Encyclopedia of Color Science and Technology
DOI 10.1007/978-3-642-27851-8_54-1

Color Dictionaries and Corpora

Angela M. Brown*
College of Optometry, Department of Optometry, Ohio State University, Columbus, OH, USA

Definition

In the study of linguistics, a **corpus** is a data set of naturally occurring language (speech or writing) that can be used to generate or test linguistic hypotheses. The study of color naming worldwide has been carried out using three types of data sets: (1) corpora of empirical color-naming data collected from native speakers of many languages; (2) scholarly data sets where the color terms are obtained from dictionaries, wordlists, and other secondary sources; and (3) philological data sets based on analysis of ancient texts.

History of Color Name Corpora and Scholarly Data Sets

In the middle of the nineteenth century, color-name data sets were primarily from philological analyses of ancient texts [1, 2]. Analyses of living languages soon followed, based on the reports of European missionaries and colonialists [3, 4]. In the twentieth century, influential data sets were elicited directly from native speakers [5], finally culminating in full-fledged empirical corpora of color terms elicited using physical color samples, reported by Paul Kay and his collaborators [6, 7]. Subsequently, scholarly data sets were published based on analyses of secondary sources [8, 9]. These data sets have been used to test specific hypotheses about the causes of variation in color naming across languages.

From the study of corpora and scholarly data sets, it has been known for over 150 years that languages differ in the number of color terms in common use. Particularly, languages differ greatly in how they name the cool colors that are called "*blue*" and "*green*" in English (Fig. 1). Some languages, such as English, use a word *BLUE* that means only blue, in conjunction with a word *GREEN* that means only green. Other languages use a single term (here and elsewhere, "*GRUE*") that means green or blue, and still other languages use a word (here, "*BLACK*") that means both black and blue, to name the cool colors, in conjunction with *WHITE*, which names the light and warm colors.

Scholars in the nineteenth century established the two general explanations for this diversity of color terms across languages, which still guide much of the research on the topic today. The first explanation was that the people who spoke languages with few color terms had deficient color vision. This speculation was at first based on the philological analysis of extinct languages and arose in part because of general interest in the theory of evolution in the latter half of the nineteenth century. Proponents of this view speculated that humans and their color vision had evolved since ancient times. The second explanation was that people living at different times and in different cultures need to differentiate between different colors, so their languages have different numbers of color terms. Particularly, ancient languages lived in simpler times and consequently had fewer color terms in their lexicons.

*Email: brown.112@osu.edu

Encyclopedia of Color Science and Technology
DOI 10.1007/978-3-642-27851-8_54-1

The Color Deficiency Explanation

The earliest scholar to study this variability across languages was **William Ewart Gladstone,** prime minister of England over the latter half of the nineteenth century and scholar of ancient Greek. Gladstone reported that Homer's epic poems used a "paucity" of color terms, mostly relating to dark and light, with a few instances of terms that may have corresponded to *YELLOW*, *RED*, *VIOLET*, and *INDIGO*, but not *GREEN* and not *BLUE* [1]. The German philologist **Lazarus Geiger** [2] reviewed evidence from even older sources: the Hindu Veda hymns of India, the Zend-Avesta books of the Parsees, and the Old Testament of the Bible, as well as ancient Greek and Roman sources. Geiger argued that color lexicons progressed over time from a *BLACK*-and-*WHITE* system to a *BLACK*-and-*RED* system (where *RED* was his term for white or warm colors), then differentiating *YELLOW*, then adding *GREEN*, then *BLUE*. "In the earliest mental productions that are preserved to us of the various peoples of the earth . . . notwithstanding a thousand obvious and often urgently pressing occasions that presented themselves, the colour blue is not mentioned at all. . . . Of the words that in any language that are used for blue, a smaller number originally signified green; the greater number in the earliest time signified black" [Ref. 2, pp. 49, 52].

Gladstone speculated that the Greek of the heroic age "had a less-evolved color sense that prevented him from seeing and distinguishing the many colors that modern people can see easily" [1] [p. 496]. Geiger came to a similar conclusion: "Were the organs of man's senses thousands of years ago in the same condition as now. . .? [p. 60] The circumstance that the colour-terms originate according to a definite succession, and originate so everywhere, must have a common cause. This cause cannot consist in the primarily defective distinction merely. . .. [W]e must assume a gradually and regularly rising sensibility to impressions of colour."

The Cultural Explanation

Under the influence of the Darwinian thinking of the day, the English writer **Grant Allen** and the German ophthalmologist **Hugo Magnus** [23] thought that "primitive tribes" who lived in modern times could provide information on the color naming and sensory color capabilities of ancient humans. Therefore, they sent questionnaires to Christian missionaries, explorers, and diplomats around the world, asking them about the color capabilities of the people they encountered and the color terms in their informants' native languages. Based on their responses, Allen wrote that "the colour-sense is, as a whole, absolutely identical throughout all branches of the human race" [Ref. 3, pp. 205] and afforded "a reasonable presumption in favour of a colour-sense in the earliest members of the human race." He therefore rejected the view, espoused by Gladstone and Geiger, that the reduced color vocabulary observed in many ancient and modern languages was due to a color vision defect. Magnus partly agreed, with qualification: "While some groups confirmed an awareness of colour, which rated in no way below that of the achievements of highly developed nations, others again gave proof of the lack of ability in identifying colours of middle- or short-wavelengths, and this was noted particularly in relationship to 'blue'" [Ref. 4, p. 145].

Based on the results of his surveys, Grant Allen proposed the second, cultural explanation of the diversity of color naming worldwide. "Words arise just in proportion to the necessity which exists for conveying their meaning. . . . Primitive man in his very earliest stage will have no colour terms whatsoever. . . . But when man comes to employ a pigment, the name of the pigment will easily glide into an adjectival sense. . . . [p. 259]. The further differentiation of the colour-vocabulary . . . is most developed among . . . dyers, drapers, milliners, and others who have to deal with coloured articles of clothing. . .." "How then are we to explain the singular fact, which Mr. Gladstone undoubtedly succeeds in proving, that the Homeric ballads contain few actual colour-epithets? In the following manner, it seems to

DOI 10.1007/978-3-642-27851-8_54-1

me. Language is at any time an index of the needs of intercommunication, not of the abstract perceptions, of those who use it."

Hugo Magnus became interested in the discrepancy between the excellent color awareness of many of the peoples in his survey and their difficult color naming. His summary of how color terms co-occur in languages is reminiscent of the results of Gladstone and Geiger: ". . .while in some. . .communities the known terminology begins and ends with 'red', it stretches in other ones well beyond the 'yellow,' and with yet others, even beyond the 'green'." Magnus began his research under the influence of Gladstone and Geiger, but in the end he was also influenced by Grant Allen's work. Magnus concluded, "one might be tempted to formulate a . . . natural law of awareness – be that linguistically engendered or physiologically-anatomically conditioned as part of the natural growth of man."

Empirical Studies

The empirical tradition in the study of color terminology began with **W. H. R. Rivers**, a medical doctor and anthropologist, who traveled as an explorer to several parts of the world on behalf of the Royal Anthropological Institute. In his book *Reports of the Cambridge Anthropological Expedition to Torres Straits* [5], he compared the color vocabularies of three languages spoken by the Kiwai, Murray Islanders, and Western Tribes of the Torres Straits. ". . .As regards blue, the three languages may be taken as representatives of three stages in the evolution of a nomenclature for this colour. In Kiwai there is no word for blue; may blues are called names which mean black. . . while other blues are called by the same word which is used for green. In Murray Island there is no proper native term used for blue. Some of the natives, especially the older men, use [a native term], which means black, but the great majority us a term borrowed from English. . .. The language of the Western Tribe of Torres Straits presents a more developed stage. . .[a native term]. . .is used definitely for blue, but is also used for green. . . .however, traces of the tendency to confuse blue and black still persist. . .." Rivers also reviewed scholarly evidence from ancient and contemporary sources, including his own work in Egypt and the Andaman Islands. All the empirical evidence he reviewed supported his view that the naming of blue is highly variable across cultures: some call blue things *BLACK*, some call them *GREEN*, and some call them with a particular word for *BLUE*. He believed that *BLACK* was the most ancient term, *GREEN* was used in more developed societies, and *BLUE* was the most advanced color term.

In the twentieth century, the large-scale study of color naming across many languages was dormant until 1969, when **Brent Berlin** and **Paul Kay** published their monograph *Basic Color Terms: Their Universality and Evolution* [6]. Berlin and Kay collected a corpus of empirical color-naming data on 20 languages on individual speakers who lived in the San Francisco Bay area in the mid-1960s. They showed each subject an array of Munsell color samples and asked them to indicate the range of colors they assigned to each color term in his or her native language. Berlin and Kay augmented their corpus with scholarly data on the color lexicons of 78 additional languages, which were obtained from dictionaries and other scholarly sources. Berlin and Kay observed that all the color terms in all the color lexicons in their data set were drawn on a superset of only 11 universal **basic color terms:** *BLACK*, *WHITE*, *RED*, *YELLOW*, *GREEN*, *BLUE*, *BROWN*, *ORANGE*, *PINK*, *PURPLE*, and *GRAY.* They also observed that these color terms occurred together in only about seven different combinations. They speculated that these seven combinations of basic color terms represented seven ordered **stages** along an **evolutionary sequence** whereby the most primitive languages distinguish only *BLACK* and *WHITE*, and other color terms are added in a fixed sequence, until all 11 color terms are present. Berlin and Kay assigned each language in their data set to one of their seven stages of color term evolution. Their idea about the

DOI 10.1007/978-3-642-27851-8_54-1

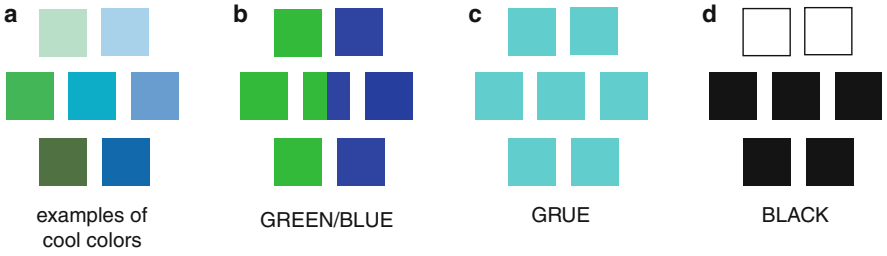

Fig. 1 (**a**) Examples of cool colors; (**b**–**d**) false color coding of the corresponding color terms. The center sample of the diagram is called *BLUE* by some informants and *GREEN* by others in (**b**), but it is called *GRUE* by all informants in (**c**), and it is called *BLACK* by all informants in (**d**)

evolution of color terms was in line with the ideas advanced by Gladstone, Geiger, Allen, Magnus, and Rivers, although their explanation for the evolution of color terms was more in line with Allen's.

The methodology of Berlin and Kay and their theoretical interpretation of their data were criticized by others [e.g., Ref. 10]. Therefore, in the 1970s, Kay, Berlin, and their colleagues collected a new corpus of data on 110 world languages: the **World Color Survey**. The languages in the World Color Survey (WCS) were mostly unwritten and were spoken in traditional societies with limited contact with Western industrialized culture. The geographical distribution of the WCS languages was generally quite similar to the worldwide distribution of all living languages (www.ethnologue.com/show_map.asp?name=World&seq=10). The WCS data set was made up of empirical color-naming data provided in face-to-face interviews by about 24 speakers of each language. Each subject viewed 330 color samples, one at a time, and provided the color term they used in everyday life. Kay, Berlin, Maffi, Merrifield, and Cook published the *World Color Survey* [7], a book-length analysis of this corpus in which they identified each color term in each language with 1 of the 11 basic color terms of Berlin and Kay and updated their theory of color term evolution. They assigned each language to one of five stages, with two stages having three versions each, in their revised theory.

The **World Color Survey** corpus of color terms is available online and has been analyzed by Paul Kay and his colleagues [11, 12], who found evidence of universal color categories across the WCS languages. Independently, Lindsey and Brown [13] performed a cluster analysis of the color-naming patterns in the WCS corpus and discovered about eight distinct clusters of chromatic color terms, which, with the addition of the three achromatic terms *BLACK*, *WHITE*, and *GRAY*, corresponded approximately to the 11 basic color terms of Berlin and Kay. They further found that these color terms fell into about four color-naming systems ("motifs") [14], which corresponded only loosely to the seven stages of Berlin and Kay or the five stages of the WCS. Similarly to all previous scholars and investigators, Lindsey and Brown found that the motifs differed most prominently in the color terms used to name cool colors that speakers of English call "*blue*." In correspondence with the four motifs, some informants called blue samples *DARK*, some called them *GRUE*, some called them *BLUE*, and a few individuals called blue samples *GRAY* (a color term that was also used for middle-value neutral samples). Almost every World Color Survey language revealed individual differences among its speakers, and at least three of the four motifs were represented among the speakers of most languages [see also Ref. 15]. Previous empirical and scholarly work that sought the color terms in each language as a whole, including the data sets in Fig. 2, could not reveal this prominent variation among individuals.

In the tradition of Allen, Kay and his colleagues [e.g., Ref. 16] argued that larger color lexicons, at a later stage along their evolutionary sequence, occur in technologically advanced, economically developed cultures, where the presence of colored artifacts and trade with other cultures requires a larger, more nuanced color vocabulary. Several authors [17, 18] have examined this hypothesis by comparing the

DOI 10.1007/978-3-642-27851-8_54-1

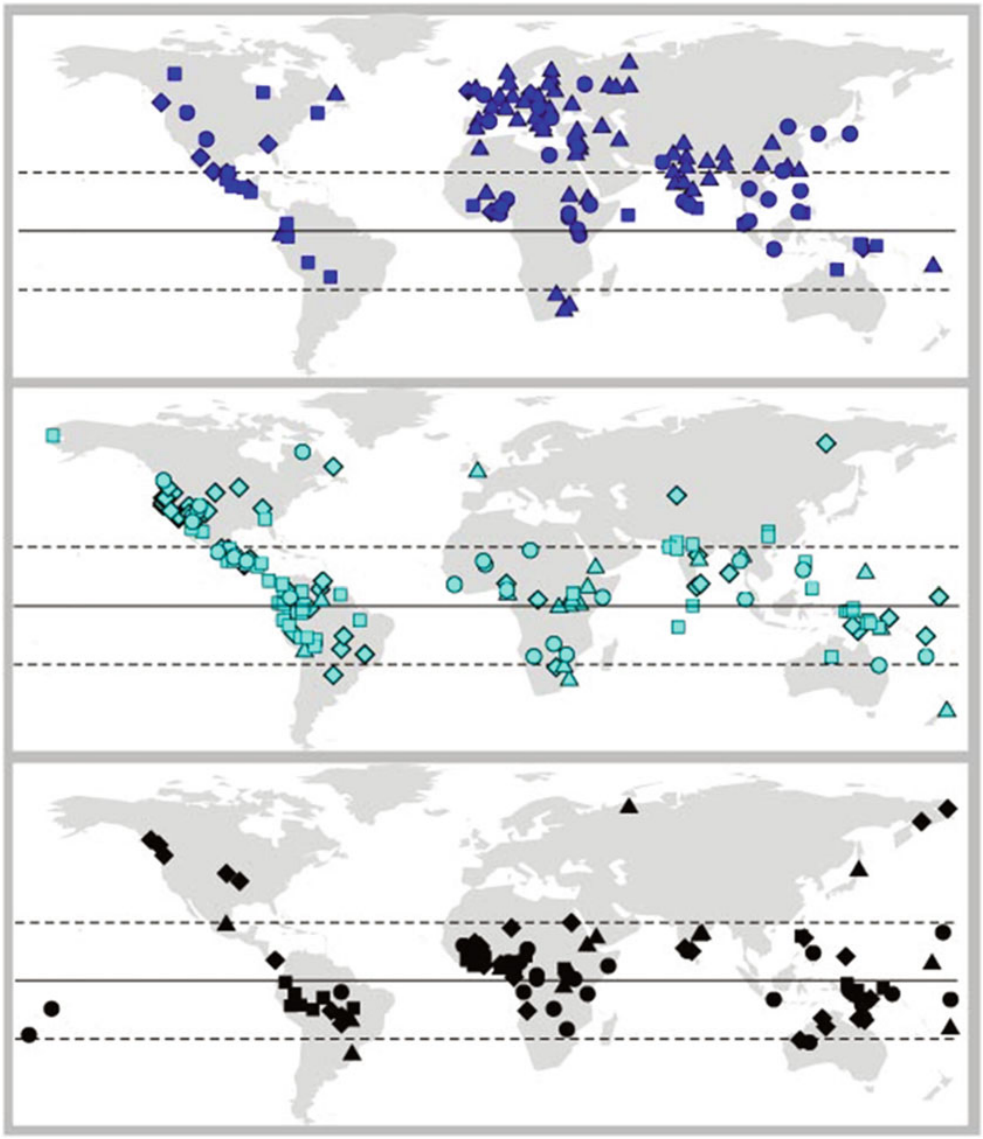

Fig. 2 The geographical distribution of languages that use BLUE/GREEN (top), GRUE (middle), or BLACK (bottom) to name the cool colors is known to be uneven worldwide. Maps show the localities of 371 living languages from color term corpora and scholarly data sets. Circles, 20 corpus languages and 77 living languages from scholarly sources from Berlin and Kay [6]; squares, 106 additional corpus languages from the World Color Survey [7], excluding two already included in [6] and two that could not be typed; diamonds, 75 additional scholarly sources from Bornstein [8]; triangles, 93 additional scholarly sources from Brown and Lindsey [9]. Colonial languages (e.g., English, French, Spanish) are plotted where they were spoken in 1492 CE. Assignment of languages to BLUE/GREEN, GRUE, or BLACK types was made by the authors of the original data sets

number of terms that Berlin and Kay ascribed to each language to its level of development as published by [19].

Modern investigators have examined this diversity across languages in the naming of blue, which was common to the analyses of Gladstone, Geiger, Magnus, Berlin and Kay, and the WCS. Marc H. Bornstein

DOI 10.1007/978-3-642-27851-8_54-1

[8] assembled a set of data on the presence or absence of *BLUE* in 145 languages from published sources and showed them on a world map. His analysis revealed a pronounced latitude effect: *BLACK* and *GRUE* languages tended to be spoken near the equator, and *BLUE* languages tended to be spoken at temperate latitudes (Fig. 2). Bornstein attributed this to the possible geographical variation in intraocular pigmentation, including the amount of melanin in the eye (the pigment epithelium and the iris) and the tint of the ocular media (the lens and macular pigment). Lindsey and Brown [20] reported that the yellow tint of the ocular lens, caused by the intense UVB in equatorial sunlight, could produce changes in color-naming behavior that were similar to those observed in *GRUE* languages. However, other work suggests that a causal link between the tint of the ocular lens and the naming of colors is at least partly modified by long-term chromatic adaptation [16, 21]. Brown and Lindsey [9] performed a geographical analysis of data on 118 ethnolinguistic groups for which both red-green color deficiency data (protan and deutan defects, not related to blue) and scholarly or dictionary data were available, also from published sources. The geographical results of Brown and Lindsey generally agreed with the results of Bornstein.

Color Naming Worldwide

There is still no single well-accepted explanation for the differences between languages in the use of *BLUE*, *GRUE*, and *BLACK*. Does *BLUE* vary across languages because of physiological differences among people, perhaps due to their different exposure to the sun, as Bornstein and Lindsey and Brown (and Gladstone and Geiger before them) suggested? If so, there might be a correlation between *BLUE* and the physical geography of the localities where these languages are spoken. Is the variation in *BLUE* due to the superior economic and cultural development of advanced nations, as Kay and his colleagues (and Allen and Magnus before them) suggested? If so, there might be a correlation between *BLUE* and the societal characteristics of the cultures where these languages are spoken. Or, is it a historical linguistic phenomenon, with the geographical patterns in the Old World being caused by the predominance of Indo-European languages in Europe? If so, then the non-Indo-European languages spoken in Europe should not show the predominance of BLUE observed in the Indo-European languages.

Figure 3, panels a, b, shows two physical geography characteristics of the individual cultures from Fig. 2, latitude and the annual dose of UVB. If the physiological hypothesis is correct, both of these graphs should correlate (with positive slope on the graph) with the use of *BLUE* (*BLACK* vs. *GRUE* vs. *BLUE*). Only latitude correlates with the use of *BLUE* in both the Old World and the New World. The idea that latitude has its effect through its effect on the annual dosage of UVB from sunlight has the difficulty that UVB dosage is correlated with *BLUE* in the Old World, but not the New World. The exposure of an individual person to UVB will be modulated by the amount of time he/she spends outdoors.

Three societal characteristics are shown in Fig. 3, panels c, d. These are Marsh's "Index of Societal Differentiation," which is a measure of the societal development of an individual culture [19]; life expectancy, which is a measure of the human development of nations; and Berry's "Technological Scale" of the economic development of nations [22]. Marsh's Index and Berry's Scale correlate with *BLUE* in the Old World but not the New World, whereas life expectancy was correlated with *BLUE* in the New World but not the Old World. These three societal effects are imperfect indicators of development. Marsh's Index is only available for the individual cultures where 166 of the languages in Fig. 2 are spoken. It is based on the per capita annual energy consumption and the fraction of males engaged in agriculture. Therefore, it will be correlated with latitude (it takes more energy to keep warm in Alaska than in Cameroon, and the short growing season makes agriculture in Siberia an unprofitable occupation). Life expectancy and Berry's Scale are available only for the nations within which the cultures are embedded. Life expectancy may show a ceiling effect in the Old World. Berry's Scale was the first principal

DOI 10.1007/978-3-642-27851-8_54-1

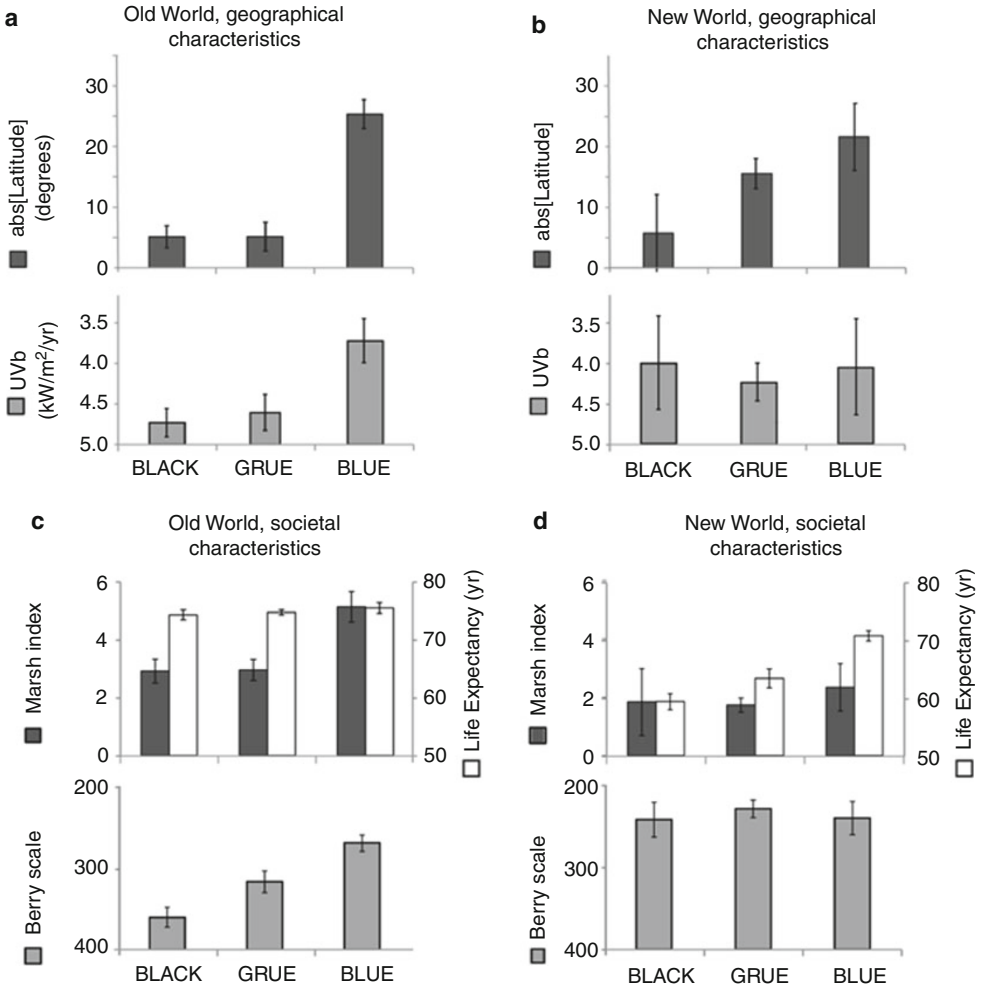

Fig. 3 Characteristics of the cultures or nations where the languages from Fig. 2 are spoken. Each bar indicates the value of the characteristic, averaged (±1 s.e.m) across cultures or nations having *BLACK*, *GRUE*, or *BLUE* as the word for blue in their languages' color lexicons. (**a**, **b**) Geographical characteristics of the localities where the 384 languages are spoken. Dark bars, absolute value of latitude, in degrees from the equator; light gray bars, annual dose of UVB. (**c**, **d**) Societal characteristics. Dark bars, Marsh's "Index of Societal Differentiation," for 166 of the cultures shown in Fig. 2 [17, 19]. White bars, life expectancy; light gray bars, Berry's "Technological Scale" of economic development [22]. Life expectancy and Barry's Scale are shown for the nations where the languages are spoken

component of an overall assessment of the economic advancement of nations, with contributions from transportation and trade, energy production and consumption, national product, communications, and urbanization. All of these societal characteristic correlates have the difficulty that none of them correlate with BLUE in both the Old World and the New World.

Languages that include BLUE greatly predominate Europe, the Mediterranean, and the Near East (Fig. 2, top panel). However, this is not entirely due to the predominance of Indo-European languages in that area. The data sets in Fig. 2 include 46 Old World languages that are spoken north of the Tropic of Cancer and west of the Ural Mountains. Of the 28 Indo-European languages and 18 non-Indo-European languages in this group, all but two use *BLUE*; one Indo-European language (Gaelic) uses *GRUE*, and one

DOI 10.1007/978-3-642-27851-8_54-1

non-Indo-European language (Nenets) uses *BLACK*. Whatever is responsible for the predominance of *BLUE* in this geographical region, it is not entirely a question of linguistic heritage.

In spite of over 150 years of research, involving empirical corpora of color-naming data, scholarly data sets of color lexicons, and philological analyses of ancient texts, it is not well understood why there is such great worldwide variation in the terms in the color lexicons of world languages. This topic continues to be the subject of much contemporary research.

Cross-References

The following links should refer to other articles in this encyclopedia:

- ▶ Berlin and Kay Theory
- ▶ Evolutionary Sequence
- ▶ GRUE
- ▶ Stages
- ▶ World Color Survey

References

1. Gladstone, W.E.: Studies on homer and the Homeric age, vol. III. Oxford University Press, London (1858)
2. Geiger, L.: Contributions to the history of the development of the human race. Trubner & Col, London (1880)
3. Allen, G.: The colour-sense: its origin and development: an essay in comparative psychology. Houghton, Boston (1879)
4. Saunders, B., Marth, I.-T.: The debate about colour-naming in 19th century German philology. Leuven University Press, Leuven (2007)
5. Rivers, W.H.R.: Vision. In: Haddon, A.C. (ed.) Reports on the Cambridge anthropological expedition of the Torres straits, pp. 1–132. Cambridge University Press, Cambridge (1901)
6. Berlin, B., Kay, P.: Basic color terms: their universality and evolution. University of California Press, Berkeley/Los Angeles (1969)
7. Kay, P., et al.: The world color survey. CSLI, Stanford (2009)
8. Bornstein, M.H.: Color vision and color naming: a psychophysiological hypothesis of cultural difference. Psychol. Bull. **80**(4), 257–285 (1973)
9. Brown, A.M., Lindsey, D.T.: Color and language: worldwide distribution of Daltonism and distinct words for “blue”. Vis. Neurosci. **21**, 409–412 (2004)
10. Hickerson, N.R.: Basic color terms: their universality and evolution by Brent Berlin; Paul Kay. Int. J. Am. Linguist. **37**(4), 257–270 (1971)
11. Kay, P., Regier, T.: Resolving the question of color naming universals. Proc. Natl. Acad. Sci. U. S. A. **100**, 9085–9089 (2003)
12. Regier, T., Kay, P., Khetarpal, N.: Color naming reflects optimal partitions of color space. Proc. Natl. Acad. Sci. U. S. A. **204**, 1436–1441 (2007)
13. Lindsey, D.T., Brown, A.M.: Universality of color names. Proc. Natl. Acad. Sci. U. S. A. **103**, 16608–16613 (2006)
14. Lindsey, D.T., Brown, A.M.: World color survey color naming reveals universal motifs and their within-language diversity. Proc. Natl. Acad. Sci. U. S. A. **206**, 19785–19790 (2009)

DOI 10.1007/978-3-642-27851-8_54-1

15. Webster, M.A., Kay, P.: Variations in color naming within and across populations. Behav. Brain Sci. **28**(4), 512 (2005)
16. Hardy, J.L., et al.: Color naming, lens aging, and Grue: what the optics of the aging eye can teach us about color language. Psychol. Sci. **16**, 321–327 (2005)
17. Hays, D.G., Margolis, E., Naroll, R., Perkins, D.R.: Color term salience. Am. Anthropol. **74**, 1107–1121 (1972)
18. Ember, M.: Size of color lexicon: interaction of cultural and biological factors. Am. Anthropol. **80**(2), 364–367 (1978)
19. Marsh, R.M.: Comparative sociology: a codification of cross-societal analysis. Harcourt, Brace &World, New York (1967)
20. Lindsey, D.T., Brown, A.M.: Color naming and the phototoxic effects of sunlight on the eye. Psychol. Sci. **13**(6), 506–512 (2002)
21. Delahunt, P.B., et al.: Long-term renormalization of chromatic mechanisms following cataract surgery. Vis. Neurosci. **21**, 301–307 (2004)
22. Berry, B.J.L.: An inductive approach to the regionalization of economic development. In: Ginsberg, N. (ed.) Essays on geography and economic development, pp. 78–107. University of Chicago, Chicago (1960)
23. Schöntag, R., Schäfer-Prieß, B.: Color term research of Hugo Magnus. In: MacLaury, R.E., Paramei, G.V., Don Dedrick, D., (eds) Anthropology of Color: Interdisciplinary Multilevel Modeling, pp. 107–122. John Benjamins, Amsterdam (2007)