The Use of English Colour Terms in Big Data

Dimitris MYLONAS,[1,3] Matthew PURVER,[1]
Mehrnoosh SADRZADEH,[1] Lindsay MACDONALD,[2] Lewis GRIFFIN,[3]
[1] School of Electronic Engineering and Computer Science, Queen Mary Univ. of London
[2] Dept. of Civil, Environmental & Geomatic Engineering, University College London
[3] Dept. of Computer Science, University College London

ABSTRACT

This study explores the use of English colour names in large datasets from informal Twitter messages and the well-structured corpus of Google Books. Because colour names in text have no directly associated chromatic stimuli, the corresponding colour categories of colour words was assessed from responses in an online colour naming experiment. A comparison of the frequency in the three datasets revealed that the mapping of colour names to perceptually uniform colour spaces does not reflect natural language colour distributions.

1. INTRODUCTION

Colour plays a central role in visual perception and can be a powerful tool to differentiate emotions, ideas and identities. We are able to see millions of different colours but we tend to organise them into a smaller set of colour categories and give them names such as yellow, peach or sky blue. There is a growing interest in the language of colour, and over recent years colour naming models have been used for gamut mapping (Motomura, 1997), image processing (Moroney *et al.,* 2008) and colour selection (Heer & Stone, 2012).

In this paper we explore natural language processing and data visualisation methods for understanding the use of colour names by analysing a large pool data from Twitter and Google Books. Because colours in language have no direct reference to chromatic stimuli, we accessed the associated colour categories of each colour name from the responses of hundreds of participants in an online colour naming experiment (Mylonas & MacDonald, 2010).

Twitter is an open micro-blogging platform that allows millions of users around the world to broadcast and receive in real time short messages, known as *tweets*, of up to 140 characters long. Twitter's conversations are public by default and organised by community-driven practices. This provides researchers with the opportunity to analyse multilingual everyday conversations outside of formal institutional environments.

In 2001, Google created a large corpus of n-grams based on ~4% of all books ever published. N-grams refer to the sequence of n words found in all digitized books. The first edition of the corpus consisted of over 500 billion words published between 1500 and 2000 in English, French, Spanish, German, Chinese, Russian and Hebrew (Michel *et al.* 2011). A new edition of the corpus provides syntactically annotated n-grams and their counts with part-of-speech (POS) tags from over 6% of all books ever published in 8 languages (Lin *et al.* 2012). POS taggers classify words as nouns, verbs and adjectives, etc. and provide an instructive form of word-category disambiguation in a given context.

An online colour naming experiment (Available at: http://colournaming.com) was designed to collect broad sets of multilingual colour names with their corresponding colour ranges in sRGB and Munsell specifications. Over the past seven years (2008-2015) the server has gathered responses from thousands of participants in fourteen languages: English, Greek, Spanish, German, Catalan, Italian, simplified and traditional Chinese, French, Korean, Danish, Lithuanian, Thai and Portuguese. The server also gathered the response time for each colour name and associated metadata regarding the cultural background, colour deficiency, hardware/software components and viewing conditions of the participants (Mylonas & MacDonald, 2010).

The Munsell system is the most widely used apparatus in colour naming research, despite its limitations, as it provides a pragmatic colour space to map colour names to perceptual colour coordinates. The system divides the colour space evenly into five primary hues (yellow, red, blue, purple and green) and five intermediate hues. Purple was included as a primary because there are about twice as many perceptible hue steps between blue and red as between red and yellow, or yellow and green, or green and blue (although in nature we might find relatively fewer purple colours). A renotation was carried out (Newhall *et al.,* 1943) with the objective to represent perceptually uniform hue, saturation and lightness spacing based on the principle of the Just Noticeable Difference (JND). The Munsell colours do not represent typical naturally occurring colours as their pigment spectra are smoothed in comparison to naturally occurring spectra and while it covers all the most important regions of colour space, some areas are not well represented (Buchsbaum & Bloch, 2002).

Considering that colour coding may reflect the colours available in our environment, previous studies have focused on uniform colour spaces and image statistics of natural scenes (McDermott & Webster, 2012). In the present study we asked instead whether colour language is efficiently represented in perceptual colour spaces and examined whether the statistics of colour in written language follow the distribution of colour names mapped to an approximately uniform colour space in an online colour naming experiment.

2. METHOD

We analysed 10,000 responses in the online colour naming experiment from 500 UK-resident English speakers, of which 90.3% reported normal colour vision. Colour name responses most often were of a single word (monolexemic) but could consist of an unlimited number of words. We identified the most frequent 50 monolexemic colour terms responded 20 times or more in responses from non-deficient observers over the age of 16. To access the associated colour categories of each colour name we retrieved all colour samples given the same name.

To explore the usage of colour names in informal, online conversations, we took the 50 most frequent monolexemic colour terms from experiment responses, and measured their probability in 1,036,103 random tweets from the Twitter API. We filtered Twitter's public stream with the geo-location coordinates of [-5.4,50.1,1.7,55.8] that correspond to a rectangle with edges approximately at the edges of Britain. We excluded tweets in other languages than English {'lang':'en'}. Each tweet was tokenised into unigrams using the Natural Language Toolkit (Bird *et al.*, 2009) and typographical conventions were removed resulting in 129,355,280 tokens.

Messages in Twitter are limited to max 140 characters and often consist of non-standard English that makes the task of word-category disambiguation challenging. For example, it is difficult to determine whether the word *orange* is being used metonymically as an adjective to describe the colour of an object, like an orange table, or is being used literally to describe a type of citrus fruit. To investigate the use of colour names in context and disambiguate their syntactic role, we also counted the probability of the 50 most frequent monolexemic colour terms from experiment responses in the syntactically annotated Google Books corpus of all digitised English books between 1500-2000. The frequency of occurrence of each unigram was counted by dividing the number of instances of each unigram in the given years by the total number of tokens (n=468,491,999,592) in the corpus for the same years (Lin *et al.,* 2012).

2.1 Sample Preparation

The 600 total test samples in the colour naming experiment were specified in the sRGB colour space and selected from the Munsell Renotation Data (Newhall *et al.,* 1943). The original dataset consisted of 2729 colour samples specified in CIE *xyY* colour space and viewed against a neutral grey background under illuminant C. Since achromatic colours were not included, nine neutral samples, one for each Munsell Value and a White and a Black sample at the extremes of the sRGB cube were added. Colour samples lying outside the sRGB gamut were discarded. Given the cylindrical coordinate system, the sub-sampling of the remaining in-gamut colour samples followed a similar approach to the advice of Billmeyer to Sturges & Whitfield (1995), namely to equalize the perceptual distances between samples (Mylonas & MacDonald, 2010).

2.2 Experimental Procedure

The procedure in the online colour naming experiment consists of six steps (Figure 1). First, we ask the observers to adjust his or her display to sRGB settings, and the brightness in order to make visible all twenty-one steps of a grey scale ramp. In the second step the participant answers questions relating to the lighting conditions, the environment and properties of the display. Then, in the third step, the participant is screened for possible colour deficiencies with a simple web-based Dynamic Colour Vision Test developed at the City University London (Barbur 2004). The fourth and main part is the *unconstrained colour-naming task*: any colour descriptor, either a single word, or a compound, or terms(s) with modifiers can be entered to describe each of twenty samples presented in sequence and randomly selected from the 600 in total samples. Along with the colour name typed on a keyboard, the response times (RTs) of onset of typing are recorded, defined as the interval between presentation of the colour stimulus and the first keystroke. In the fifth step we collect information about the participant's residency, nationality, language proficiency, educational level, age, gender and colour experience. In the last step we provide the participant with a summary of the responses.

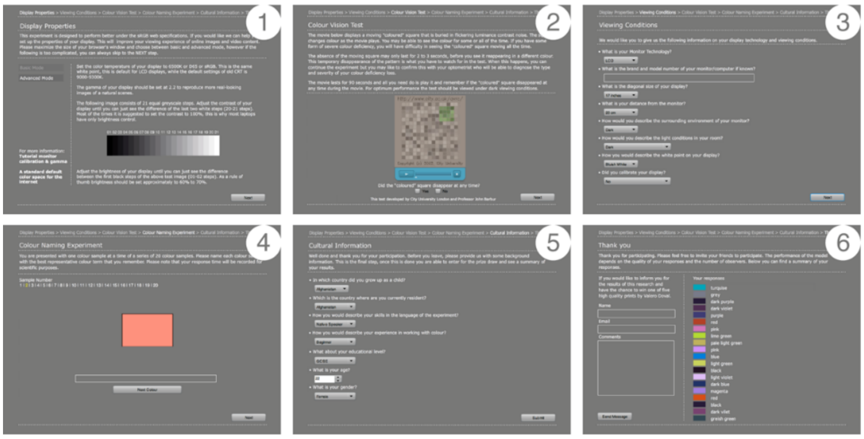

Figure 1: Flowchart of the experimental procedure (Available at: http://colornaming.com)

3. RESULTS AND DISCUSSION

The probability of the most frequent monolexemic colour terms from experiment responses in Twitter messages is shown in Figure 2. For clarity, we have chosen a cut-off of 30 most frequent terms in Twitter given that non-expert observers are able to identify 30 colour names in their native language without training (Derefeldt & Swartling, 1995). *Black* and *white* were the most frequent colour terms followed by *red*, *cream* and *blue*. *Yellow* was found in the 9th position while *indigo* and *teal* were ranked at the bottom of the list. The absence of context in this approach produced an issue of word-category disambiguation. For example, we were not able to disambiguate whether *cream*, *orange* and *salmon* were used as nouns or as adjectives. In Twitter messages this is a particularly challenging problem as the character limit and conventions of text communication forces users to compress more information into fewer characters without conventional use of grammar and syntax. For well-structured corpora such as books and articles, POS taggers achieve higher accuracy.

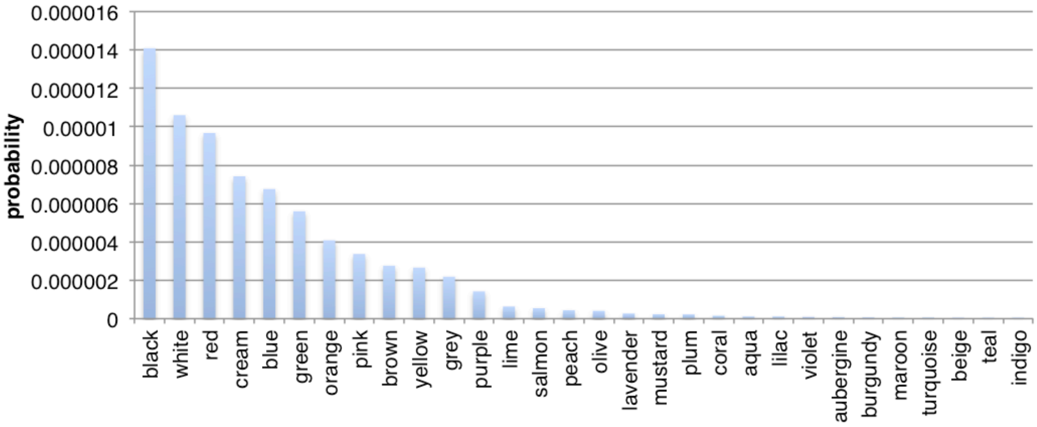

Figure 2. Top 30 most frequent colour terms in ~1 million Twitter conversations

Figure 3 shows the probability of the 30 most frequent English colour terms used as adjectives in the syntactically annotated Google Ngrams Corpus from the 50 most frequent monolexemic colour terms from experiment responses. *White, brown* and *red* were the most frequent colour terms followed by *blue*, *black, green* and *yellow*. The least frequent terms were *khaki*, *turquoise* and *maroon*.

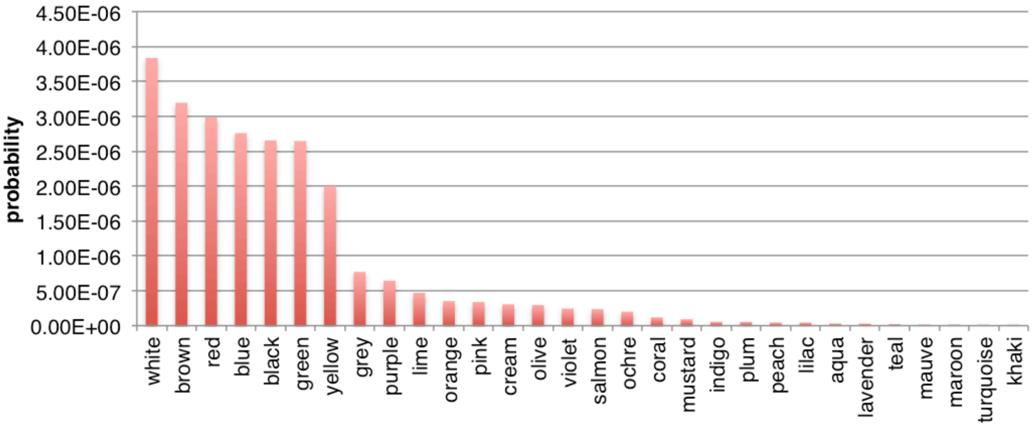

Figure 3. Top 30 most frequent English colour terms used as adjectives in Google books Ngrams between 1500-2000.

We retrieved the 30 most frequent colour terms with the highest average rank across Twitter and Google Books (*white, black, red, blue, brown, green, cream, yellow, orange, grey, pink, purple, lime, olive, salmon, mustard, peach, coral, violet, plum, lavender, lilac, aqua, indigo, maroon, teal, turquoise, burgundy, aubergine and beige*) and obtained all colour samples given the same name by hundreds of participants in the online colour naming experiment.

Figure 4 shows these colour categories by the size and their associated colour names. *Purple* was the largest colour category in the experiment followed by *blue* and *pink*. *Lilac* and *turquoise* were found in the 6th and 7th positions respectively. The colour categories with the smallest size were *coral, cream* and *lime*.

Comparing the distributions of colour names in the three datasets shows that while *white* was the most frequent colour term in Google Books and second in Twitter, in the online experiment it was found in the 26th position. *Purple* on the other hand was the largest category in the online experiment while in Twitter was the 12th and in Google Books the 9th most popular colour term. *Red* was found in the 3rd position in both Twitter and Google Books but in the experimental dataset was found in the 12th position.

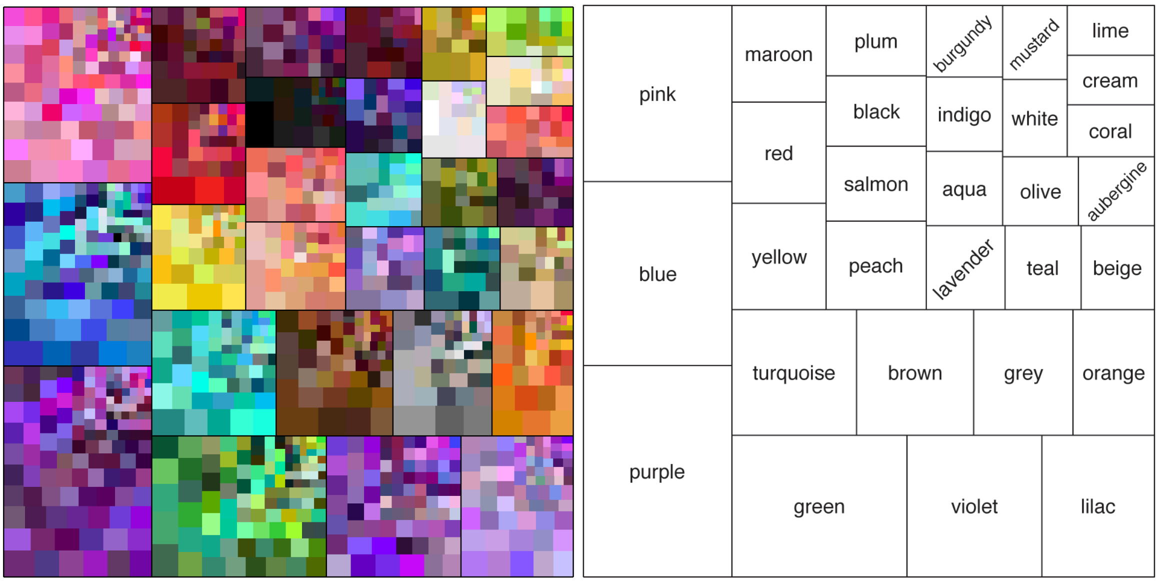

Figure 4. Treemap associated with the size of colour categories in the online colour naming experiment: (left) colour samples of each colour name; (right) colour name.

4. CONCLUSIONS

In this study we have presented the use of colour names in Twitter messages and Google Books and we visualised their associated colour categories using responses from an online colour naming experiment. The comparison of the colour distributions in the three datasets revealed that the mapping of colour names to perceptually uniform colour coordinates does not reflect natural language colour distributions. Future plans include the examination of the geometry of lexical colour spaces.

ACKNOWLEDGEMENTS

We thank all the volunteers who have contributed in this study as participants in the online experiment. This work was supported in part by the QMUL Media and Arts Technology programme, EPSRC Doctoral Training Centre and by the UCL Computer Science, EPSRC Doctoral Training Grant.

REFERENCES

Buchsbaum, G. and O. Bloch 2002. Color categories revealed by non-negative matrix factorization of Munsell color spectra. *Vision Research*, 42(5), 559-563.

Barbur J.L., A.J. Harlow, G.T. Plant 1994. Insights into the different exploits of colour in the visual Cortex. *Proc Royal Society London B Biology Sci*. 258: 327-334.

Bird, S., E. Klein & E. Loper 2009. *Natural Language Processing with Python*. Beijing; Cambridge Mass.: O'Reilly Media.

Derefeldt, G., & T. Swartling 1995. Colour concept retrieval by free colour naming. Identification of up to 30 colours without training. *Displays*, 16(2), 69–77.

Heer, J. and M. Stone 2012. Color naming models for color selection, image editing and palette design. *In Proc. CHI 2012*, Austin, TX, USA, ACM Press: 1007-1016

Lin, Y., J.-B. Michel, E.L. Aiden, J. Orwant, W. Brockman and S. Petrov 2012. Syntactic Annotations for the Google Books Ngram Corpus. *In Proc. ACL 2012 System Demonstrations*. Stroudsburg: Assoc. Comp. Linguistics.

Michel, J.-B., Y.K. Shen, A.P. Aiden, A. Veres,… E.L.Aiden 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331 (6014), 176-182.

McDermott, K. C., & M.A. Webster 2012. Uniform color spaces and natural image statistics. *Journal of the Optical Society of America A*, 29(2), A182–A187.

Moroney, N., P. Obrador and G. Beretta 2008. Lexical Image Processing. *In Proc. 16th Color Imaging Conf.*, Portland: IS&T, pp. 268-273

Motomura, H. 1997. Categorical Color Mapping for Gamut Mapping. *In Proc. 5th Color Imaging Conf.,* Scottsdale: IS&T/SID, pp. 50-55.

Mylonas, D. and L. MacDonald 2010. Online Colour Naming Experiment Using Munsell Samples. *In Proc 5th European Conf. on Colour in Graphics, Imaging and Vision (CGIV)*, Joensuu: IS&T, pp. 27-32

Newhall, S.M., D. Nickerson and D.B. Judd 1943. Final Report of the O.S.A. Subcommittee on the Spacing of the Munsell Colors. *Journal of the Optical Society of America A*, 33(7), 385-411.

Sturges, J. and A. Whitfield 1995. Locating basic colours in the Munsell Space. *Color Research and Application*, 20(6): 364-376.

Address: Dimitris MYLONAS, Department of Computer Science,
University College London, Gower Street, London, WC1E 6BT, UK
E-mails: d.mylonas@ucl.ac.uk, m.purver@qmul.ac.uk, mehrnoosh.sadrzadeh@qmul.ac.uk,
lindsay.macdonald@ucl.ac.uk, l.griffin@cs.ucl.ac.uk