Studying the Effect of Globalization on Color Perception using Multilingual Online Recruitment and Large Language Models

Jakob Niedermann[1,*], Ilia Sucholutsky[2,*], Raja Marjieh[3], Elif Celen[1]
Thomas L. Griffiths[2,3], Nori Jacoby[1,], Pol van Rijn [1,**]**
[1]Computational Auditory Perception Group, Max Planck Institute for Empirical Aesthetics
[2]Department of Computer Science, Princeton University
[3]Department of Psychology, Princeton University
{jakob.niedermann, elif.celen, nori.jacoby, pol.van-rijn}@ae.mpg.de
{raja.marjieh, is2961, tomg}@princeton.edu
[*,**] Equal contribution

Abstract

How does globalization impact the interaction between perception and language? Building on Berlin and Kay's foundational study of color naming, we recruited 2,280 online participants speaking 22 different languages. We show that color naming maps differ structurally across languages, even among internet users living in (mostly) industrial societies. We use Large Language Models (LLMs) to simulate the limits of globalization by reproducing the naming task with a highly multilingual artificial agent with access to global digital information. We show that while the LLM has access to all languages, it has language-specific color representations and the number of color terms is correlated across humans and LLMs. However, LLMs use more color terms than humans, indicating differences in the representation. These results suggest that globalization has not removed cultural distinctions in color concepts, as language continues to be a key factor in the diversity of perception and meaning.

Keywords: Culture, Cross-linguistic analysis, Color naming, Large language models, Language and Thought

Introduction

Characterizing the role of culture in shaping cognition is one of the fundamental questions in cognitive science (Barrett, 2020; Henrich, Heine, & Norenzayan, 2010). Culture is influenced by many factors, including language (Blasi, Henrich, Adamou, Kemmerer, & Majid, 2022; Kramsch, 2014), values (Inglehart, Basanez, Diez-Medrano, Halman, & Luijkx, 2000; Triandis, 2018), and economy (Fine, 2016; Henrich et al., 2001). More recently, globalization, as well as the spread of information over the internet, have had an increasing impact on societies around the world (Barrett, 2020; Pieterse, 2019). Here, we study the effect of globalization on cognition using the paradigmatic case of color terms.

Color concepts are a classic domain for studying the interaction between language and perception across cultures (Whorf, 2012). This is exemplified by the seminal works of Berlin and Kay (1969) and the World Color Survey (WCS; Kay and Cook 2016) on characterizing basic color concepts in written and unwritten languages, respectively. Their key findings are that a) speakers of different languages produce different color maps when describing the same set of colors using their respective basic color terms, and b) there is a large overlap between color maps in languages that share the same number of basic color terms. Surprisingly, there is no large-scale, standardized, contemporary color-naming data at a global scale investigating industrialized languages. The largest color naming study, the World Color Survey (Kay & Cook, 2016), had fairly standardized procedures for each language but focused on collecting data for unwritten languages, mostly from small-scale societies. Majid et al. (2018) collected data from 20 languages but did not focus on industrialized (or written) languages. They used a different set of stimuli than the WCS, and participants were allowed to use broader vocabulary (not only color terms). Many other contemporary studies that did collect some data from written languages are either not publicly available or cover only a few languages, each with different methodologies (for example Lindsey & Brown, 2014; Al-rasheed, 2014; Davies & Corbett, 1994; Kuriki et al., 2017; Ozgen & Davies, 1998; Thierry, Athanasopoulos, Wiggett, Dering, & Kuipers, 2009; Zollinger, 1984; Winawer et al., 2007; Berlin & Kay, 1969). Furthermore, even if color-naming data were available for written languages, it is still difficult to define a suitable control class that could reflect high levels of globalization and multilingualism. As a result, the effect of globalization on color naming has yet to be tested at scale.

To address this, we leverage Lucid[1], a non-traditional diverse recruiter, to collect color-naming data for 22 languages with a unified methodology. The Lucid recruiter provides access to significantly larger participant pools from a diverse range of countries and languages than most traditional recruitment platforms such as Amazon Mechanical Turk or Prolific. We also use multilingual state-of-the-art Large Language Models (LLMs, here GPT-4; Achiam et al., 2023) and Vision Language Models (VLMs, here GPT-4V) that have been trained on virtually all digitally available text in many languages (as well as a large number of images in the case of the VLMs) as a control class to simulate the ultimately globalized agent (highly multilingual, heavily exposed to the internet, with access to digital content from around the world).

We find that even among internet users, we still see structural variation in the number of basic color terms and maps across languages. Even when using LLMs and VLMs to simulate the limits of globalization, we see analogous differences when the agents are queried in different languages. Taken together, our findings demonstrate how online experiments can be coupled with recent advances in machine learning to better understand classical questions in cogni-

[1]A service provided by CINT https://www.cint.com/

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

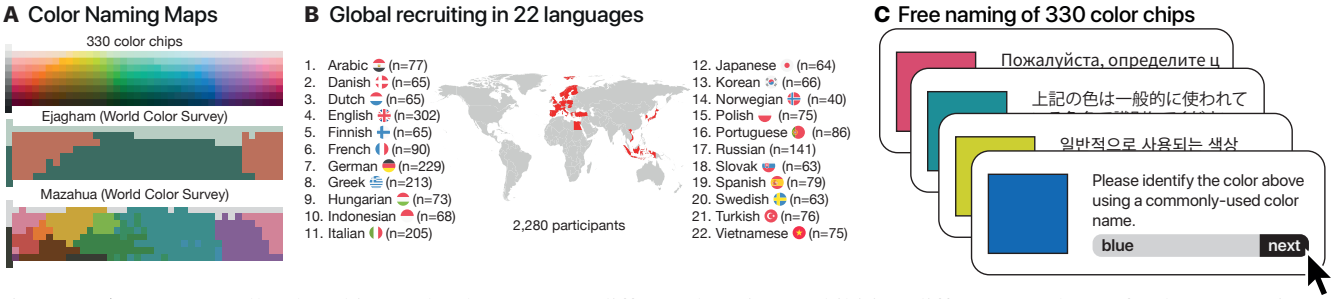

Figure 1: **A** 330 Munsell color chips and color maps on different locations exhibiting different numbers of color categories. **B** Worldwide recruitment in 22 different languages. **C** Participants were asked to name the color chip using a basic color term.

tive science. We provide an interactive visualization of our data here: `https://global-colors.s3-eu-central-1.amazonaws.com/index.html`.

Background

The World Color Survey

Berlin and Kay (1969) collected color naming data on 320 chromatic and nine non-chromatic Munsell chips (Lenneberg & Roberts, 1956; see Figure 1A) from speakers of 20 languages in the Bay Area. The majority of these speakers were bilingual. In a follow-up project, the "World Color Survey" (WCS), linguists and anthropologists from around the world studied color naming patterns in 110 unwritten languages, many of them in small societies. Here, participants provided color-naming responses for a set of 330 Munsell chips. This study concluded that having "basic colors" – a small set of commonly used color terms – is universal. The systems of color terms themselves, however, varied greatly across languages. Nevertheless, they showed that languages with similar numbers of basic color terms tend to have similar mappings of terms to colors. Based on this, Berlin and Kay suggested that systems of color terms develop from the simplest two color-term systems to more complex color systems with up to eleven color terms (Berlin & Kay, 1969; Kay & Cook, 2016).

Subsequent theoretical work provided an alternative interpretation of the WCS findings. According to Regier, Kay, and Khetarpal (2007), the results of the WCS can be explained from two principles: a) a universally shared visual system, and b) an "efficient" organization of the lexicon. This idea was further developed by Zaslavsky, Kemp, Regier, and Tishby (2018), proposing a unified formalization of "efficiency" using information theory. According to Zaslavsky et al., efficient communication between speakers is characterized by an optimal trade-off between the complexity of the lexicon (as measured by the number of color terms) and the accuracy of communication.

Despite the impact of the World Color Survey, the data in the survey is limited to unwritten languages, mostly from small-scale societies and non-industrial countries. Languages with many speakers, such as English, Spanish, French, and Portuguese, are missing from the survey. Part of the motivation for this omission was a concern that systems of color terms in industrial societies may have converged to similar structures as a result of contact through globalization (this probably was also the motivation for language selection in Majid et al. (2018)). Data for these languages do exist in the literature but are scattered over different papers with variations in methodology (for example Forbes, 1976; Josserand, Caparos, Pellegrino, & Dediu, 2022; Lindsey & Brown, 2014; M. Xu, Zhu, & Benítez-Burraco, 2023).

Several follow-up studies have further tested the findings of the WCS in a number of other languages and complementary paradigms. Most of these studies involved a small number of languages (e.g., English: Lindsey & Brown, 2014; German: Zollinger, 1984; Greek: Thierry et al., 2009; Turkish: Ozgen & Davies, 1998). It is also well documented that some languages have greater diversity in terms of their shades of blue, in particular, Russian, Hebrew, and Italian (Winawer et al., 2007; Cerqueglini, 2021; Paggetti, Menegaz, & Paramei, 2016). Winawer et al. (2007) showed that Russian and English speakers differ not only in their color terminology, but also in the speed and accuracy of their discrimination of shades of blue.

Interestingly, several studies showed evidence for changes in the number of color terms within periods of just a few decades. For Japanese, for example, Kuriki et al. (2017) have reported that a large number of color categories have significantly evolved in less than fifty years. It was also proposed that the basic color term "pink" is being borrowed from English. In another paper, Gibson et al. (2017) documented changes in the number of color terms produced by Tsimane' participants in the Bolivian Amazon and attributed these changes to industrialization due to the increased use of diversely-colored artificial objects and growing cultural exchange.

This leads us to hypothesize that increased cultural exchange through globalization might lead to an increase in the number of color terms. For example, companies like "Orange" might introduce global concepts of colors adding or changing existing basic color terms. Subsequently, languages might align their color vocabulary stepwise towards the most influential language in a particular region or global languages such as English and Spanish in a way that influences not only niche color terms such as "navy" or "magenta"

Large Language Models

Large language models (LLMs) are a class of deep learning models that have powered a lot of the recent progress in natural language processing and machine learning more broadly (Devlin, Chang, Lee, & Toutanova, 2018; Achiam et al., 2023). Beyond their impact on machine learning, LLMs have drawn considerable attention in the cognitive sciences (Hardy, Sucholutsky, Thompson, & Griffiths, 2023) due to their flexible prompt comprehension capabilities, which allow them to effectively simulate human participants (Marjieh, Sucholutsky, Sumers, Jacoby, & Griffiths, 2022; Marjieh, Rijn, et al., 2023; Acerbi & Stubbersfield, 2023; Dillion, Tandon, Gu, & Gray, 2023) and to serve as a control class against which human behavior can be contrasted.

LLMs are particularly interesting from the perspective of the study of the interaction between perception and language (Chen, Sucholutsky, Russakovsky, & Griffiths, 2024). This is because these models are trained on a substantial chunk of human language and can be used to interrogate the limits of perceptual information that can be extracted from language. Marjieh, Sucholutsky, Rijn, Jacoby, and Griffiths (2023) showed that GPT variants can deliver accurate psychophysical judgments across six perceptual modalities. In a subsequent experiment, the authors also showed that when GPT-4 is subjected to a limited color naming task in English and Russian analogous to the one discussed in the present work, GPT-4 successfully replicated cross-cultural variation in the blue color categories. Zooming out further, a new line of research situates LLMs within a broader lens of "machine culture" (Brinkmann et al., 2023), namely, as interacting agents that mediate cultural processes and innovation. As such, the question of what kind of human cultures are approximated by such agents becomes imperative (Atari et al., 2023). Indeed, by administering questions from the World Values Survey (WVS) to GPT, Atari et al. (2023) showed that GPT's responses aligned best with those of Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies. The observed WEIRD bias of LLMs and their ability to flexibly capture the implications of linguistic variation makes them an ideal control class for studying the effect of globalization on color concept representations. However, LLMs can behave quite differently from humans. For example, they are exposed to a wealth of data from the web from many cultures and might have a single representation for color, leading to similar color maps across all languages. It might also be that due to the rich training data, LLMs behave more like human professionals such as painters or designers, resulting in color maps containing highly specified color terms. This comparison will be the focus of the remainder of the paper.

Methods

Recruiting

We recruited participants from two different crowdsourcing platforms: Prolific[2] and Lucid. Lucid currently supports the recruitment of participants speaking 73 different languages and makes it possible to target speakers in their native language (whereas Prolific has a limited number of supported languages, and all communication on Prolific is in English). We integrated Lucid recruiting with our open-source framework for designing complex online experiments, PsyNet (Harrison et al., 2020)[3].

Participants

Participants interact with the study via a user interface in their browser. Prior to participating, all participants gave informed consent in accordance with the Max Planck Ethics Council (application 2021-42) approved protocol. To participate in our study, participants had to speak the language as their mother tongue, be raised monolingually, hold nationality, and be born in the target country. Before starting the task, participants had to pass a vocabulary test (van Rijn et al., 2023) to make sure they were indeed speakers of the designated language. Data of color-blind participants were excluded from the analysis based on the results of a color blindness test (Clark, 1924; Harrison et al., 2020). We collected data from 25 groups (21 in Lucid and 4 in Prolific, with 22 unique languages across both recruiters; see Figure 1B). We recruited between 40 and 229 participants per group (mean: 91.2, sd: 41.7). Altogether, we collected data from 2,280 participants (Lucid: 1,763, Prolific: 517). The mean age across all participants was 42 (sd: 15) and 31% of the participants have at least a Bachelor's degree. The wage per hour was adapted to the local minimal wage.

Procedure

In accordance with the literature, the participants were presented with 330 Munsell color chips (Cook, Kay, & Regier, 2005). To prevent fatigue, each participant was only presented with a random subset of 50 color chips.

The prompt had to match the following criteria: First, it had to be general enough to be precisely translated into all languages. Second, the prompt needed to convey the concept of basic color terms to restrict the responses and facilitate comparison with the literature. Third, we applied the same prompt for human and LLM agents. We tested various wordings to get responses that elicited reasonably constrained results both from GPT and humans. Thus, we converged on the following: "Please identify the color above using a commonly used color name. The color name should be the one you would normally use in everyday life to describe that color. Avoid using compound words. The color name should be a single word" (see Figure 1C). Since we did not have access to specialists for each language, and in order to maintain translation quality, the prompt was then translated by a professional translation service to all 22 languages[4]. To guarantee high-quality translations, this service involves an initial translation process and validation by a different translator from the same language. Overall, we collected 121,108

[2] https://www.prolific.co

[3] PsyNet is available here: https://www.psynet.dev/.

[4] https://www.translated.com

naming responses from the participants (Mean number of answers per color chip: 13.2 ± 6.4).

Language Models

We conducted our LLM experiments with OpenAI's GPT-4 (Achiam et al., 2023) using the Microsoft Azure OpenAI API (version 0613 of the model). The majority of our GPT-4 experiments were conducted using the default temperature parameter (0.7). We first presented GPT-4 with the following system prompt in English to prepare the model for the experiment: "Follow all instructions that users provide to you in their own language. Respond to users in their own language using only a single word and no other text. Do not use any compound words." For every language and every color, we would then query GPT-4 with the corresponding translation of the following user prompt: "COLOR: <hexcode> Please identify the color above using a commonly-used color name. The color name should be the one you would normally use in everyday life to describe that color. Avoid using compound words. The color name should be a single word." Since GPT-4 responses are stochastic at non-zero temperatures, we sampled 50 responses for every color for each language.

For our corresponding experiments with a Vision Language Model (VLM), we used OpenAI's GPT-4V (version gpt-4-vision-preview on Microsoft Azure's OpenAI API). We similarly first presented GPT-4V with the same system prompt as GPT-4 (for all languages but English as the model would otherwise get confused about which language to respond in). We would then query GPT-4V with a Base64 encoded image containing only a square of the relevant color and the relevant translation of the same user prompt as GPT-4 with the first line containing the hex code removed. Due to low rate limits, we could only sample a single response per color per language so we conducted these experiments with temperature set to zero to elicit the most probable response.

Preprocessing

Since participants and LLMs provided free text, the responses had to be processed. We excluded responses that contained spaces, digits, or punctuation marks. Furthermore, the word had to be written in the expected script (e.g., a Russian color term in Cyrillic). We performed lemmatization to remove word variants (Barbaresi, 2024) and replaced them with the most common variant. In the next steps, we removed diacritics (e.g., "rosá" to "rosa") and replaced characters with smaller units in Korean (Hangul) and Japanese (Katakana) to detect the same word written differently. For character-based languages, we also removed the word "color" since the word was often added to the color term. We replaced all variants under the same simplified form with the most common variant. To detect compound words, we identify the top color terms (occurring in $> 1\%$ of all responses) and check all other color terms for whether they end with this term. If they do and also co-occur, we replace the compound word with the top color term. For example, in Dutch, we would replace "donkerblauw" (dark blue) with "blauw" (blue). For all words, we look up their word frequency in a large text corpus (Speer, 2024). Words that do not occur in the corpus are unlikely to be color terms "you would normally use in everyday life" and are likely to be typos. For each of the typos, we obtain a list of color terms it co-occurs with. For each of those terms, we compute the Levenshtein Distance (Seatgeak, 2024) and merge if they match (score $> 80\%$). We exclude all color terms that are used less than five times in total. For all minority colors (occurring less than 1% of all responses), we merge them with the majority color it co-occurs most with. Terms without co-occurrence are removed. Terms are only removed if this would not lead to removing all color terms in one chip (e.g., if a particular term is used only for one chip and never for the remaining 329 chips). Since we only have a single response per chip from GPT-4V, we did not apply the pipeline to the VLM.

Results

Human Data

Figure 2A shows the English color maps obtained from Prolific and Lucid and compares them to the maps by Lindsey and Brown (2014). In these maps, each chip is colored by the majority vote for this particular chip. We plot the color of the region associated with a term to be the average color (in RGB space) across all winning chips. As indicated by the legends, participants mention similar color categories. The Adjusted Rand Index (ARI) is a measure of how similar different clusterings are (Hubert & Arabie, 1985). We found a high ARI between Lindsey and Brown (2014) and Prolific and Lucid (.72 [.67, .77] and .67 [.62, .72], respectively, CIs via bootstrapping). We further found that for languages that were tested on both recruiters (English, Italian, and Greek), the ARIs were also high (.70 [.65, .74], .61 [.56, .66], and .63 [.57, .68], respectively). Overall, we found the split-half ARI value for all languages to be high (range: .61-.80, mean: .69 sd: .05), with some exceptions in Korean (.42), Japanese (.43), and Norwegian (.56). These findings suggest that the human maps were reliable, and our findings are consistent with prior literature for English. As a sanity check, we also validated the plausibility of the answers with native speakers of 10 of the 20 languages.

Figure 2B provides a few examples of color maps for languages other than English. We showed that while Dutch still resembles the English color maps (.70 [.65, .76]), Italian (.59 [.55, .63]), Russian (.55 [.49, .60]), and Japanese (.42 [.38, .46]) show a more distinct pattern with more differentiation between light blue colors (Winawer et al., 2007). In Italian, for example, we see a distinction between three shades of blue: "blu", "azzurro", and "celeste". Comparing the distance between maps within our dataset and WCS we observed that the mean ARI of our online groups is higher (.64 [.43, .85]) than within the color maps of the WCS (.43 [.18, .68]), suggesting that our color maps were more uniform compared with the WCS. The ARIs across both datasets were significantly lower (.30 [.05, .55]) than within the newly collected

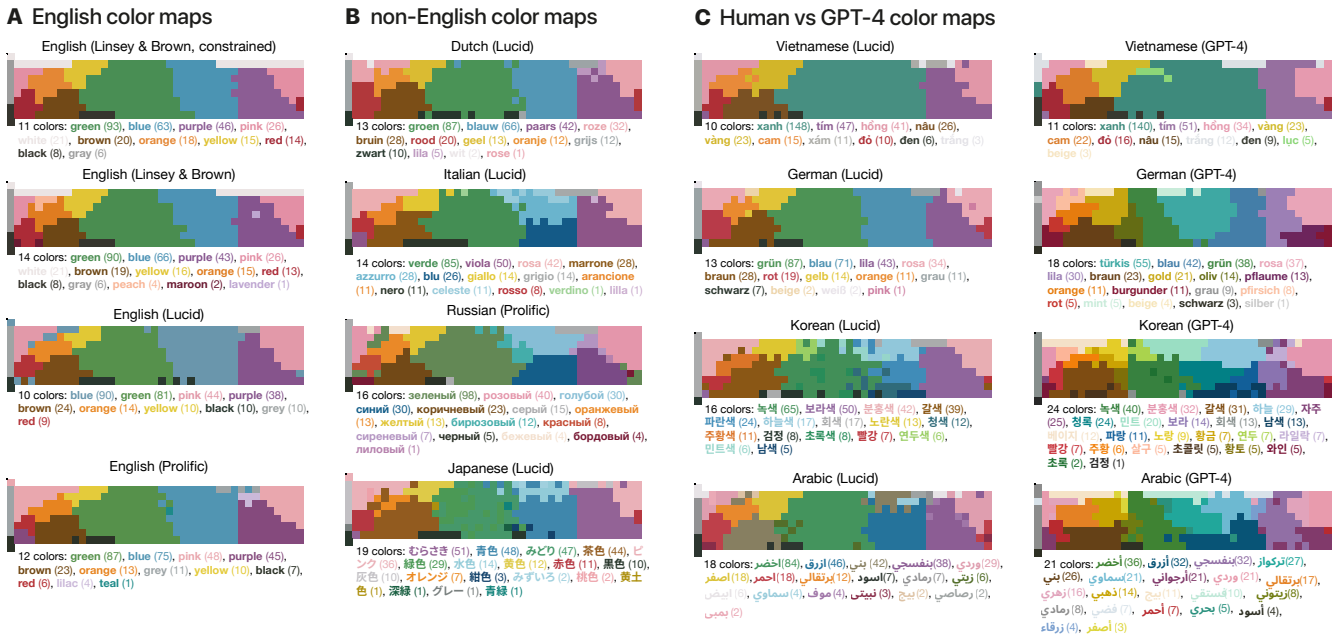

Figure 2: Example color maps. The number after the color term indicates the number of chips. Human color maps: **A** Different color maps were collected on Prolific and Lucid and by Lindsey and Brown (2014). Data reproduced with permission. **B** Collected maps for Dutch, Italian, Russian and Japanese. **C** Comparison of languages in humans and GPT-4. All maps can be viewed interactively on: `https://global-colors.s3-eu-central-1.amazonaws.com/index.html`.

data and the WCS ($p < 0.001$ for both cases via the Mann-Whitney U test), highlighting the difference between the two datasets.

Importantly, the number of color terms in our new data was significantly larger on average (8.54 [5.94, 11.13]) compared to the WCS (5.71 [2.57, 8.85], $p < 0.001$ via the Mann-Whitney U test), where we computed the number of color terms by computing the exponent of the entropy on all chips (Figure 3A x-axis). This is in line with the proposal of Gibson et al. (2017) that industrial societies exhibit a larger number of color terms. However, a broader set of languages needs to be examined to more clearly separate the effects of these and other demographic factors. Furthermore, we found that the variance in the number of color terms in the new data (1.32) was smaller but still amounts to 83% of the variance in the WCS (1.60). This suggests internet users still use varying numbers of color terms to describe colors.

Human and Large Language Models Comparison

Figure 2C shows examples of human and LLM color maps from the same language. While the maps look qualitatively similar and have similar color terms for the same language, GPT-4 maps exhibit a larger number of color terms.

To quantify this effect, we plotted in Figure 3A the number of color terms and the number of distinct responses per chip. We found that the newly collected color maps (red) vary in their number of color terms but are significantly higher (mean: 8.5, SD: 1.3) than the WCS color maps (mean: 5.7, SD: 1.6; $p < 0.001$). Furthermore, we see that both GPT-4 (dark blue) and GPT-4V (light blue) contain more color categories than humans. The vertical-axis of Figure 3A represents the average consensus (number of distinct responses) within each chip. The WCS data showed the most diverse consensus, while GPT-4 provided far fewer distinct answers compared with our new human data. Due to the way the experiment was conducted (see Methods), GPT-4V always provided a single answer.

In Figure 3B we compare the number of color terms per language for GPT-4 and humans. We found that the number of color categories is significantly smaller in humans compared to GPT-4 and GPT-4V (Wilcoxon signed rank test: $t(24) = 5.0$, $p < 0.001$, and $t(24) = 3.0$, $p < 0.001$). However, we see that the number of color categories is correlated across GPT-4 and humans ($r = .39$), indicating that languages with a small number of categories in humans (such as Vietnamese) also tend to have fewer categories in GPT-4. Also, on a single language level, humans fairly overlap with the labels proposed by GPT-4 (mean: 59%, SD: 16%), and the overlapping colors occur at a similar frequency except for Arabic ($r = .06$) and Dutch ($r = .32$) (.44-.90, mean: .65, SD: .13).

In a follow-up analysis, we computed the average word frequency (Speer, 2024) of the colors chosen by humans and those by GPT-4 and weighted them by the frequency of the color category (Figure 3C). We found that the color categories proposed by GPT-4, on average, are less frequent words in that language compared to the color terms used by humans ($t(24) = 53.0$, $p = 0.002$) but still highly correlated ($r = .70$). To illustrate the effect in English, we measure the distance between the colors in RGB space for Lucid and GPT-4 (Figure 3D). While human participants restrict themselves to basic color terms such as "yellow", "red", or "blue", GPT-4 uses colors such as "peach", "bronze", "gold", "olive", "mint",

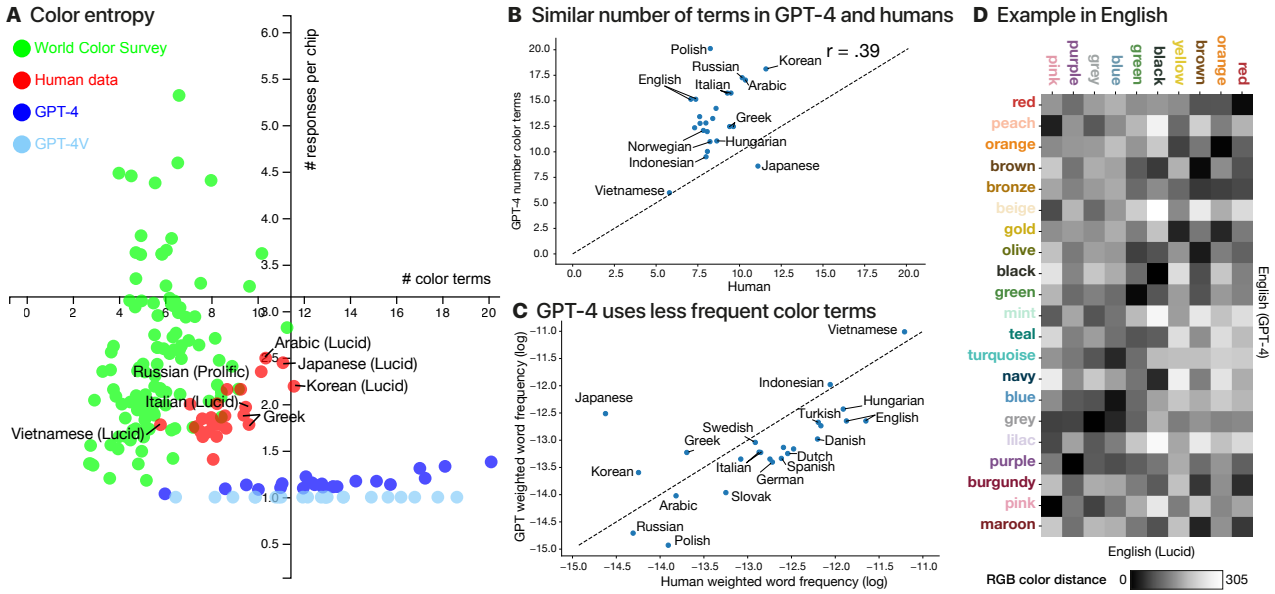

Figure 3: **A** Color entropy space. X-axis: the number of color terms. Y-axis: the number of distinct responses per chip . **B** Number of color terms in humans and GPT-4. **C** Word frequency taken from a corpus of selected color terms. The frequency is weighted by the frequency of the color term. **D** Comparison of color terms in humans (Lucid) and GPT-4 for English.

"navy", "burgundy", and "maroon". These results suggest that because GPT-4 lacks common sense knowledge of what should be conveyed in everyday conversation and is mainly exposed to written text, it uses color terms that are uncommon in everyday language (Zaslavsky et al., 2018; Regier & Kay, 2009).

Discussion

Our recruiting technology allowed us to recruit participants from 22 languages across the world. Even though we tested online participants with considerable access to the internet and global media, the results exhibited significant diversity in color maps and the number of color terms. Shades of blue, for instance, were differentiated in some languages (e.g., Italian, Russian, and Japanese) but not in others (e.g., Dutch and English). Some languages showed a large number of color terms (e.g., Korean) while others had a smaller number (e.g., Vietnamese). We also found that participants recruited from our internet pool had more color terms, on average, than WCS participants who were recruited from small-scale societies. This supports the idea that globalization and industrialization contribute to an increase in the number of color terms (Gibson et al., 2017). Similarly, we found that GPT agents exhibited a diversity in vocabulary size. Across languages, we found a correlation between the number of color terms in humans and GPT. GPT, however, almost always produced maps that contained more color terms than humans. Furthermore, GPT tended to use less frequent words than humans such as maroon and burgundy. These results suggest that globalization has not homogenized cultural distinctions, as language continues to be a key factor in the diversity of perception, even for online participants and machine learning agents with multilingual training.

Our work has limitations that point toward future research directions. First, there are some inherent differences between the online color naming task and its in-lab equivalent. Specifically, there is no control or calibration of the color presentation, and the physical apparatus of color chips typically used in WCS has no online equivalent. However, the fact that we found consistency with the in-lab data in English suggests that these approaches are compatible, and future work should look into that more systematically. Second, while 22 languages is a substantial set, our sample does not have enough representation of the global South, in particular, South America and Africa (Barrett, 2020; Blasi et al., 2022). We believe that covering such locations should be possible with our recruitment strategy, and we plan to follow up on that in the near future. Third, our analysis was focused on data at the population level, but it is reasonable to assume that there is some degree of variation in color concepts across individuals from within the same culture (Lindsey & Brown, 2014). Our sample size was not sufficient to probe individual-level variation in color naming, however, this should be feasible with larger online recruitment batches. Fourth, color naming is only one out of multiple possible experimental paradigms for studying color representations. It remains to be seen whether other paradigms such as color discrimination (Winawer et al., 2007), serial reproduction (J. Xu, Dowman, & Griffiths, 2013), and similarity judgments (Marjieh, Sucholutsky, et al., 2023) would lead to similar conclusions. In particular, discrimination experiments can identify the contribution of perceptual sensitivity. We hope to engage with all of these directions in future work. More broadly, our work showcases how progress in online, diverse recruiting and advances in machine learning can be harnessed to revisit classical cognitive science questions.

Acknowledgments

This work was supported by Microsoft Azure credits supplied to Princeton and a Microsoft Foundation Models grant, as well as an NSERC fellowship (567554-2022) to IS. We extend our gratitude to Lindsey et al. (Lindsey & Brown, 2014) for granting access to the data collected in their study.

References

Acerbi, A., & Stubbersfield, J. M. (2023). Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, *120*(44). doi: 10.1073/pnas.2313790120

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... others (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Al-rasheed, A. S. (2014). Further Evidence for Arabic Basic Colour Categories. *Psychology*, *5*(15), 1714–1729. doi: 10.4236/psych.2014.515179

Atari, M., Haidt, J., Graham, J., Koleva, S., Stevens, S. T., & Dehghani, M. (2023). Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*, *125*(5), 1157–1188. doi: 10.1037/pspp0000470

Barbaresi, A. (2024). *simplemma.* `https://github.com/adbar/simplemma`. GitHub.

Barrett, H. C. (2020). Towards a cognitive science of the human: cross-cultural approaches and their urgency. *Trends in cognitive sciences*, *24*(8), 620–638.

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution.* Berkeley, CA: University of California Press.

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on english hinders cognitive science. *Trends in cognitive sciences*, *26*(12), 1153–1170.

Brinkmann, L., Baumann, F., Bonnefon, J.-F., Derex, M., Müller, T. F., Nussberger, A.-M., ... Rahwan, I. (2023). Machine culture. *Nature Human Behaviour*, *7*(11), 1855-1868. (Number: 11 Publisher: Nature Publishing Group) doi: 10.1038/s41562-023-01742-2

Cerqueglini, L. (2021). Changes in Mutallat Arabic color language and cognition induced by contact with Modern Hebrew. *Studies in the Linguistic Sciences: Illinois Working Papers*.

Chen, A., Sucholutsky, I., Russakovsky, O., & Griffiths, T. L. (2024). Analyzing the roles of language and vision in learning from limited data. *arXiv preprint arXiv:2403.19669*.

Clark, J. (1924). The Ishihara test for color blindness. *American Journal of Physiological Optics*.

Cook, R. S., Kay, P., & Regier, T. (2005). The World Color Survey Database. In *Handbook of categorization in cognitive science* (pp. 223–241). Elsevier.

Davies, I., & Corbett, G. (1994). The basic color terms of Russian. *Linguistics*, *32*, 65–90. doi: 10.1515/ling.1994.32.1.65

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600. doi: 10.1016/j.tics.2023.04.008

Fine, B. (2016). *The world of consumption: the material and cultural revisited*. Routledge.

Forbes, I. (1976). *Structural semantics with particular reference to the vocabulary of colour in modern standard French* (Unpublished doctoral dissertation). The University of Edinburgh.

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., ... Conway, B. R. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, *114*(40), 10785–10790. doi: 10.1073/pnas.1619666114

Hardy, M., Sucholutsky, I., Thompson, B., & Griffiths, T. (2023). Large language models meet cognitive science: LLMs as tools, models, and participants. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 45).

Harrison, P. M. C., Marjieh, R., Adolfi, F., van Rijn, P., Anglada-Tort, M., Tchernichovski, O., ... Jacoby, N. (2020). *Gibbs Sampling with People.* arXiv. (arXiv:2008.02595 [cs, q-bio, stat]) doi: 10.48550/arXiv.2008.02595

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of homo economicus: behavioral experiments in 15 small-scale societies. *American Economic Review*, *91*(2), 73–78.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83. (Publisher: Cambridge University Press) doi: 10.1017/S0140525X0999152X

Hubert, L., & Arabie, P. (1985, December). Comparing partitions. *Journal of Classification*, *2*(1), 193–218. doi: 10.1007/bf01908075

Inglehart, R., Basanez, M., Diez-Medrano, J., Halman, L., & Luijkx, R. (2000). World values surveys and European values surveys, 1981-1984, 1990-1993, and 1995-1997. *Ann Arbor-Michigan, Institute for Social Research, ICPSR version*.

Josserand, M., Caparos, S., Pellegrino, F., & Dediu, D. (2022, September). The Colour Lexicon Is Shaped by Environment and Biology: Comparing Himba and French Colour Perception. In A. Ravignani et al. (Eds.), *Joint Conference on Language Evolution* (p. 368-370). Kanazawa, Japan: Joint Conference on Language Evolution (JCoLE).

Kay, P., & Cook, R. S. (2016). World Color Survey. In M. R. Luo (Ed.), *Encyclopedia of Color Science and Technology* (pp. 1265–1271). New York, NY: Springer New York. doi: 10.1007/978-1-4419-8071-7_113

Kramsch, C. (2014). Language and culture. *AILA review*, *27*(1), 30–55.

Kuriki, I., Lange, R., Muto, Y., Brown, A. M., Fukuda, K., Tokunaga, R., ... Shioiri, S. (2017). The modern Japanese color lexicon. *Journal of Vision*, *17*(3), 1. doi: 10.1167/17.3.1

Lenneberg, E. H., & Roberts, J. M. (1956). *The language of experience: a study in methodology* (No. Memoir 13). Waverly Press.

Lindsey, D. T., & Brown, A. M. (2014). The color lexicon of American English. *Journal of Vision*, *14*(2), 17–17. doi: 10.1167/14.2.17

Majid, A., Roberts, S. G., Cilissen, L., Emmorey, K., Nicodemus, B., OâGrady, L., ... Levinson, S. C. (2018). Differential coding of perception in the worldâs languages. *Proceedings of the National Academy of Sciences*, *115*(45), 11369-11376. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.1720419115 doi: 10.1073/pnas.1720419115

Marjieh, R., Rijn, P. V., Sucholutsky, I., Sumers, T., Lee, H., Griffiths, T. L., & Jacoby, N. (2023). Words are all you need? language as an approximation for human similarity judgments. In *The Eleventh International Conference on Learning Representations* .

Marjieh, R., Sucholutsky, I., Rijn, P. v., Jacoby, N., & Griffiths, T. L. (2023). Large language models predict human sensory judgments across six modalities. *arXiv preprint arXiv:2302.01308*.

Marjieh, R., Sucholutsky, I., Sumers, T., Jacoby, N., & Griffiths, T. (2022). Predicting human similarity judgments using large language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44).

Ozgen, E., & Davies, I. (1998). Turkish color terms: Tests of Berlin and Kay's theory of color universels and linguistic relativity. *Linguistics*, *36*, 919–956. doi: 10.1515/ling.1998.36.5.919

Paggetti, G., Menegaz, G., & Paramei, G. V. (2016). Color naming in Italian language. *Color Research and Application*, *41*, 402-415.

Pieterse, J. N. (2019). *Globalization and culture: Global mélange*. Rowman & Littlefield.

Regier, T., & Kay, P. (2009). Language, thought, and color: Whorf was half right. *Trends in Cognitive Sciences*, *13*(10), 439–446. doi: 10.1016/j.tics.2009.07.001

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, *104*(4), 1436–1441. doi: 10.1073/pnas.0610341104

Seatgeak. (2024). *Thefuzz*. https://github.com/seatgeek/thefuzz. GitHub.

Speer, R. (2024). *wordfreq*. https://github.com/rspeer/wordfreq. GitHub.

Thierry, G., Athanasopoulos, P., Wiggett, A., Dering, B., & Kuipers, J.-R. (2009). Unconscious effects of language-specific terminology on preattentive color perception. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *106*(11), 4567–4570. (Place: US Publisher: National Academy of Sciences) doi: 10.1073/pnas.0811155106

Triandis, H. C. (2018). *Individualism and collectivism*. Routledge.

van Rijn, P., Sun, Y., Lee, H., Marjieh, R., Sucholutsky, I., Lanzarini, F., ... Jacoby, N. (2023). Around the world in 60 words: A generative vocabulary test for online research. In *Proceedings of the 45th annual conference of the cognitive science society*. (Vol. 45).

Whorf, B. L. (2012). *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf* (J. B. Carroll, S. C. Levinson, & P. Lee, Eds.). The MIT Press.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, *104*(19), 7780–7785. (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.0701644104

Xu, J., Dowman, M., & Griffiths, T. L. (2013). Cultural transmission results in convergence towards colour term universals. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1758), 20123073. doi: 10.1098/rspb.2012.3073

Xu, M., Zhu, J., & Benítez-Burraco, A. (2023). A comparison of basic color terms in Mandarin and Spanish. *Color Research & Application*, *48*(6), 709–720. doi: 10.1002/col.22863

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, *115*(31), 7937–7942. doi: 10.1073/pnas.1800521115

Zollinger, H. (1984). Why just turquoise? Remarks on the evolution of color terms. *Psychological Research*, *46*(4), 403–409. (Place: Germany Publisher: Springer) doi: 10.1007/BF00309072