neural networks: Tricks of the trade*. Springer, 2012, pp. 599–619.

[7] Carlos Enmanuel Soto López. *ThesisQLS*. `https://github.com/carlossoto362/QLS2021-2022Diploma/tree/main/ThesisQLS`. 2022.

[8] Matteo Marsili and Yasser Roudi. “Quantifying relevance in learning and inference”. In: *Physics Reports* 963 (2022), pp. 1–43.

[9] Frank J Massey Jr. “The Kolmogorov-Smirnov test for goodness of fit”. In: *Journal of the American statistical Association* 46.253 (1951), pp. 68–78.

[10] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[11] Charles A Poynton. “A guided tour of colour space”. In: *New Foundation for Video Technology: The SMPTE Advanced Television and Electronic Imaging Conference*. SMPTE. 1995, pp. 167–180.

[12] Anina N Rich, John L Bradshaw, and Jason B Mattingley. “A systematic, large-scale study of synaesthesia: implications for the role of early experience in lexical-colour associations”. In: *Cognition* 98.1 (2005), pp. 53–84.

[13] Nicholas B Root et al. “Why is the synesthete’s “A” red? Using a five-language dataset to disentangle the effects of shape, sound, semantics, and ordinality on inducer–concurrent relationships in grapheme-color synesthesia”. In: *Cortex* 99 (2018), pp. 375–389.

[14] Amadou T Sanda Mahama, Augustin S Dossa, and Pierre Gouton. "Choice of distance metrics for RGB color image analysis". In: *Electronic Imaging* 2016.20 (2016), pp. 1–4.

[15] Julia Simner et al. "Non-random associations of graphemes to colours in synaesthetic and non-synaesthetic populations". In: *Cognitive neuropsychology* 22.8 (2005), pp. 1069–1085.

[16] Tijmen Tieleman. "Training restricted Boltzmann machines using approximations to the likelihood gradient". In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 1064–1071.

[17] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[18] Jamie Ward. "Synesthesia". In: *Annual review of psychology* 64 (2013), pp. 49–75.

[19] Jamie Ward and Julia Simner. "Synesthesia: The current state of the field". In: *Multisensory Perception*. Elsevier, 2020, pp. 283–300.

[20] Eagleman DM Witthoft N Winawer J. "Prevalence of Learned Grapheme-Color Pairings in a Large Online Sample of Synesthetes". In: *PLoS ONE* (2015). URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118996.