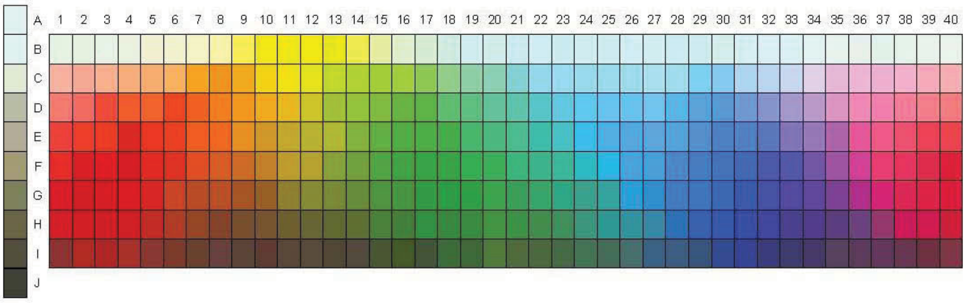
Gerhard Jäger
University of Tübingen, Institute of Linguistics
gerhard.jaeger@uni-tuebingen.de

Color naming universals: A statistical approach

The paper describes a quantitative investigation of the distribution of color naming systems across the languages of the world, using the data from the World Color Survey (WCS, see Cook et al. 2005 for details). Working with data from individual speakers, we found that (1) different categorization systems are distributed according to a power law, and (2) the systems of possible color naming patterns proposed in the literature so far only provide an imperfect description of the data. We (3) propose an alternative system of universals that provide a better fit of the empirical findings.

The WCS researchers collected field research data for 110 unwritten languages, working with an average of 24 native speakers for each of these languages. During this investigation, the Munsell chips were used, a set of 330 chips of different colors that cover 322 colors of maximal saturation plus eight shades of gray. Figure 1 displays them in form of the Munsell chart.

FIGURE 1. THE MUNSELL CHART.



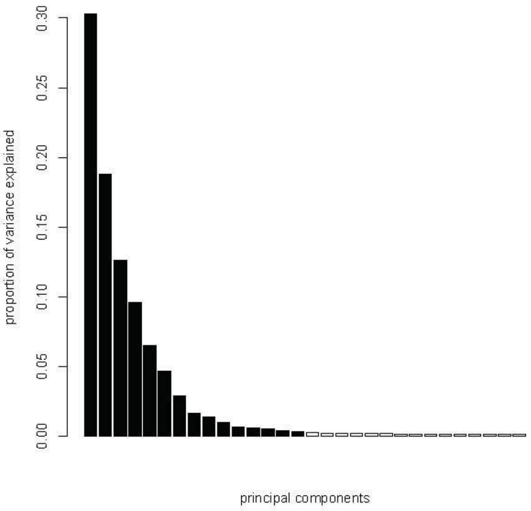
For the WCS, each test person was “asked to name each of 330 Munsell chips, shown in a constant, random order” (quoted from the WCS homepage). The data from this survey are freely available from the WCS homepage http://www.icsi.berkeley.edu/wcs/data.html.

For each informant, the outcome of the categorization task defines a partition of the Munsell space into disjoint sets — one for each color term from their idiolect.
For the present study, these data were organized in a contingency table. It has 1,601 rows — one for each term that was used by at least one of the 1,771 test persons, and 330 columns — one for each Munsell chip. Each cell contains the number of test persons that used the term corresponding to the row to name the chip corresponding to the column.

To normalize vector lengths while at the same time preserve the statistical weight of often-used color terms, each row was divided by the number of speakers that used the corresponding term at least once, and afterwards copied as many times as there were speakers that used the corresponding term.

Figure 2 depicts the proportion of the total variance in the data that are explained by the principal components. The first 15 principal components jointly explain about 91.6% of the total variance in the data. After applying the Varimax algorithm, the resulting 15 extracted features have a clear interpretation: green, white, red, yellow, black, blue, purple, pink, brown, light blue, olive green, gray, orange, violet, pastel green. Further details on the feature extraction procedure can be found in Jäger (2010).

FIGURE 2. PROPORTION OF TOTAL VARIANCE EXPLAINED BY PRINCIPAL COMPONENTS.



The first six features thus extracted correspond to the primary colors that play a central role in Kay et al.'s (1997) system of color naming universals. They assume that a system of basic color terms partitions the color space into disjoint but jointly exhaustive regions. It is claimed that almost all languages partition the primary colors into one of the systems given in Table 1 (above the horizontal line).

For each informant, each term is represented as a 15d vector after dimensionality reduction. Likewise, each primary color is a vector in this low-dimensional space. Thus the data for each speaker implicitly define a partition over the six primary colors: each primary color is assigned to (the vector representing) the term with which it has the largest cosine, i.e. the largest similarity.

The frequencies of the different partition types are distributed approximately according to a power law, as can be seen from Figure 3. The partition types are ordered according to their frequency. The distribution almost follows a straight line. This means that it can be approximated by a power law

$$fr \sim r^c,$$

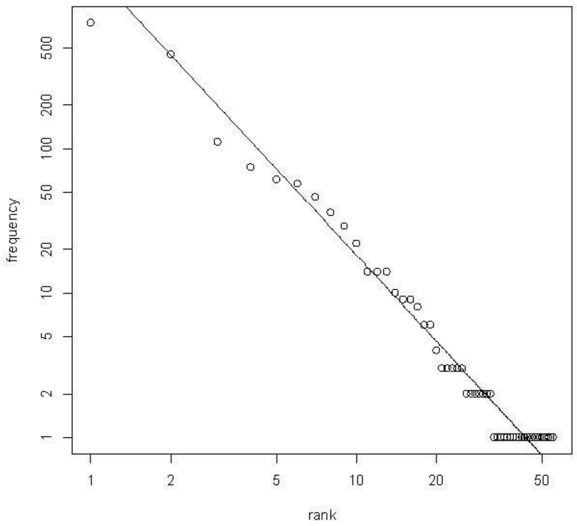
where *fr* is the frequency, *r* is the rank, and *c* is a constant coefficient. Using the methods described in Clauset et al. (2009), we estimate $c \approx -1.99$.

All of the partition types from Table 1 except the leftmost one are attested. The frequencies of each type are indicated in the table. Additionally, there are two partition types not mentioned by Kay et al. (1997) (or any other author that I would be aware of) which occur in substantial numbers. They are displayed in the table below the horizontal line.

TABLE 1. PARTITION HIERARCHY ACCORDING TO (KAY ET AL. 1997).

I	II	III	IV	V
		white red/yellow green/blue black 45	white red yellow green/blue black 737	
white/red/yellow black/green/blue 0	white red/yellow black/green/blue 111	white red/yellow green black/blue 7		white red yellow green blue black 446
		white red yellow black/green/blue 57	white red yellow green black/blue 79	
		white red yellow/green/blue black 29	white red yellow/green blue black 9	
		white red yellow/green black/blue 22		
	white/yellow red green/black/blue 36	white/yellow red green/blue black 61		

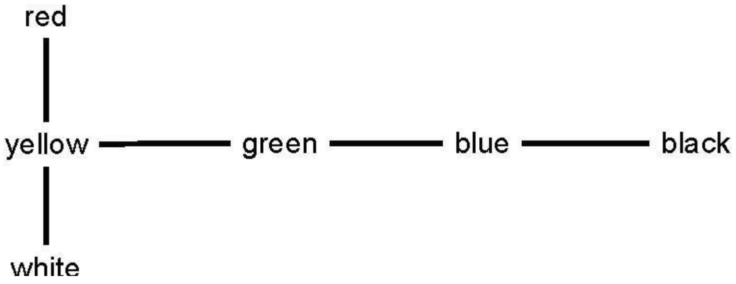
FIGURE 3. FREQUENCY DISTRIBUTION OF PARTITION TYPES OVER THE PRIMARY COLORS.



So while Kay et al.'s model is not completely off the mark, the data from the WCS do not confirm it very strongly either. Only 87.1% of all informants conform to their model. The ten partitions that, according to their model, are the only ones possible occupy the ranks 1, 2, 3, 4, 6, 7, 9, 10, 16, and 18 in the list of the attested partition types.

A closer manual inspection of the partitions in Table 1 shows a certain pattern though. All partition cells are continuous sub-graphs of the connection graph given in Figure 4. Also,

FIGURE 4. CONNECTION GRAPH.



all attested partitions in Table 1 obey the following constraints:

(1) a. All partition cells are continuous sub-graphs of the connection graph.

b. No partition cell has more than three elements.

c. *Red* and *white* only occur in cells with at most two elements.

Next to the 12 partitions in Table 1, there are three partitions obeying these constraints, which are all attested in the data:

- {green}, {white/yellow}, {red}, {black/blue} (14 occurrences)
- {green}, {white/yellow}, {red}, {black}, {blue} (8 occurrences)
- {green}, {white}, {red/yellow}, {black}, {blue} (2 occurrences)

About 94% of all data points in the WCS are captured by this model, and all partition types that are predicted to be possible are in fact attested.

References

Clauset, Aaron, Shalizi, Cosma Rohilla and Newman, M. E. J. 2009. Power-law distributions in empirical data. *SIAM Review*, 51(4):661-703.

Cook, Richard, Paul Kay and Terry Regier 2005. The world color survey database: History and use. In *Handbook of Categorisation in the Cognitive Sciences*, Cohen, Henry and Lefebvre, Claire (eds.), 223-242. Amsterdam: Elsevier.

Jäger, Gerhard 2010. Natural color categories are convex sets. In *Logic, Language and Meaning. 17th Amsterdam Colloquium*, Aloni, Maria, Bastiaanse, Harald, de Jager, Samson Tikitu and Katrin Schulz (eds.) 11-20, New York/Heidelberg: Springer.

Kay, Paul, Berlin, Brent, Maffi, Luisa and Merrifield, William 1997. Color naming across languages. In *Color categories in thought and language*, Hardin, C. L. and Luisa Maffi (eds.) 21-58, Cambridge: Cambridge University Press.