Automatic construction of semantic maps: topological machine learning on WCS color data

Mikael Vejdemo-Johansson and Susanne Vejdemo

Computer Vision and Active Perception Lab
KTH Royal Institute of Technology

Department of Linguistics
Stockholm University

June 6, 2013

KTH Computer Science and Communication

Outline

Color universals

Fundamental question

Are there universal tendencies in color naming?

Rephrased question

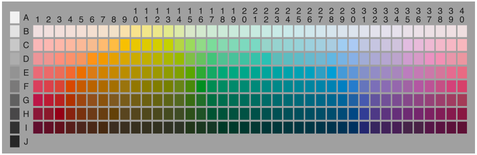
Are there recurring structures or patterns in data representing color naming across languages?

This rephrased question opens up for computational and quantitative methods.

KTH Computer Science and Communication

The World Color Survey

Study



Elicit *color names* for 330 Munsell chips and *color focus* for each used color name.

20 languages for Berlin & Kay 1969; 110 non-written languages for WCS. Dataset released 2003.

Universal inventory of 11 basic color categories, labelled: WHITE, BLACK, RED, GREEN, YELLOW, BLUE, BROWN, PURPLE, PINK, ORANGE, and GREY.

Subsequent work

Group 1:

Kay, Berlin, Maffi & Merrifield 1997 Study the World Color Survey. Refined hierarchy from Berlin & Kay 1969.

Regier & Kay 2003 Used statistical methods to demonstrate support in WCS for the universal color categories in Berlin & Kay 1969.

Group 2:

Lindsey & Brown 2006 Expanded previous studies to study *distributions* of color name responses, and not only single point representatives. Recovered 8 categories. Used *k*-means clustering.

Jäger 2012 Tried to automate a quantitative analysis with statistical tools. Recovered 15 categories. Used PCA.

KTH Computer Science and Communication

Outline

Group 1: ignores distribution shapes

Lindsey & Brown (2006) point out that the earliest work on WCS focuses too hard on single point averages as replacements for entire distributions.

The two first papers to come out of the WCS dataset both compute a single representative color point for each lexeme, and compares these.

Group 2: ignores perceptual distances

The later work by Lindsey & Brown (2006) and by Jäger (2012) have their own methodological issues.

Most fundamental is the inherent choices in the use of k-means and of PCA. These choices carry assumptions about the data.

In particular, the L_2 metric these methods use assumes that different parts of a response are independent.

PCA also carries an interpretative burden.

Colors are not independent

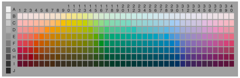
Consider the following example:

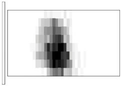
RED	RED	BLUE
RED	PINK	BLUE
RED	RED	BLUE

An L_2 metric assumes that the responses in all three columns are uncorrelated.

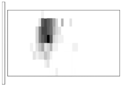
Unexpected results

Using L_2 one might rank some of the 20 occurring color terms from Amuzgo by their similarity to cachuii like this: from left to right the color terms are less and less similar to cachuii. The rank of each term is stated over its distribution.



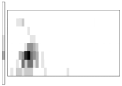


3



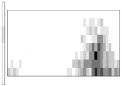
’tsco-tasa-’ndaa

10



calu

11

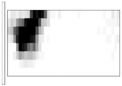


’catsioo

18

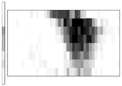
china

19



cajan

20



tsa

KTH Computer Science and Communication

L_2 properties

L_2 produces these results by measuring *overlap* between distributions; and secondarily promote *small* distributions with few responses.

Independence between colors also means that a disjoint distribution can move about arbitrarily wihout affecting distances or ranking. Hue or lightness have no effect on the distance.

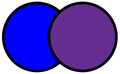
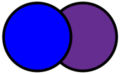
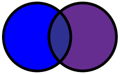
These properties are a particularly large problem for global methods such as PCA.

Interpreting PCA

Subsets as studied by Berlin & Kay are straightforward to interpret.

Principal components do not necessarily represent observed color footprints.

One possibility for the BLUE, INDIGO, PURPLE constellation Jäger observed could be that the indigo footprint is an area that shifts allegiance between languages:

  both combinations of 

but not occurring as a color of its own.

KTH Computer Science and Communication

Outline

1. The World Color Survey
2. A methodological critique
3. Solution attempts

How can we measure color distributions?

First, we need a measure of color differences.

The Commission Internationale de l'Éclairage (CIE)[1] has defined a perceptual color space: CIELAB, with a perceptually constructed color distance measure ΔE.

Second, we need to be able to compare distributions using an underlying distance measure.

There are several possibilities available:
Quadratic form distance
Quadratic χ^2
Earth Mover's Distance

[1] International commission on illumination

Earth Mover's Distance

The Earth Mover's Distance has a particularly accessible interpretation: measures minimal *work* to redistribute piles of sand from one shape to another.

Can be computed using a *linear program solver*. Using IBM's industrial strength solver, we computed this metric for mass distributions assigning to a Munsell cell the % of speakers of that language who used that term for that cell.

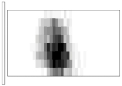
Total computation took just under 1 week.

We will be releasing the resulting dataset for research use.

KTH Computer Science and Communication

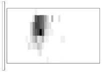
Amuzgo revisited

Again, we rank some color terms from Amuzgo by their similarity to cachuii with their ranks.



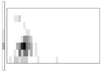
L_2

3



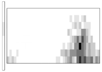
’tsco-tasa-’ndaa

10



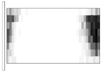
calu

11



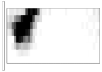
’catsioo

18



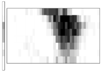
china

19



cajan

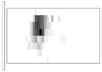
20



tsa

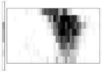
Earth Mover’s Distance

2



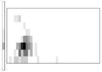
’tsco-tasa-’ndaa

4



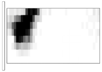
tsa

7



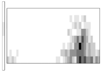
calu

13



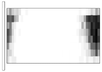
cajan

16



’catsioo

17



china

Additional properties of Earth Mover's Distance

Since the Earth Mover's Distance has a high level of abstraction, very little is assumed about the datapoints compared.

In particular, since ΔE provides a distance measure between any CIELAB colors, EMD can measure distances between distributions on different color grids in CIELAB space.

Thus, data sets such as Berlin & Kay (1969) with 329 chips, the World Color Survey with 330 chips and EoSS with 84 chips can all be compared in the same framework.

KTH Computer Science and Communication

Complementary approach

Instead of, or as well as, changing the metric, we can use methods that work on a local scale.

Topological methods are robust to metric flaws

- Data analysis based on *similarity* not *distance*: inherently local.
- Less sensitive to choice of metric.
- Less sensitive to metric being a *nice type*, as opposed to PCA or machine learning methods.

Mapper – structured clustering

Clustering methods work locally; but are sensitive to connected data. One alternative to clustering methods is Mapper:

Mapper

- Analysis technique invented at Stanford 2008.
- In use for *knowledge discovery* in bio-informatics.
- Clusters, but locally with a view towards cluster connectivity.
- Detects flares, that correspond to potential universals.

In the following demonstration, we use Mapper to discover structures in WCS data. Each point represents a collection of lexemes. The size encodes the number of lexemes. Blue (low) to Red (high) indicates a *measurement function* used in the analysis. Lines encode connectivity between clusters.

Thank you for listening

Our data and tools will be available soon through http://wcs.appliedtopology.org.

Thanks also go to:

- EU FP7 and the TOPOSYS project.
- Knut och Alice Wallenbergs Stiftelse – Unga Forskare.
- Special doctoral programme in language and linguistics – FoSprak
- Carl-Henrik Ek & Omid Aghazadeh for methodological ideas
- The Department of Linguistics, Stockholm University for presentation feedback