*ICIC Express Letters*
Volume **11**, Number **2**, February **2017**

ICIC International ©2017 ISSN 1881-803X
pp. 343–347

# TWITTER-BASED COLOR TERM ANALYSIS IN ENGLISH, GERMAN, AND JAPANESE

Daniel Moritz Marutschke[1], Yancong Su[2,*] and Hitoshi Ogawa[1]

[1]College of Information Science and Engineering
Ritsumeikan University
1-1-1 Nojihigashi, Kusatsu, Shiga 525-8577, Japan
moritz@fc.ritsumei.ac.jp; ogawa@is.ritsumei.ac.jp

[2]School of Digital Arts and Communication
Xiamen University of Technology
No. 600, Ligong Road, Xiamen 361024, P. R. China
*Corresponding author: syc@xmut.edu.cn

Received July 2016; accepted October 2016

Abstract. *Color continues to be an important topic and the cultural identification plays a significant role in society. This research focuses on combining known facts about cultural responses to colors by data-mining social media. To differentiate the use of 11 basic color terms in English, Japanese, and German Twitter feeds, word clusters and co-occurrences are analyzed.*

**Keywords:** Color analysis, Cultural analysis, Twitter

1. **Introduction.** Colors and their perception and use in different cultures have been researched, mostly by affective engineering and social experiments with typical subsets of society [1, 2, 3, 4]. Other researches which analyze the cultural effect on the usage of colors are done manually on limited number of data points [5, 6]. The color red is often the focus of this type of research as it is one of the most distinct colors and has clear cultural differentiation as well as meanings shared in most cultures [7, 8]. A set of 11 colors (Table 1) can be considered a *basic color set* that is most common in numerous cultures.

Usual approaches in analyzing cultural differences related to color are limited due to the amount of data that can be processed manually. Commonly the data sets consist of several hundred data points and one-digit number of variables on a two-dimensional data set. This could be the manual viewing of letter-coloring on birthday cards or New Years cards. The effort in examining each card individually sets obvious constraints. Another drawback is the exponential workload when considering multiple colors in various cultures. For data samples such as acquired by social media, the task would be impossible. An automated approach opens capabilities of many order of magnitudes larger. Subjective biases can also be avoided. Especially when considering textual data, human error would inevitably be introduced. Some drawbacks of introducing Information and Communication Technologies (ICT) to this area are less powerful semantic analytics. Another drawback is the lack of complex pattern recognition – or a combination of both.

Other research uses questionnaires or experiments with sample sizes of several hundred participants. Results can be skewed by the type of questionnaire, perception of the questions, the targeted subject demography, etc. One recent study researches the gender based preference on Twitter [9].

This research presents a data mining procedure to find word-association and co-occurrence to examine the differences in color-term usage in different languages. Differences are established in many instances by small-scale experiments in psychology, but not in the larger scale in multiple languages and the use of ICT and social media. By analyzing 1,320,000 individual tweets, 40,000 tweets of each 11 basic color keywords in English, Japanese, and German, this paper provides results to support an ICT approach to these questions. As Twitter is one of the most popular social networks world wide and the limited number of characters per message, this medium was found to be useful to assess this methodology.

The paper is structured as follows: Section 2 introduces the methods used in getting the clustering results and assessment with subsections about data acquisition and details about the methodologies for analysis; Section 3 discusses the results and implications for future research, difficulties in natural language processing and cultural influences; Section 4 concludes the paper.

## 2. Methods.

The methodology used in this research consists of four steps:

1) Collection and pre-processing of keyword related tweets
2) Natural language processing
3) Clusters of co-occurring phrases with keywords
4) Assessing clusters using *t*-test

Basic colors exist differently in cultures [7, 10, 11, 12]. These differences may be based linguistically (in Japanese the word for green and blue) or from their cultural usage (e.g., religious meaning). The data set of this research consists of 11 basic colors used in German, Japanese, and English (see Table 1).

Table 1. German, Japanese, and English basic color terms

| German | Japanese | English |
|---|---|---|
| schwarz | kuro | black |
| weiss | shiro | white |
| rot | aka | red |
| gruen | midori | green |
| blau | ao | blue |
| grau | haiiro | grey |
| braun | chairo | braun |
| orange | orenji | orange |
| gelb | kiiro | yellow |
| violett | murasaki | purple |
| pink | pinku | pink |

### 2.1. Data acquisition.

Twitter was used as data source because of its widespread international presence and amount of textual data produced each day. Due to the popularity of the microblogging platform in German, Japanese, and English speaking countries, fine-grained and targeted data mining is possible.

A `Python` algorithm was written to access the `Twitter API` (Application Programming Interface), filter the tweets, pre-process the data, and store the results to a text file. API access was done via the `TwitterSearch` package to facilitate filters such as language, region, exclusion of retweets, and necessary keywords.

Tweets were filtered by language rather than by GPS (Global Positioning System) data. GPS data is not always enabled on the user account. Moreover, in cultural terms, language is assumed to be more prevalent than location. Retweets, a simple forwarding of messages, were removed to reduce duplicates.

In pre-processing, URLs and usernames (beginning with an @-symbol) were stripped from the tweets to facilitate the data mining.

Each language was mined using natural language processing (Python implementation of `nltk` for English and German, Python implementation of `janome` for Japanese) to obtain a new data set with only relevant words in their basic form. This pre-processing step minimizes the effects of semantic misinterpretation of words like "orange", which could mean the fruit or the color. The authors acknowledge the introduction of other effects specific to the natural language processing tool used, as they behave differently in each language.

For each keyword (basic color term), a timeframe was set to target 40,000 unique tweets containing that keyword. The data-set dimension is 3 by 11 by 40,000 (language, colors, tweets), resulting in a total of 1,320,000 individual tweets.

2.2. **Word-association and co-occurrence.** As computation has become increasingly powerful, especially textual data-mining has experienced improvements in recent decades. Linguistic and lexical analyses such as concordance were prohibitive in pre-computer era and reserved for major religious or historic works. Nowadays, any standard desktop computer or even laptop has the capability to perform text-mining on order or magnitudes larger than ever before.

Using the software tool `CasualConc` (v.1.9.8), a KWIC (Keyword in Context) concordance was performed to manually search for most used phrases connected to a color keyword. Some of these phrases were expected, such as the keyword *red* in combination with *roses* or *dress*.

The clusters relevant for this analysis were formed by word co-occurrences by word frequency based on each color keyword using `CasualConc`.

After the initial examination, the algorithm was integrated in the Python script to provide a more accessible data file for further analysis.

The algorithm was then implemented in Mathematica to evaluate the output.

2.3. **Cluster assessment.** The number of clusters selected depended on each color keyword. A joint set of all co-occurring words was generated to compare the results for German, Japanese, and English sets. All the colors showed distinct clusters for German, Japanese, and English words with $p < 0.05$.

Table 2. Word co-occurrence related to the keyword *red* in German, Japanese, and English

| German | Japanese | English | Meaning |
|---|---|---|---|
| haare | kami | hair | |
| – | pazudora | – | Puzzle & Dragons (game) |
| kreuz | – | cross | |
| – | – | video | |
| mond | – | moon | |
| rosen | bara | roses | |
| kleid | doresu | dress | |
| ampel | shingo | light | |
| zahlen | akaji | – | in the red (financial) |
| – | akaten | – | failed test (expression) |
| teppich | – | carpet | |
| lippen | – | – | |
| Wein | wain | wine | |
| faden | – | thread | |
| – | tesuto | – | exam |

An exhaustive list of all the clusters would be beyond the scope of a publication, and an example is given from the most frequent keywords in the color red as in Table 2. This table lists the first 15 co-occurring phrases related with the *red* in German, Japanese, and English.

The example for the keyword *red* was selected as an example intentionally as red is the most prevalent in everyday life and can be identified easily. Lists of the other ten colors would again be beyond the scope of this paper.

## 3. Discussion.

Research in cultural differences based on color schemes is an established topic in liberal arts. With the help of ICT, especially trend related social networking services such as Twitter, the use of colors in different cultures can be assessed in more depth.

The results for the color red show known associations with words like *rose*, *dress*, or *lips*. Unique results were *(red) lips* in German, *video* in English, and *exam* or *akaten* in Japanese.

As colors play an important role in product design and marketing [13, 14, 15], viability for system implementation is deemed promising.

## 4. Conclusions.

Differences in the use of colors in context depending on three languages – German, Japanese, and English – are explored. Clustering words, which co-occur with one of the basic 11 colors, show reliable differences in cultural classification.

The use of Twitter allows to analyze large amounts of data points and a view on co-occurrence clusters that are not possible with methodologies used in most color and culture related research.

After a deeper understanding of the use of colors in social media, the results could help avoid cultural dissonance caused by colors. These findings are also considered to improve timely reaction on cultural trends, e.g., for product design.

Future research for culturally different use of colors is set for four main goals – in-depth natural language processing and evaluation to address meaning of expressions; more languages; colors co-occurring with other colors; colors co-occurring with emojis.

**Acknowledgement.** This research was partly supported by the Social Sciences Planning Projects of Fujian Province (No. FJ2015C075) and the High-level Talents Projects of Xiamen University of Technology (No. YKJ14023R).

## REFERENCES

[1] K. Uchikawa and R. M. Boynton, Categorical color perception of Japanese observers: Comparison with that of Americans, *Vision Research*, vol.27, no.10, pp.1825-1833, 1987.

[2] E. R. Heider and D. C. Olivier, The structure of the color space in naming and memory for two languages, *Cognitive Psychology*, vol.3, no.2, pp.337-354, 1972.

[3] A. Majid and S. C. Levinson, The senses in language and culture, *The Senses and Society*, vol.6, no.1, pp.5-18, 2011.

[4] O. Song, W.-H. Lee and J.-Y. Kim, A study on color features based on color classification of the Korean royal costumes during Choseon Era, *Art, Culture, Game, Graphics, Broadcasting and Digital Contents*, pp.62-67, 2015.

[5] M. Saito, Comparative studies on color preference in Japan and other Asian regions, with special emphasis on the preference for white, *Color Research & Application*, 1996.

[6] D. Roberson, I. Davies and J. Davidoff, Color categories are not universal: Replications and new evidence from a stone-age culture, *Journal of Experimental Psychology: General*, vol.129, no.3, pp.369-398, 2000.

[7] K. A. Jameson, Culture and cognition: What is universal about the representation of color experience? *Journal of Cognition and Culture*, 2005.

[8] D. Roberson, Color categories are culturally diverse in cognition as well as in language, *Cross-Cultural Research*, vol.39, no.1, pp.56-71, 2005.

[9] S. Fortmann-Roe, Effects of hue, saturation, and brightness on color preference in social networks: Gender-based color preference on the social networking site Twitter, *Color Research & Application*, vol.38, no.3, pp.196-202, 2011.

[10] J. Winawer, N. Witthoft, M. C. Frank and L. Wu, Russian blues reveal effects of language on color discrimination, *Proc. of the National Academy of Sciences of the United States of America*, pp.7780-7785, 2007.

[11] R. Hanley and D. Roberson, Color vision: Color categories vary with language after all, *Current Biology*, vol.17, no.15, pp.R603-R605, 2007.

[12] D. Roberson, J. Davidoff, I. R. L. Davies and L. R. Shapiro, Color categories: Evidence for the cultural relativity hypothesis, *Cognitive Psychology*, vol.50, no.4, pp.378-411, 2005.

[13] T. J. Madden, K. Hewett and M. S. Roth, Managing images in different cultures: A cross-national study of color meanings and preferences, *Journal of International Marketing*, vol.8, no.4, pp.90-107, 2000.

[14] M. M. Aslam, Are you selling the right colour? A cross-cultural review of colour as a marketing cue, *Journal of Marketing Communications*, vol.12, no.1, pp.15-30, 2006.

[15] S. Singh, Impact of color on marketing, *Management Decision*, vol.44, no.6, pp.783-789, 2006.