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 18, 2022

Introduction

The hypothesis tested is that the frequency of letters in English picture books is an explanatory factor in letter-color synesthetes. To test it, four methods are used,

1. Linear Regression
2. Analysis based on PCA
3. Clustering
4. Analysis based on training a RBM

Índice

- Letter-Color Synesthesia
    - Letter-Color Synesthesia
    - Explanatory Factors
- Data and Hypothesis
    - Data
    - Hypothesis
- Methods and results
    - Methods
- Appendix
    - Fraction of letters used
    - Computations

Índice

Letter-Color Synesthesia

- Letter-Color Synesthesia
- Explanatory Factors

Letter-Color Synesthesia

Synesthesia can be described as an inducer-cuncurrent pairing, characterized by internal consistency and automaticity. Is not a pathological condition and can exist as a developmental and acquired condition (Ward, 2013).

Índice

Letter-Color Synesthesia

- Letter-Color Synesthesia
- Explanatory Factors

How random it looks like?

For grapheme-color synesthetes, a grapheme will elicit the perception of a color. The color perceived by different graphemes is very consistent over periods of time if tested for the same person, but inconsisten between people. However, the heterogeneity is far from been random (Root et al., 2018).

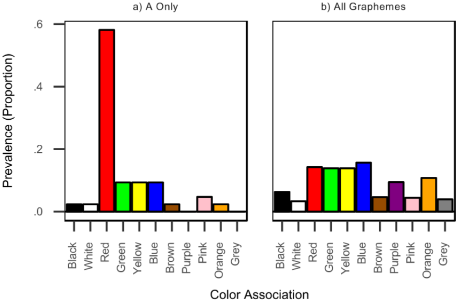


Figure: Proportion of association in 11 colors for the letter “A” (a) and for all the letters in the English alphabet (b) (Root et al., 2018).

Explanatory Factors

Root and collaborators (Root et al., 2018) proposed to called Explanatory Factors to properties that have been shown to influence inducer-concurrent relationships, with synesthetes and non synesthetes.
For example, (Simner et al., 2005) 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.

Índice

Data and Hypothesis

- Data
- Hypothesis

Data

In a paper called "Prevalence of Learned Grapheme-Color Pairings in a Large Online Sample of Synesthetes" (Witthoft N, 2015), Witthoft and collaborators showed that the coincidence between the grapheme-color associations and the color of a letter toy, at least 6% of a total of 6588 synesthetes, is not a random coincidence. They collected the data from a synesthesia battery www.synesthete.org from 6588 synesthetes, which is available online.

Índice

- Data and Hypothesis
  - Data
  - Hypothesis

Hypothesis

The hypothesis is 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.
We used the color associations in the $\mathrm{L^*a^*b^*}$ color space from (Witthoft N, 2015) and the frequency of letters in English picture books studied in (Fears & Lockman, 2020) to test it.

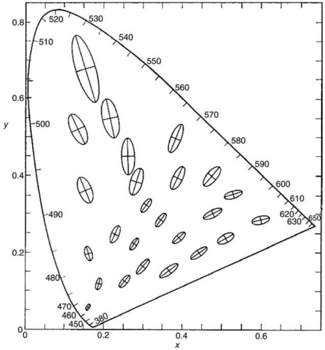


Figure: "MacAdam ellipses 10-times enlarged" Poynton, 1995.

Índice

Methods and results

- Methods

Linear Regression

The R^2 and Kendall Tau statistics were computed between the three components of the cylindrical representation of each color in the $\mathrm{L^*a^*b^*}$ color space, “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” $\mathrm{h_{a,b}} = \arctan\left(\mathrm{a^*}/\mathrm{b^*}\right)$.

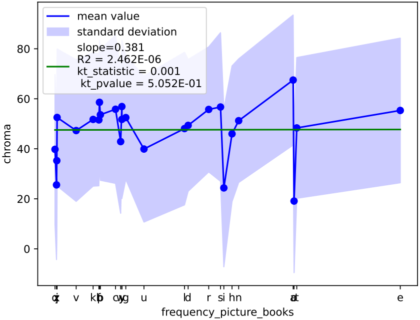


Figure: Linear Regression between the mean value of chroma and the frequency of picture books (Fears & Lockman, 2020).

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.

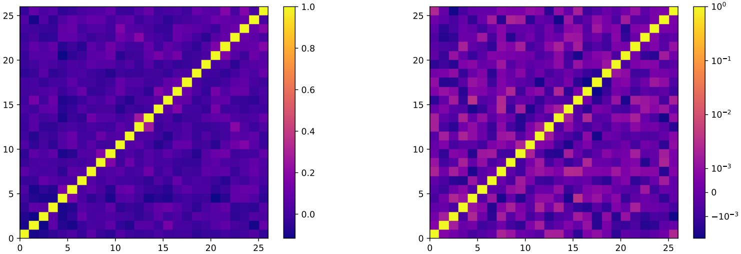


Figure: correlation matrix $\mathbf{C} = \mathbf{X}^T\mathbf{X}$ computed with the three components in the Cartesian representation (left), and the correlation matrix of the shuffle data, where the color that each synesthete associate to each letter where shuffle between all the synesthetes (right).

Principal Component Analysis

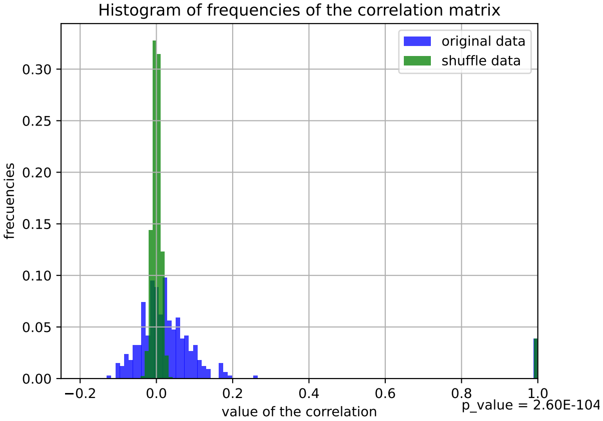


Figure: Histogram of values in the correlation matrix calculated with the three components in the Cartesian representation (blue) and the shuffle data (green). At the right bottom of each histogram, the p-value of the Kolmogorov-Smirnov Test (Massey Jr, 1951).

Principal Component Analysis

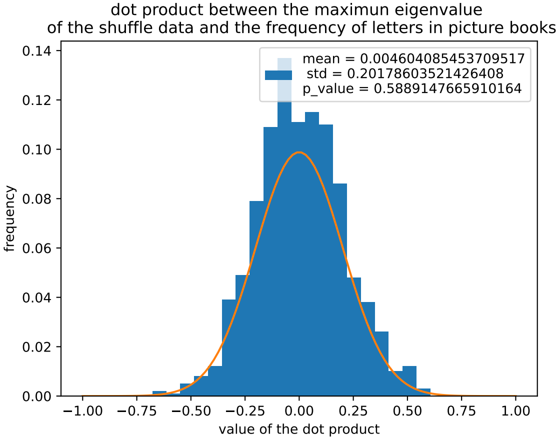


Figure: Histogram of the dot product between the frequency of letters and the eigenvector corresponding with the biggest eigenvalue of 1000 correlation matrices of shuffle sets of data. Dot product obtained for chroma principal value: -0.408.

Clustering

We did nine clusters using 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 (Virtanen et al., 2020). Non of the clusters returned a significant result for our hypothesis.

Restricted Boltzmann Machine

From the EMNIST data Cohen et al., 2017, a RBM was trained, with 47,984 images of lower case letters, intending to match the frequency of letters in English picture books. 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.

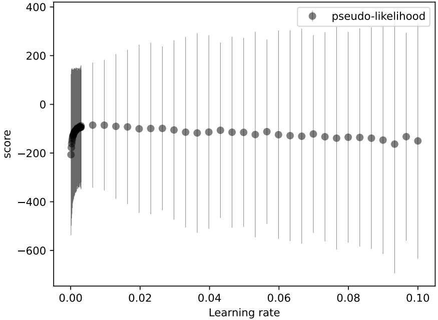
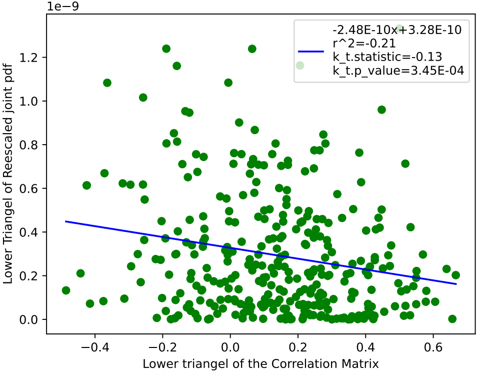


Figure: The points represent the pseudo-likelihood as described in the method score_samples of the BernoulliRBM function from Sklearn (Pedregosa et al., 2011).

Restricted Boltzmann Machine

Then we estimated the joint probability distribution over letters and 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.



Conclusion

- There is a strong probability of been a negative correlation between the chroma that the synesthetes associate to each letters with the frequency of letters in English picture books.

Thanks for your attention...

Índice

Appendix

- Fraction of letters used
- Computations

Fraction of letters in English picture books

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

Table: Mean weighted frequencies of English picture books (Fears & Lockman, 2020).

Índice

Appendix

- Fraction of letters used
- Computations

R square and Kendall Tau

The R^2 statistic between two variables, X and Y is computed as

$$\begin{aligned} R^2 &= \frac{\mathrm{Cov}(X,Y)}{\mathbf{V}(X)\mathbf{V}(Y)}, \\ \mathrm{Cov}(X,Y) &= \mathbf{E}\left[(X-\mathbf{E}[X])(Y-\mathbf{E}[Y])\right] \\ \mathbf{V}(W) &= \mathbf{E}\left[(W-\mathbf{E}[W])^2\right] \end{aligned}$$

The Kendall Tau statistic between two variables, X and Y is computed as

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{(\text{number of pairs})}.$$

The computations and the p-values where obtained using an implementation on SciPy (Virtanen et al., 2020).

Lab Color Space Transformations

The transformations between the $\mathrm{L^*a^*b^*}$ coordinates and the RGB coordinates would follow the relations

$$
\begin{aligned}
L^* &= 116 * f(Y/Y_0) - 16 \\
a^* &= 500[f(X/X_0) - f(Y/Y_0)] \\
b^* &= 200[f(Y/Y_0 - f(Z/Z_0))], \\
f(w) &= (w/w_0)^{1/3}, \text{ if } f(w) > (24/116)^3 \\
f(w) &= (841/108)(w/w_0) + 16/116, \text{ if } f(w) < (24/116)^3
\end{aligned}
$$

$$
\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 2.768892 & 1.751784 & 1.130160 \\ 1.000000 & 4.590700 & 0.060100 \\ 0 & 0.056508 & 5.594292 \end{bmatrix} \cdot \begin{bmatrix} R \\ G \\ B \end{bmatrix}
$$

With X_0, Y_0, Z_0 the values of white D65 $(0.950456, 1, 1.088754)$ (Sanda Mahama et al., 2016).

Estimation of the joint probability distribution

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

With

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

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

and

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

Bibliography

Cohen, G., Afshar, S., Tapson, J., & Van Schaik, A. (2017). Emnist: Extending mnist to handwritten letters. *2017 international joint conference on neural networks (IJCNN)*, 2921–2926.

Fears, N. E., & Lockman, J. J. (2020). Case-and form-sensitive letter frequencies in children's picture books. *Early Childhood Research Quarterly*, *53*, 370–378.

Massey Jr, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, *46*(253), 68–78.

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

Poynton, C. A. (1995). A guided tour of colour space. *New Foundation for Video Technology: The SMPTE Advanced*