The dependence of the blue vs. green distinction in the language on certain linguistic and geographical factors

A Hypothesis

The hypothesis of this study is that the distinction between blue and green colours in the language depends on:

- the number of basic colour categories in the language,
- the number of non-derived colour names in the language,
- red vs. yellow differentiation in the language,

and possibly depends on:

- genealogical classification of the language,
- the macroarea where the language is spoken,
- geographical characteristics: longitude, latitude, altitude, and average annual precipitation.

Research Design

During the research the following variables will be used (the target variable is marked in italics):

№	name	type
1	number of base colour categories in the language	numeric
2	number of non-derived colour names in the language	numeric
3	red vs. yellow differentiation in the language	binary
0	*blue vs. green differentiation in a language*	*binary*
4	genealogical classification of the language (family)	categorical
5	genealogical classification of the language (group)	categorical
6	macroarea	categorical
7	latitude	numeric

8	longitude	numeric
9	altitude	numeric
10	average annual precipitation	numeric

Initially, the red vs. yellow and blue vs. green distinctions are nominal variables with a set of variants containing composites with other colours (for example, black/blue vs. green), but for this study, all variants will be reduced to binary categories. Data on the number of basic colour categories is partially fuzzy and contains intervals from the minimum to the maximum value (for example, 7-7. 5). In this case, the average values will be used.

Among the available data, there are three languages in which one of the studied colours does not exist among the basic ones, so the classification is not applicable to it. It would be interesting to explore these cases, but for the purpose of this research, these objects will be deleted. In addition, the number of objects in different data differs slightly: 120 languages for chapters 132-133 and 119 languages for chapters 133-134. Information about the blue vs. green distinction and yellow vs. red distinction (or the absence of these categories) was not found for the Cahuilla language, so it will be excluded from the study.

To better understand the data, the first stage of the study will be to visualize various combinations of features (inc. using dimensionality reduction methods).

Then statistical tests will be performed to assess the significance of some of these features:

№	**Null hypothesis ($H_{0)}$**	**Alternative hypothesis ($H_{A)}$**	**test**
1	There is no relationship between the number of non-derived colours and the blue vs. green distinction in the language	There is a statistically significant relationship between the number of non-derived colours and the blue vs. green distinction in the language	t-test

2	There is no relationship between the territory altitude and the blue vs. green distinction in the language	There is a statistically significant relationship between the territory altitude and the blue vs. green distinction in the language	
3	There is no relationship between the average annual precipitation and the blue vs. green distinction in the language	There is a statistically significant relationship between the average annual precipitation and the blue vs. green distinction in the language	
4	There is no relationship between the red vs. yellow distinction and the blue vs. green distinction in the language	There is a statistically significant relationship between the red vs. yellow distinction and the blue vs. green distinction in the language	Chi-square, Cramer's V, odds
5	There is no relationship between the macroarea and the blue vs. green distinction in the language	There is a statistically significant relationship between the macroarea and the blue vs. green distinction in the language	
6	There is no relationship between the language family and the blue vs. green distinction in the language	There is a statistically significant relationship between the language family and the blue vs. green distinction in the language	

Two classification models are supposed to be used: logistic regression and decision tree. For logistic regression it is assumed, in addition to the usual formula, to explore the interaction of the number of base colour categories with other features. The significance of the variables obtained by the two models will be compared with each other and with the

results of statistical tests. In addition, the resulting decision tree will be plotted and interpreted.

It will be interesting to predict the blue vs. green distinction for the Cahuilla language removed from the data, although the accuracy of the prediction cannot be verified.

The final stage of the research will be the generalization and interpretation of the results obtained.

The following R packages are to be used for research purposes: tidyverse, ggplot2, vcd, ggfortify, ca, stats, party. Data collection and preprocessing will be performed using python.

Description of Data Collection Method

The research is based on data obtained from the open database of structural linguistic features *World Atlas of Language Structures* (WALS), chapters 132 — 135: *Number of Non-Derived Basic Colour Categories, Number of Basic Colour Categories, Green and Blue, Red and Yellow*. These chapters provide the following information for languages:

1) the number of colours whose names are not derived from other names;

2) the number of basic colour categories;

3) whether red and yellow exist as two separate base colours or a composite (perhaps with some third colour);

4) whether blue and green exist as two separate base colours or a composite (possibly with some third colour).

Information about genealogy, macroarea, latitude and longitude will also be obtained from the WALS site using python. In addition, information about altitude and precipitation for all coordinates will be collected manually using specialized maps and / or from the Internet using python.