Why us? Factors Influencing Membership in Color Naming Clusters

Simin Liu, Applied Data Analysis, Wesleyan University
Quantitative Analysis Center Summer 22 Apprenticeship

Introduction

- Color is a universal concept that exists in every language [1] and color naming provides an excellent case for studying linguistic properties across groups (Figure4).
- Clustering the World Color Survey (WCS) data [2] of 104 monolingual, pre-industrial language groups has found universal pattern that exist across these languages [1]. Specifically, 85% of WCS participants fall into 18 clusters.
- Here, we study the top 6 clusters to understand the common features within each cluster (Figure3). Interestingly, participants from different languages fall into the same clusters based on their characteristics of color [1].

Research Questions

- What are the characteristics of each cluster?
- Who is in each cluster?
- Which factors contribute most to the formation of each cluster?

Methods

Data

N=991 WCS participants of the 6 clusters with members above n>157.

1. Variable	*Age*	*Gender*	*Language name*	*Stage*	*Family*	*Basic color terms*	*Unique words*
2. Type	N	C	C	C	C	N	N
3. Source	WCS [2]	WCS	WCS	WCS	WCS	WCS	WCS
1. Variable	*Country*	*Cluster_id*	*Climate*	*Latitude*	*Longitude*	*Occupation*	*Temp diff/yr*
Type	C	C	C	N	N	D	N
Source	WCS	Study2[1]	Research[3]	Research	Research	Research	Research

Table 1: (N=Numeric, C=Categorical, D=Dummy)

The data is drawn from the WCS archive [2], a past study on WCS participants [1], and personal research into each language group [3].

Measures

- *Cluster membership* is the dependent variable and therefore most of the analyses use chi-square tests and multinomial regression to explore possible relationships between demographic features.
- Additionally, we use the chi-square contingency tables to explore variables that have the largest difference between expected values and actual values.

Figure 1: gender distribution across 6 clusters

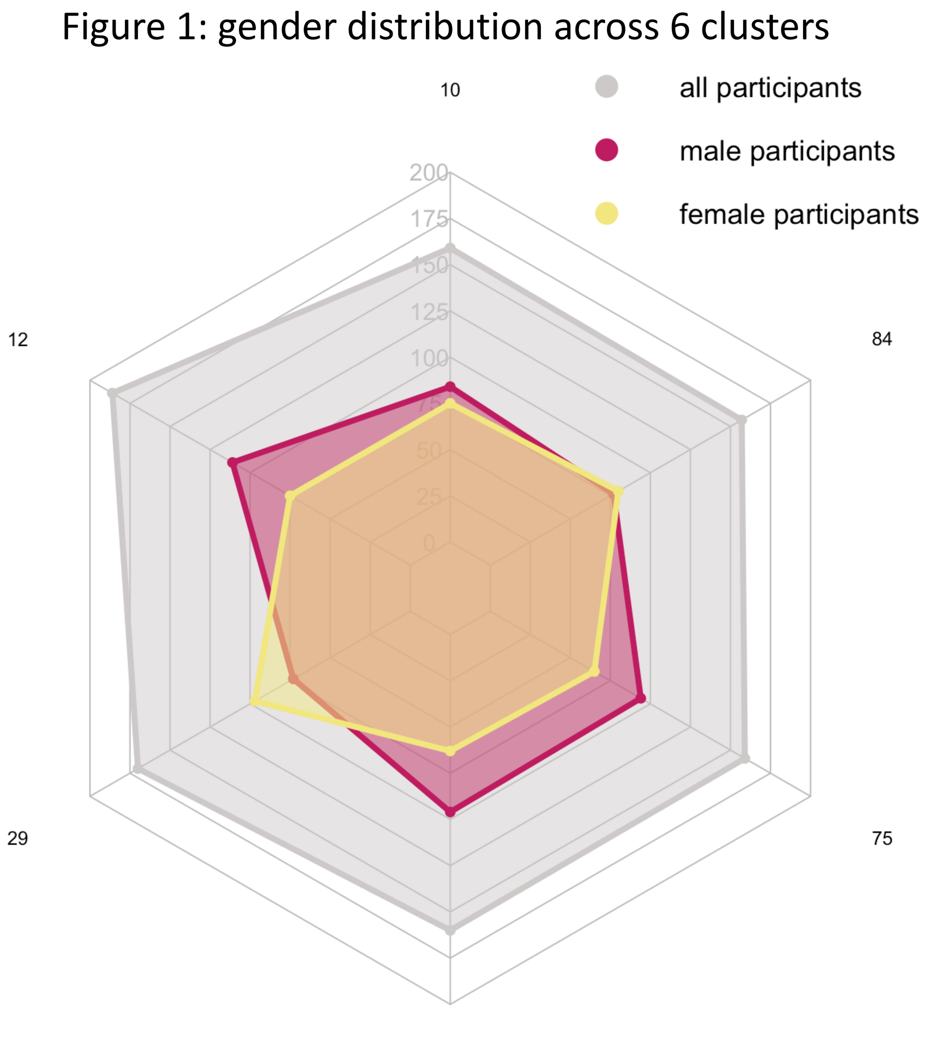
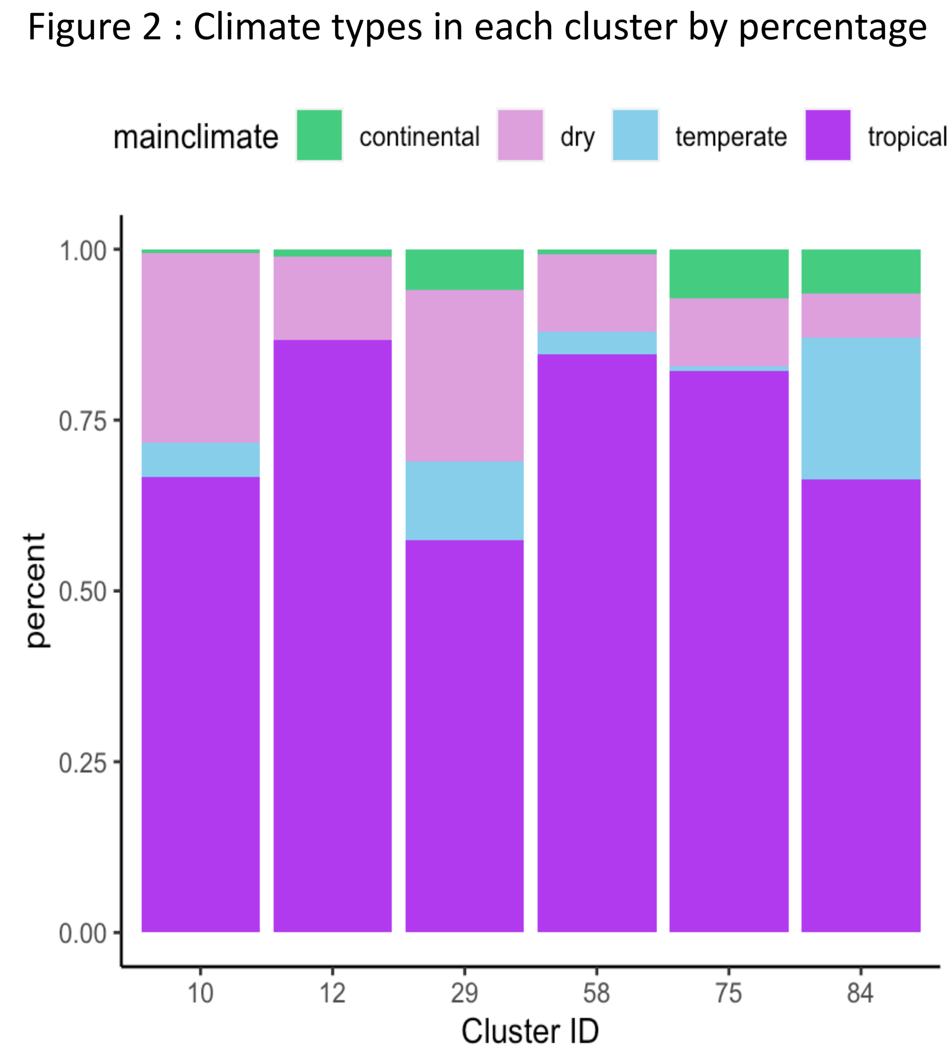


Figure 2 : Climate types in each cluster by percentage



Results

Univariate

- 54.04 % of participants are male (Figure1).
- 74.52 % of participants are from a tropical climate (Figure2).
- 44.5 % of participants have 6-8 unique words (Figure5).

Bivariate

Cluster ID	Main Climate	GenderF	Agriculture1	Hunting1	Gathering1
10(n=159)	R(dry:-18.85)	R(-1.92)	R(4.73)	R(-3.61)	R(-7.03)
12(n=186)	(tropical: -21.63)*	(10.48)	**(-15.31)***	(-22.17)	**(21.22)***
29(n=170)	**(dry:-16.33)***	(-18.87)	**(-13.95)***	(14.05)	**(19.11)***
58(n=160)	(tropical: -17.51)*	(10.07)	(16.4)	(10.81)	(-16.84)
75(n=159)	**(tropical: -14.24)***	(8.07)	(19.73)	(-19.61)	**(-24.03)***
84(n=157)	**(temperate: -21.27)***	(-7.84)	**(-11.6)***	**(20.54)***	**(7.58)***

Cluster ID	Mercantilism1	Latitude	Longitude	Temp Diff
10	R(-0.19)	R	R	R
12	(0.93)	**[-0.014]***	[0.0004]	**[-0.020]***
29	(0.85)	**[0.025]***	[-0.0002]	**[0.049]***
58	(-3.19)	[-0.009]	**[0.0042]***	[-0.002]
75	(0.8)	[0.008]	**[0.0036]***	[-0.012]
84	(0.79)	[0.007]	[0.0008]	**[0.022]***

Table 2: *=significant for multinomial regressions; R=reference level; ()=results from contingency difference tables(expected-actual value) ; []=coefficients from multinomial regressions;

Multivariate

- The model of *cluster_id* as a function of their occupation indicates that having hunting and gathering as occupations plays a significant role in determining unique words ($p < 0.05$).

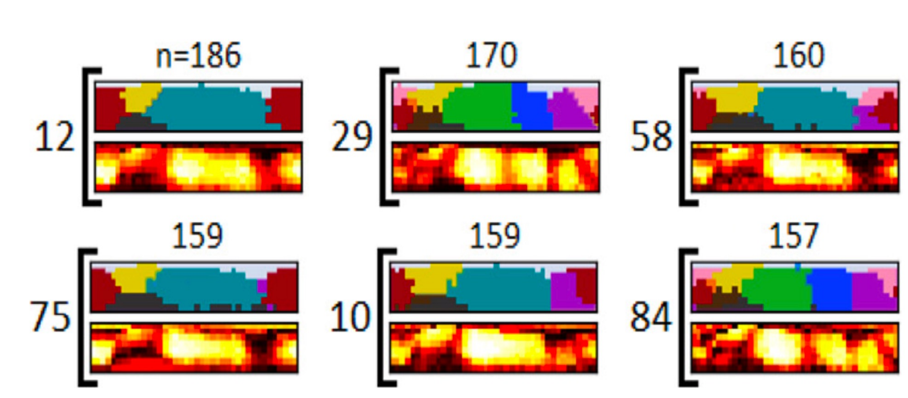


Figure 3: Color naming presentation[1]

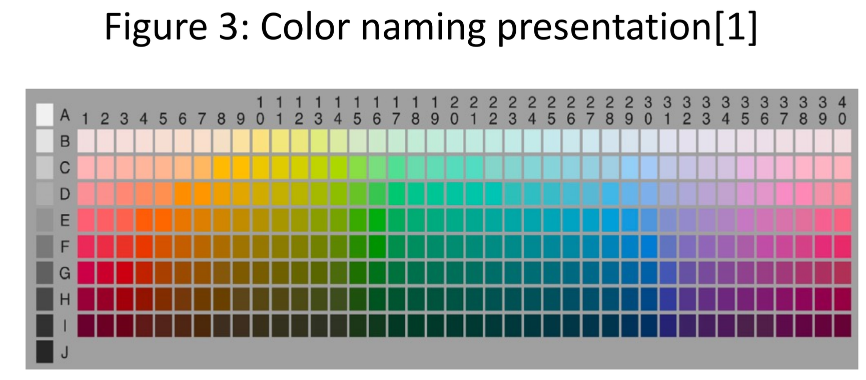
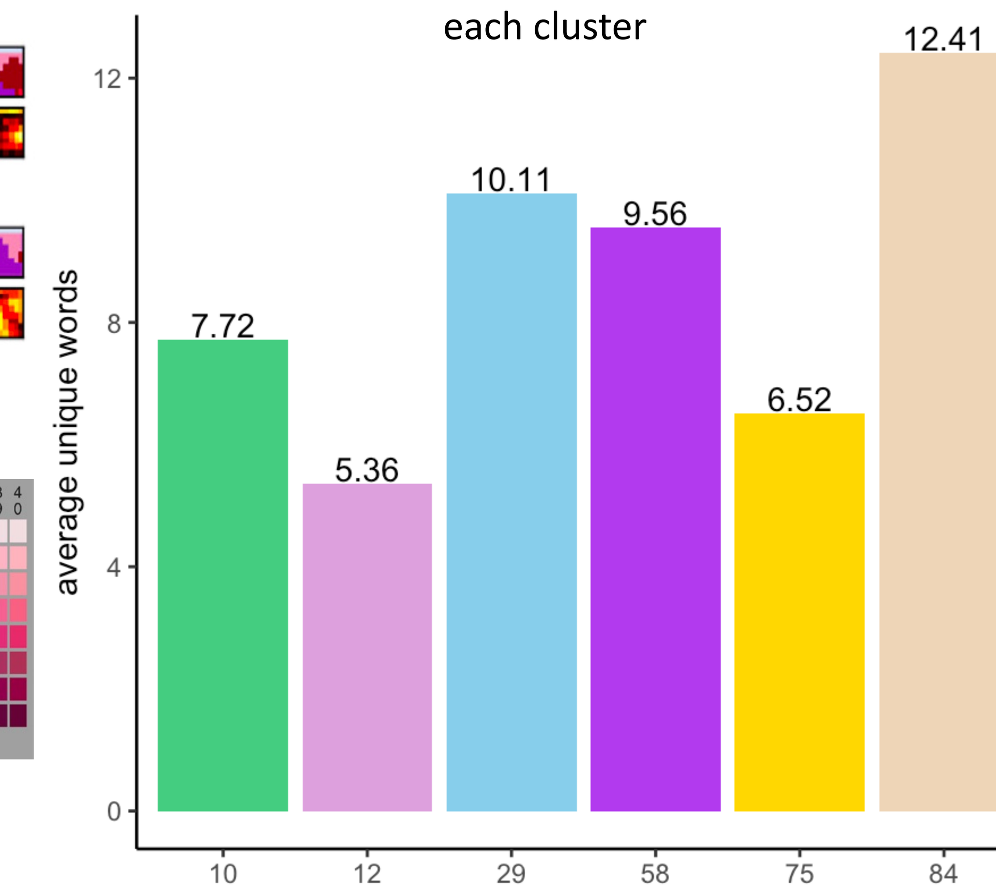


Figure 4: Sample color grids[1]

Figure 5: Average number of unique words of each cluster



Discussions

- Many predictor variables are correlated to each other (e.g. unique words per person and basic color terms for a language). Therefore, this study only focuses on how some independent demographic features influence color naming. Further study is needed to better understand the mechanism behind color naming systems.
- Gender seems to be significant for cluster 29, but the regression result rejects the guess. Further study is necessary.
- Studying the relationships between some demographic features and color naming may help build new hypotheses that will lead to future empirical studies and understand social phenomena[1].

[1]: Joe, K., & Gooyabadi, M. (2021). A bayesian nonparametric mixture model for studying universal patterns in color naming. *Applied Mathematics and Computation*, *395*, 125868. https://doi.org/10.1016/j.amc.2020.125868
[2]: Cook, R., Kay, P., & Regier, T. (2003, June 1). WCS Data Archives. Retrieved July 17, 2022, from http://www.icsi.berkeley.edu/wcs/data.html
[3]: Hammarström, Harald & Forkel, Robert & Haspelmath, Martin & Bank, Sebastian. 2022. Glottolog 4.6. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.6578297 (Available online at http://glottolog.org, Accessed on 2022-07-22.)