

# Colour in Translation: Data, Models, and Benchmarking for Cross-Linguistic Colour Naming

Dimitris Mylonas  
dimitris.mylonas@nulondon.ac.uk  
Northeastern University London  
London, United Kingdom

Akvile Sinkeviciute  
a.sinkeviciute@nulondon.ac.uk  
Northeastern University London  
London, United Kingdom

Rafique Ahmed  
ahmed.rafi@nulondon.ac.uk  
Northeastern University London  
London, United Kingdom

Alexandros Koliouis  
alexandros.koliouis@nulondon.ac.uk  
Northeastern University London  
London, United Kingdom

## Abstract

Colour naming links vision and language. Yet, effective cross-linguistic colour communication is limited by the lack of multilingual data and computational models for comprehensive colour name translation. We collected 6,408 unique colour naming responses in five languages using online experiments and fieldwork. For each language, we train a *spin colour forest*, a novel partially rotated decision trees model that accurately estimate colour naming distributions across the full gamut, consistently outperforming existing methods. Unlike prior work that assumed 11 universal colour categories, our results reveal cross-linguistic variation in naming granularity: American English uses 47 indispensable colour names, British English 32, French 27, Greek 32, and the Himba 7 to categorise the same perceptually uniform colour space. Building on these findings, we develop a colour translation benchmark, which we demonstrate by evaluating both the lexical and perceptual accuracy of a large language model. Our evaluation reveals a critical lexical-perceptual disconnect, demonstrating that language models lack perceptual grounding in colour translation. Our data, models, and benchmark provide an empirical foundation for inclusive design that reflects how people communicate colour across cultures.

## CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **Visualization**; • **Computing methodologies** → **Machine learning**; • **Applied computing** → **Language translation**.

## Keywords

colour, naming, cross-lingual, translation, benchmark

### ACM Reference Format:

Dimitris Mylonas, Rafique Ahmed, Akvile Sinkeviciute, and Alexandros Koliouis. 2026. Colour in Translation: Data, Models, and Benchmarking for Cross-Linguistic Colour Naming. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*, April 13–17, 2026, Barcelona, Spain. ACM, New York, NY, USA, 21 pages. <https://doi.org/10.1145/3772318.3791626>



This work is licensed under a Creative Commons Attribution 4.0 International License. *CHI '26, Barcelona, Spain*

© 2026 Copyright held by the owner/author(s).  
ACM ISBN 979-8-4007-2278-3/2026/04  
<https://doi.org/10.1145/3772318.3791626>

## 1 Introduction

Colour naming sits at the intersection of perception and cognition with direct implications for human-computer interaction (HCI). Humans compress thousands of perceivable colours into a smaller set of names such as *yellow*, *turquoise*, or *olive green* [31, 42, 58, 92]. These lexicons vary across languages in size, lexicalisation, and category boundaries. For example, English distinguishes red and pink, while some languages such as Himba uses a single term, *serandu*, to name both regions of colour space [63, 72]. Colour names are not mere labels: they mark perceptually meaningful regions of the colour space that support memory, recognition, and discrimination in HCI tasks such as data visualisation, interface design, and cross-cultural communication [19, 27, 30, 32, 51].

Colour naming experiments provide an empirical foundation to investigate the mapping between colour names and the corresponding regions of perceptual colour spaces. Since the 1950s, researchers have elicited names from speakers using controlled colour samples such as colour chips from the “Munsell Book of Color” [57] to study how languages partition colour space and how naming relates to perception and cognition [6, 10]. The World Colour Survey [38] and subsequent laboratory [8, 78], field [21], and online colour naming experiments [54, 56, 61] have revealed both cross-linguistic diversity and recurrent regularities in colour categorisation. For example, while languages differ in vocabulary size, many converge to a shared subset of colour terms. These experimental methods form the methodological foundation for our own cross-lingual colour naming experiments in British English, American English, French, Greek, and Himba [31, 59, 61, 63, 66].

Computational colour naming models aim both to assign names to colours and to interpret colour names that are meaningful across languages. We can, for instance, ask a model to translate *turquoise* into another language, or to provide its best colour example—the *focal colour*. Yet, alignment with human colour categories is not guaranteed. Most models are shaped by the Basic Colour Terms (BCT) framework, assuming a universal inventory of up to eleven colour categories: in English, these are *black*, *white*, *red*, *yellow*, *green*, *blue*, *brown*, *orange*, *purple*, *pink*, and *grey* [6, 38]. However, by presuming universal categorisation, such models miss the richer, language-specific partitioning observed in actual language use. Figure 1 illustrates this key limitation: given the continuous surface of a perceptual colour space (Figure 1, left), the 9 (excluding white and



**Figure 1: Segmentation of perceptual colour space showing the gap between assumed eleven universal categories and actual language use: (left) perceptually continuous surface colours; (middle) nine Basic Colour Terms in British English (excluding white and black from the canonical eleven); (right) sixteen indispensable British English colour names.**

black) basic terms in British English provide only coarse segmentation (Figure 1, middle), while, based on our data, British English speakers use 16 colour names to naturally describe the same surface colours, revealing finer-grained boundaries than BCT (Figure 1, right) [63]. Consequently, BCT-based models limit the effectiveness of natural language [32, 51] and complicate direct translation [41] for colour selection and data visualisation in HCI.

Recent advances in multilingual large language models [11, 23] and chatbots [47] have raised expectations for effective machine colour translations. Colour names, however, are not standard lexical items: category regions vary across languages, and many terms lack one-to-one equivalents. For example, English *blue* maps to two distinct basic terms in Greek (*γαλάζιο/galazio* for light and *μπλε/ble* for dark blue) and Lithuanian (*žydra/zydra*, *mėlyna/melyna*) [41, 76]. Machine translation or text generation systems typically map words rather than align colour categories, leading to systematic perceptual mismatch [65, 81]. For example, *periwinkle* in English is challenging to translate into languages without a corresponding flower reference. Prior work also shows that language models achieve competitive quality for high-resourced languages but perform less reliably in low-resource settings [33, 35]. Existing terminology translations focus mainly on technical domains such as medicine and finance [84], leaving colour largely unexplored. To date, no systematic multilingual colour naming benchmark exists.

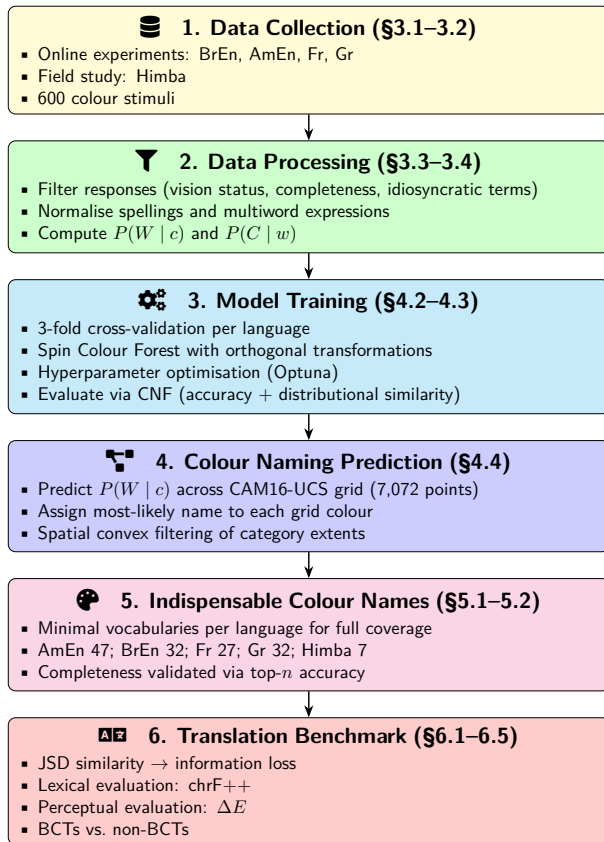
Effective colour communication requires more than lexical translation: it requires understanding of how different communities perceive and talk about colour [62]. Without it, cross-cultural visual communication and interaction risk misinterpretation or exclusion. We present new and existing systematic data, models, and benchmarks for cross-lingual colour naming. Our three key contributions provide an empirical foundation for inclusive design that reflects how people communicate colour across cultures:

- (1) **Multilingual datasets.** We release unconstrained colour naming datasets (i.e., freely named colours without predetermined category restrictions) for American English, British

English, French, Greek and Himba, collected through a combination of online crowdsourcing and offline fieldwork. The dataset consists of 70,052 responses to 600 colour stimuli, extending beyond basic colour terms and enabling richer cross-linguistic comparison in CAM16-UCS, a perceptually uniform colour space.

- (2) **Spin Colour Forests.** We introduce a novel ensemble regression method for naming colours. By applying partial orthogonal transformations in perceptual colour space, our model accurately captures colour category boundaries, leading to the identification of the indispensable vocabularies required to name the entire colour gamut in each language: 47 for American English, 32 for British English, 27 for French, 32 for Greek, and 7 for Himba.
- (3) **Colour translation benchmark.** Using our indispensable colour name sets, we establish translation pairs across languages, proposing the first, to our knowledge, colour translation benchmark. As a demonstrator, we measure whether AI translation systems respect cultural diversity in colour categorisation or merely substitute words while breaking perceptual correspondence.

Figure 2 presents our complete methodological workflow, organised into six main stages: (1) we collected colour naming data through online experiments for four languages and field studies with the Himba people of Namibia; (2) we then processed these responses through filtering and normalisation procedures to compute probabilistic naming patterns; (3) using processed data, we trained Spin Colour Forest models via cross-validation and hyperparameter optimisation to (4) predict colour naming across the colour space; (5) we identified the indispensable colour vocabularies—the minimal set of terms required for complete colour space coverage in each language; finally, (6) we developed a translation benchmark that evaluates large language models on both lexical accuracy using character-level similarity scores (chrF++) [70] and perceptual accuracy using CAM16-UCS colour space distance ( $\Delta E$ ) [48].



**Figure 2: Methodological workflow from multilingual data collection to model training, indispensable vocabulary identification, and translation benchmarking.**

## 2 Related Work

Our work builds on prior work on colour naming (§2.1), colour-naming experiments (§2.2), computational models (§2.3) and machine translation (§2.4). We review these foundations and highlight the limitations that motivate our datasets, model and translation benchmark.

### 2.1 Colour Naming

The number of colour names in languages is often large, which improves colour communication precision [45]. Yet, only a smaller number of colour names are shared and well understood by speakers in each language [8, 10, 21, 92]. Berlin and Kay’s study [6] reported shared regularities on the surface of the colour solid across cultures and proposed a universal inventory of eleven basic colour categories, the Basic Colour Terms (BCTs). Their criteria for the identification of BCTs were based on multiple factors (e.g. single word terms that are not the name of an object) judged by experts as not being equally applicable across languages [7, 18, 74]. Others have segregated BCTs on more rigorous behavioural criteria such as frequency, response time and consistency [8, 21, 61, 78].

Explanations for the development of colour naming systems range from universalist accounts grounded in physical, optical and

physiological processes [39, 68, 71, 75] to cultural accounts emphasising the communicative need to describe object properties [10, 26, 29, 77]. Current consensus suggests that both perception and language influence the cognitive organisation of colour, leading to a growing view that basicness is not a simple dichotomy between basic and non-basic terms, but rather a continuous, gradual characteristic of lexical colour spaces [86]. The quest for a more cross-culturally legitimate approach to identify the minimal set of names that can name all colours thus remains open.

### 2.2 Colour Naming Experiments

Understanding how languages categorise colour requires systematic empirical methods for eliciting and analysing naming behaviour across diverse linguistic and cultural contexts. Colour naming experiments have long provided the empirical foundation for studying how languages partition colour space. Early work by Brown and Lenneberg [10] introduced the concept of linguistic codability, demonstrating systematic links between naming behaviour and recognition. Berlin and Kay extended this approach cross linguistically in 110 languages, using Munsell samples to elicit best examples of BCTs, what they termed colour foci. Their methodology inspired the World Color Survey [38], which gathered constrained naming and focus data in 110 unwritten languages, revealing both broad commonalities and some variation in categorical boundaries.

Subsequent laboratory studies refined these methods by mapping BCTs in perceptual spaces. Boynton and Olson [8] established a quantitative approach to deriving colour foci in the OSA system as the stimuli receiving both the highest naming consensus (agreement across participants) and shortest response times within each colour category, demonstrating that these focal colours are named more frequently, quickly, and consistently than peripheral category members. Sturges and Whitfield [78] extended this work in the Munsell system, while Davies and Corbett [21] introduced measures of dominance and stability to quantify consensus in the field. Zeki and colleagues [93, 94] used Land Colour Mondrian experiments to name and match colours to Munsell chips under different viewing conditions in a Bayesian framework. Heer & Stone’s [32] probabilistic approach for determining foci used the average of the 4 top most likely colours of each name. In parallel, field studies adapted elicitation procedures to smaller communities, showing the importance of local saliency alongside cross-linguistic comparability [15, 20].

More recently, online experiments have scaled colour naming to large populations: Moroney’s distributed psychophysics [54] reproduced lab results at scale, Munroe’s XKCD dataset [56] generated millions of unconstrained responses for computational modelling, and the ongoing multilingual experiment by Mylonas [61, 63, 66] has collected tens of thousands of responses across more than twenty languages with associated metadata and compared them against consistent laboratory [60] and fieldwork settings [59].

Together, these traditions, from laboratory precision to field adaptability and online scalability, have established colour naming as a robust paradigm for probing both universal tendencies and language-specific variation. However, most existing datasets are either restricted to single languages or to BCT inventories, limiting their usefulness for cross-lingual modelling and translation.

Our work builds on this foundation by providing new and existing unconstrained multilingual datasets that extend beyond BCTs and by introducing models and benchmarks designed to capture and evaluate perceptual meaning across languages.

### 2.3 Colour Naming Models

A wide range of computational colour naming models have been developed to support colour communication across domains and modalities. Early efforts, such as the ISCC-NBS dictionary [40], mapped thousands of English colour names to regions in the Munsell system to enable cross-disciplinary colour standardisation. However, its ad hoc vocabulary and lack of formal syntax limited usability. The Colour Naming System [5] introduced a more structured approach by aligning with the BCTs and allowing intermediate hues to be defined through systematic combination in HSL colour space.

Subsequent models incorporated perceptual and probabilistic techniques to improve accuracy and robustness. Tominaga [79] proposed a multi-level naming hierarchy derived from digitised Munsell samples, while Lammens [44] applied fuzzy set theory to model BCT membership distributions. Motomura [55] used Mahalanobis distances to preserve BCT category mapping across media. Lin *et al.* [49, 50] and Menegaz *et al.* [53] focused on defining category boundaries using experimental data and fuzzy interpolation in CIELAB space. Benavente *et al.* [4] extended Lammens' computational model with Gaussian-Sigmoid distribution functions, while Parraga and Akbarinia [67] introduced a biologically inspired model based on cone contrast encoding.

In parallel, data-driven models leveraged image statistics and large-scale naming datasets. Weijer *et al.* [83] estimated colour distributions for the BCTs from Google Image search results using probabilistic latent semantic analysis (PLSA). While rapid automated training approaches scale efficiently compared to expensive colour naming experiments, they suffer from poor correspondence with psychophysical data [67] and cannot capture cross-linguistic differences that are crucial for multilingual applications. Chuang *et al.* [14] and Heer & Stone [32] used existing colour naming data [38, 56] to develop probabilistic models of naming saliency that corresponded well to the BCTs and demonstrating practical applications in colour selection, image editing, and palette design.

The first significant extension to multilingual contexts came from Kim *et al.* [41], who created models across 14 languages and revealed systematic cross-linguistic differences in colour categorisation. Like Heer & Stone [32] and Chuang *et al.* [14], their translation models [41] were narrowed down to BCTs with only 10 clusters corresponding to English BCTs and 16 for Korean, limiting applicability to the rich colour vocabularies people use in their native languages. Mylonas *et al.* [64] applied a Maximum a Posteriori estimator to crowdsourced colour naming data, enabling classification into a broader set of 47 commonly used English terms.

These models have supported diverse applications across human-computer interaction, including colour selection interfaces [32], categorical palettes for data visualisation [30, 51], and categorical perception in interactive graphics [80]. They have also been applied in computer vision domains such as image understanding, enhancement, and person re-identification [13, 89, 91].

Most existing systems remain limited to BCTs and operate under the assumption of universal colour categories. This creates significant challenges for cross-cultural interface design and international user experiences, where the number, boundaries, and semantics of colour categories vary widely across languages.

### 2.4 Machine Translation Tools

Modern Machine Translation (MT) relies on three main approaches: encoder-decoder Neural Machine Translation (NMT), which captures cross-lingual patterns; decoder-only Large Language Models (LLMs) such as Generative Pre-trained Transformer (GPT), which generate context-driven output; and Sparsely Gated Mixture of Experts (MoE) models like NLLB-200 [16], which enhance efficiency and reduce multilingual interference. While purpose-built NMT systems such as NLLB-200 remain strong baselines, Hendy *et al.* [33] showed that LLMs like GPT can achieve competitive performance in high-resource languages (e.g., German) despite being trained largely on monolingual data, though their performance declines for low-resource languages (e.g., Icelandic). The latter reflects a broader issue in low-resource MT: data collection is both resource-intensive and logistically complex, slowing overall progress [43].

In this study, we expect LLMs to find difficulties in translating to Himba, an unwritten language, whilst performing best on English and French, which are consistently classified as high-resource; Greek, however, is labelled high-resource in NLLB [16] but medium-resource in Euas-20 [35]. To systematically quantify language resource availability, we use Wikipedia article counts as a transparent, reproducible proxy measure [85]. English Wikipedia contains 7.1M articles, French 2.6M, Greek 235K, whilst Himba has no presence. This 30,000-fold difference directly reflects the training data disparities that affect LLM performance across languages [37, 46]. Whilst Wikipedia coverage does not capture all dimensions of language resources, it serves as a reliable indicator of textual training data availability that fundamentally shapes neural translation systems [69].

Our focus on colour terminology translation represents a specialised domain where lexical equivalence does not guarantee perceptual or cultural correspondence. Wassie *et al.* [84] provide evidence for such domain-specific challenges, showing that although open-source LLMs exhibit potential for general translation tasks, their performance deteriorates when translating specialized terminology. Their experiments on medical domain translation across four language pairs (English-French, English-Portuguese, English-Swahili, and Swahili-English) revealed that the task-oriented NLLB-200 3.3B model outperformed all evaluated LLMs, particularly evident in medium-resource and low-resource language settings. Enis & Hopkins [24] reported that Claude 3 Opus exhibits exceptional resource efficiency and robustness across domains, outperforming baselines like NLLB-54B and Google Translate in many  $xxx \rightarrow eng$  pairs (e.g., Bengali-English, Maltese-English). Its translation quality is less dependent on resource level compared to other LLMs such as GPT-4 and Llama, and it generalizes well to diverse datasets like BBC News and MASRI-HEADSET. While weaker in  $eng \rightarrow xxx$  overall, Claude still surpasses baselines for some pairs, including English-Korean and English-Thai and will be evaluated here as a case study.

Critical gaps remain in handling cultural diversity. Yao *et al.* [90] introduced benchmarks for evaluating machine translation with cultural awareness, focusing specifically on culture-specific items that often lack direct translations across languages. Their findings reveal that current systems, both NMT and LLM-based, frequently fail to preserve culturally specific conceptual boundaries, with LLMs showing superior capability only when provided with external cultural knowledge through prompting strategies. This limitation is directly relevant to colour naming translation, where cultural variation in colour categorisation creates fundamental challenges beyond simple lexical mapping. Since colour names are culturally specific, context-aware translation has become a practical necessity for creating inclusive and effective multilingual systems.

### 3 Data Collection

We collected colour naming data through online experiments for written languages and field studies for unwritten languages, designed to capture unconstrained colour naming responses across the colour gamut. Our methodology provides consistency while accommodating different cultural and technological contexts.

#### 3.1 Colour Stimuli

Colour stimuli consisted of 600 samples drawn from the Munsell Renotation dataset, including 589 chromatic and 11 achromatic samples, providing approximately uniform coverage of the colour solid. Following Billmeyer's sampling recommendations [78], we used variable numbers of hues at different Value and Chroma levels: 10 hues at Chroma 2, increasing by 10 hues at each successive Chroma step, reaching all 40 hues from Chroma 8 to the sRGB boundaries [64]. This approach maintains an approximately perceptually uniform distribution of samples in the cylindrical Munsell system. Colour stimuli were 2-degree uniformly coloured patches with 1-pixel black outlines displayed for online experiments on participants' own devices and on calibrated monitors for the field studies, both targeting sRGB colour specifications. All presentations used mid neutral grey backgrounds to enable surface colour perception necessary for appropriate naming of colours like grey and brown and to avoid perceptual compression of dark colours as occurs when viewed against a white background [32, 56]. In this study, colour stimuli were transformed to CAM16-UCS [48] that offers state-of-the-art perceptual uniformity for measuring colour differences. Yet, the  $a'$  and  $b'$  opponent chromatic dimensions of this uniform space do not align well with unique hue directions [87, 88], meaning that colour naming boundaries can appear complex in standard coordinates.

#### 3.2 Language-Specific Data Collection

British and American English, French and Greek data were collected through web-based experiments (available at [colornaming.net](http://colornaming.net)) Participation was anonymous, and the experimental sessions were voluntary and conducted after obtaining informed consent. To minimize the number of disruptive and poorly motivated participants a high entrance barrier technique [90] was used requiring participants to adjust the brightness of their display for viewing a greyscale test image and provide metadata on colour vision status, software and hardware specifications, and viewing conditions. Each



**Figure 3: Colour naming task of online colour naming experiment (available at [colornaming.net](http://colornaming.net)).**

participant named sequentially one-at-a-time presented colours using unconstrained text input (Figure 3; for all steps see Supplementary Materials Figure S1).

In total data were collected from 598 American English, 600 British English, 153 French and 527 Greek observers providing 12,000, 11,940, 13,138, 9,908 raw responses with 2,780, 2,369, 1,447, 2,439 unique colour terms respectively. The study received ethical approval from Northeastern University London (No.0001 16th May 2023).

For Himba, a field study in Namibia was conducted using a computerized but offline approach to maintain methodological consistency [59]. Two Asus Transformer Mini T102HA tablets were calibrated towards sRGB specifications using ColorCal CRS colorimeter and RadOMA spectroradiometer. The measured CIE 1931 chromaticity coordinates of the white point of the monitors were  $x = 0.3067$ ,  $y = 0.3318$ , and  $x = 0.3055$ ,  $y = 0.3296$  with a correlated temperature of 6816 K and 6907 K, respectively within 0.003 drift over the fieldwork period. Stimuli were presented using PsychoPy software with observers seated 80cm from monitors inside tents during daytime. Participants named colours aloud in their native language, with responses audio-recorded and transcribed by research assistants fluent in Himba to ensure appropriate cultural context and handling. No sensitive, offensive, or culturally inappropriate terms were encountered in responses. The 55 Himba observers provided 33,000 raw responses across the 600-colour stimulus set. Participant metadata were assigned anonymous identifiers with no linkage to individuals. Participants were compensated for their time with gifts of flour. The field study received ethical approval from Goldsmiths College, University of London (No.1390, 4th of June 2018). Anonymised datasets and analysis code for all languages and variables are available at <https://doi.org/10.17605/OSF.IO/3BQMP> under a CC BY 4.0 licence.

#### 3.3 Data Pre-processing

We excluded responses from observers with possible diverse colour vision (5.59%) and removed incomplete, vernacular, numerical and

empty responses (1.84%). Spelling errors were regularized by native speakers, with hyphenated words treated as multiword expressions. Spelling corrections were applied to 5.09% of American English unique raw terms, 4.78% of British English, 4.66% of French, 2.02% of Greek, and 0.02% of Himba unique terms. For Greek specifically, 16.6% raw responses were offered in Latin characters (Greeklish) rather than Greek script, requiring transliteration to standard Greek orthography before normalisation. Multiword colour expressions (e.g., *dark blue*, *light green*) were preserved with space-to-underscore standardization for data formatting consistency. Multi-word terms comprised 34.3% of American English clean vocabulary (363 terms), 37.6% of British English (395 terms), 32.8% of French (229 terms), 22.7% of Greek (224 terms), and none in Himba, reflecting cross-linguistic variation in colour naming granularity. Responses unique to single participants represented 5.32% of total responses and were eliminated as potentially idiosyncratic. These singleton terms comprised substantial proportions of raw responses: 1,617 terms in American English (68.2% of raw unique responses), 2,007 in British English (72.2%), 871 in French (60.2%), 1,727 in Greek (70.8%), and 10 in Himba (21.3%). This filtering primarily removed typographical errors, ambiguous entries, and data input noise rather than legitimate rare colour terms. The resulting 70,052 clean responses yielded 510, 492, 276, 356, 24 unique colour names for British English, American English, French, Greek and Himba respectively. French, Greek and Himba terms are presented in Latin script transliteration for readability, while preserving the original native language colour names used by participants.

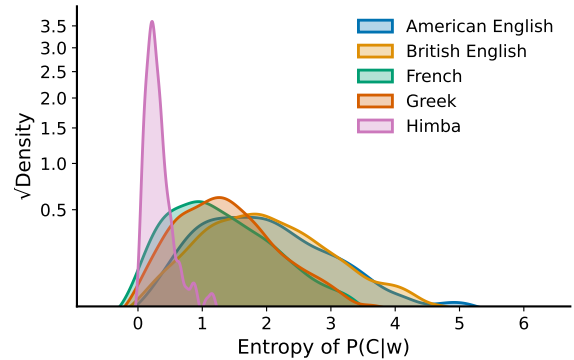
### 3.4 Diversity of Languages

To quantify colour vocabulary diversity within each language, we modelled these multilingual datasets as conditional probabilities  $P(W|c)$  and  $P(C|w)$ , representing the likelihood of word  $w$  given colour  $c$ , and colour  $c$  given word  $w$  respectively. Vocabulary richness was quantified using the Shannon entropy of  $P(C|w)$ , which captures how broadly each name spans the colour space, that is, the specificity of individual terms. Higher average entropy across words indicates greater lexical diversity in colour naming. Vocabulary richness varied across languages (Figure 4). British and American English exhibit the highest average entropies (1.94 and 1.89 bits) and largest vocabularies (510 and 492 distinct terms), reflecting diverse colour naming systems. Greek and French show intermediate complexity (1.34 and 1.21 bits; 356 and 276 terms respectively). In contrast, Himba demonstrates the lowest entropy (0.26 bits) with only 24 distinct terms, consistent with a compact colour naming system. These differences establish the baseline diversity that our multilingual colour naming model must accommodate, spanning from highly detailed industrial vocabularies to efficient traditional systems. Data preprocessing reduced entropy by 21.2–42.6% whilst preserving linguistic complexity ordering (see Supplementary Materials Table S1).

## 4 Mapping Colour Names to Colours

### 4.1 Motivation and Problem Formulation

Understanding how people name colours across cultures is fundamental for designing inclusive interfaces and culturally aware systems. Our challenge lies in modelling colour naming behaviour



**Figure 4: Diversity of colour lexicons in American and British English, French, Greek and Himba measured as the square root of the  $P(C|w)$  Entropy in bits.**

across diverse linguistic communities, from languages with very limited vocabularies to those with much richer lexicons, using computational methods that can generalise to unseen colours in the entire perceptual colour space. Our training data comprises 600 colour samples approximately perceptually uniformly distributed in CAM16-UCS colour space ( $J', a', b'$ ), with naming responses collected from speakers of five languages. From participant responses, we model conditional probabilities  $P(W|c)$ , which serve as regression targets for predicting human colour naming distributions. Our goal is to learn a function  $f: \mathbb{R}_3 \rightarrow [0, 1]^N$ ,  $f(c) \rightarrow P(W|c)$  that maps colour coordinates  $c = (J', a', b')$  across CAM16-UCS to probability distributions over  $N$  colour names.

### 4.2 Spin Colour Forest

We introduce Spin Colour Forest (SCF), a novel ensemble method that applies random orthogonal transformations to the three-dimensional perceptual colour space before training the individual decision trees. The name reflects how the model spins perceptual colour space to find simpler views of category boundaries. The key insight is that colour naming boundaries often look complex in the CIECAM16-UCS ( $J', a', b'$ ) coordinates but in transformed coordinate systems, these boundaries may appear simpler and better aligned with the perceptual structure of human categorisation. For example, boundaries for Greek  $\gamma\alpha\lambda\acute{\alpha}\zeta\iota\omicron$  (light blue) or Himba *serandu* (pink-red) may follow diagonal paths in standard coordinates but appear as clean, axis-aligned splits in rotated views. By training trees in randomly transformed coordinate systems, each tree discovers boundaries oriented at arbitrary angles in the original space, rather than being constrained to splits along lightness, redness-greenness, or yellowness-blueness axes alone. Our Spin Colour Forest extends traditional ensemble approaches [9, 28, 73] by: (a) applying random orthogonal transformations to the full three-dimensional colour space rather than using Principal Component Analysis (PCA) on feature subsets; (b) targeting regression to probability distributions  $P(W|c)$  rather than classification; (c) treating CIECAM16-UCS ( $J', a', b'$ ) coordinates as a single perceptual unit, using dense  $3 \times 3$  transformation matrices; and (d) adopting a fractional transformations strategy, where only a proportion of

the trees receive transformed inputs while the rest use the original inputs, to accommodate varying vocabulary sizes and preserve interpretability.

### 4.3 Algorithm Description

Spin Colour Forest applies systematic orthogonal transformations to training data in CAM16-UCS colour space to enable optimal decision tree construction (Figure 5). For each tree  $t$  in our ensemble of  $T$  trees, we generate orthogonal matrix  $R_t \in \mathbb{R}^{3 \times 3}$  using QR decomposition of random  $3 \times 3$  matrix  $A_t$  with entries drawn from  $\mathcal{N}_{(0,1)}$  following standard practice for random orthogonal matrix generation  $O$  [34]. Orthogonal matrices  $R_t$  are generated once during tree initialization and applied consistently to all training samples for tree  $t$ , ensuring each tree learns decision boundaries in a fixed transformed coordinate system. We apply column sign correction to ensure positive diagonal elements in the upper triangular matrix  $R$ : if  $R[i, j] < 0$  then  $Q[:, i] \leftarrow -Q[:, i]$ . This procedure generates orthogonal matrices  $Q_t$  in  $O(3)$  that include both proper rotations and reflections, enabling exploration of decision boundaries with fixed random geometric orientations including mirror symmetries.

The partial transformation strategy is controlled by  $\rho \in [0, 1]$ :

$$R_t = \begin{cases} Q_t \in O(3) & \text{if } t \leq \lfloor \rho T \rfloor \\ I_3 & \text{if } t > \lfloor \rho T \rfloor \end{cases}$$

For transformed features:

$$X'_t = X_{\text{scaled}} \times R_t$$

where  $X_{\text{scaled}}$  represents standardized CAM16-UCS coordinates ( $J'$ ,  $a'$ ,  $b'$ ). The ensemble prediction combines both geometric perspectives:

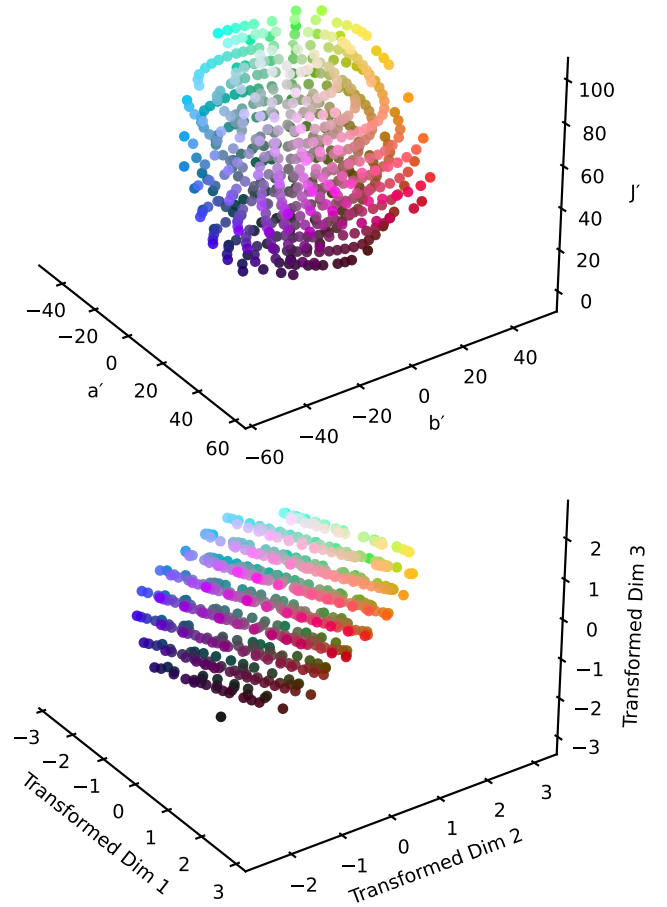
$$\hat{y} = \frac{1}{T} \left[ \sum_{t=1}^{\lfloor \rho T \rfloor} f_t(X \cdot R_t) + \sum_{t=\lfloor \rho T \rfloor + 1}^T f_t(X \cdot I_3) \right]$$

This combination is crucial for handling diverse vocabularies ranging from 24 to over 510 colour names in our data, as it allows the ensemble to capture both simple contrasts aligned with perceptual dimensions (e.g., light vs. dark along  $J'$ -axis) and more nuanced distinctions requiring oblique boundaries (e.g., *turquoise* vs. *teal* in the  $a'$ ,  $b'$  plane). The parameter  $\rho$  controls the exploration-exploitation trade-off: higher values prioritize geometric flexibility through transformation, while lower values maintain interpretability through axis-aligned cuts.

We employ Optuna’s Tree-structured Parzen Estimator [2] to optimise hyperparameters including number of estimators (200-400), maximum depth (10-25), minimum samples for splitting (2-8), leaf samples (1-4), maximum features (0.6-1.0), and transformation fraction  $\rho \in [0.3, 0.9]$ . Through systematic evaluation, we found that Random Forest’s deterministic split optimization synergises effectively with orthogonal transformation, as optimal thresholds can be better identified within transformed coordinate systems compared to random split selection (see Supplementary Materials Table S3).

The optimization objective maximizes our Colour Naming Fidelity (CNF) score, which combines two complementary metrics:

- (1) *Peak accuracy* measures whether the model correctly predicts the most frequently used colour name. For each test colour



**Figure 5: Visualisation of partial orthogonal transformation strategy in CAM16-UCS space. (top) Original coordinate system, (bottom) Transformed coordinate system.**

$c$ , we select the term  $W$  with highest predicted probability  $P(W|c)$ , using their overall frequency for tie-breaking. Peak accuracy is the proportion of correct predictions.

- (2) *Distributional similarity* quantified by the Bhattacharyya coefficient between predicted and observed naming patterns over colours, ranging from 0 (no overlap) to 1 (perfect match). It captures the full spectrum of naming variability (e.g., whether a blueish colour sample is called *blue* 60%, *navy* 30%, *dark blue* 10%).

High CNF requires both correctly identifying the most conventional term and representing naming variation. Model performance was evaluated using 3-fold cross-validation within each language dataset. For each model-language combination, we computed performance metrics (peak accuracy, distributional similarity, and combined CNF) by averaging results across the three folds, with hyperparameter optimization conducted to maximize CNF scores. Statistical comparisons between models were conducted using paired t-tests at the language level ( $n = 5$  languages), where each language

contributes one aggregated performance value per model, ensuring independent observations for statistical testing.

#### 4.4 Grid Construction for Indispensable Colour Name Identification

To identify the minimal set of indispensable colour names required to comprehensively name all colours within the perceptual gamut, we constructed a systematic test grid within CAM16-UCS colour space assuming sRGB viewing conditions. Lightness ( $J'$ ) was sampled at 9 levels corresponding to Munsell values:  $J' = 15, 24, 34, 45, 55, 65, 75, 83,$  and  $92$ , with chromaticity coordinates ( $a', b'$ ) sampled at 2-unit spacing across  $\pm 65$  unit ranges. All coordinates were validated for sRGB compatibility using round-trip conversion testing with  $\Delta E \leq 1$  tolerance. Here,  $\Delta E$  represents the Euclidean distance in CAM16-UCS space [52], and  $\Delta E = 1$  approximates a just-noticeable difference, yielding 7,072 perceptually uniform colour samples.

We identified indispensable colour vocabularies through a two-step filtering process. First, for each test colour sample of our grid, we selected the colour name with the highest predicted probability. Second, we retained only terms appearing as the top prediction at  $\geq 4$  grid locations, filtering out names with insufficient spatial extent. The  $\geq 4$ -grid-cells criterion is theoretically motivated by the principle that colour category extensions form convex sets in perceptually uniform three-dimensional colour spaces across all languages [25, 36]. Grid points where the top-predicted term was filtered out were reassigned to their next-highest probability term from the retained vocabulary. This methodology eliminates isolated colour names while identifying the minimal vocabulary needed to assign meaningful names to all perceptual colours.

For determining focal colours for each indispensable colour name within each language, we follow Boynton and Olson's [8] logic of identifying foci as the colour samples with maximum within-category agreement, selecting the single highest  $P(C|w)$  colour for each name. When multiple colours share the maximum agreement value, we compute the focal coordinates as their mean, avoiding the arbitrary thresholds of previous studies [32]. This approach facilitates the convex hull analysis of colour category volumes and determines their focal colours for practical applications.

## 5 Evaluation

Having established our methodological framework, our evaluation procedures address three key questions: (i) how effectively does SCF learn colour name distributions  $P(W|c)$  compared to baseline ensemble methods? (§5.1); (ii) how well does our indispensable colour name set represent the entire colour naming space? (§5.2); and, (iii) how do our indispensable sets of colour names compare with previous models? (§5.3).

### 5.1 Model Performance

We evaluated our SCF model across five languages, comparing standard ensemble methods against transformed-enhanced variants (Table 1). SCF consistently achieved the highest combined fidelity scores across all languages, with improvements ranging from 2.7% (Himba) to 4.4% (French and American English). Transformations improved both peak accuracy and distributional similarity

(Bhattacharyya coefficients) across all languages. American English improved from 1.259 to 1.315 CNF score with coverage expanding from 40 to 47 valid names (17.5% increase). French achieved the strongest distributional gains (0.758 to 0.776 Bhattacharyya) whilst maintaining competitive peak accuracy. Optimal transformation fractions clustered at  $\rho = 0.615$  (American English, French, Himba) and  $\rho = 0.837$  (British English, Greek), suggesting distinct geometric complexity patterns in colour naming systems requiring different degrees of coordinate transformation. Valid name coverage on systematic test grids improved consistently with orthogonal transformation. Himba achieved the highest coverage percentage (29.2% vs 25.0% baseline) despite having only 24 colour terms, whilst specialised vocabulary languages showed substantial absolute gains in valid name identification. Extra Trees generally outperformed Random Forest, particularly for French (1.538 vs. 1.498) and Greek (1.321 vs. 1.306). Himba achieved exceptionally high baseline performance (1.858-1.889) reflecting regular structure in basic colour term systems yet still benefited from transformation (1.901). These findings demonstrate that transformed ensemble methods provide systematic improvements across diverse linguistic colour naming systems, with benefits scaling appropriately to vocabulary complexity and geometric structure within each cultural-linguistic context.

Statistical analysis using paired  $t$ -tests demonstrated that our transformation-enhanced approach (SFC) significantly outperformed all baseline methods across five languages. The SFC method achieved improvements of 0.87% over Extra Trees ( $t(4) = 4.62, p = 0.010$ , Cohen's  $d = 2.07$ ), 2.65% over Random Forest ( $t(4) = 8.42, p = 0.001$ , Cohen's  $d = 3.76$ ), and 4.18% over the unrotated variant ( $t(4) = 19.28, p < 0.001$ , Cohen's  $d = 8.62$ ). All comparisons remained statistically significant after conservative Bonferroni correction for three pairwise tests (adjusted  $\alpha = 0.017$ ): SFC vs. Extra Trees ( $p = 0.030$ ), SFC vs. Random Forest ( $p = 0.003$ ), and SFC vs. Unrotated ( $p < 0.001$ ), with large effect sizes indicating substantial practical differences. SFC shows universal improvements across all tested languages, suggesting robust cross-cultural applicability. The magnitude of improvements ranged from moderate gains in structurally regular languages like Himba (2.7% improvement) to substantial enhancements in specialised vocabulary systems like French and American English (4.4% improvement each).

We conducted three validation analyses to confirm our methodological choices. Data normalisation substantially improved model performance (mean CNF increase: 11.3%,  $t(4) = 3.20, p = 0.033$ , Cohen's  $d = 1.43$ ) and enabled identification of larger indispensable vocabularies (Supplementary Materials Table S2). Threshold sensitivity analysis across values 1–6 confirmed our  $\geq 4$  grid-cell criterion as optimal: thresholds  $\geq 5$  showed 100% intersection with our reference (only reducing inventory), whilst threshold 1 captured 5.9–16.1% spurious names (Supplementary Materials Figure S2). Finally, systematic evaluation across 10 random seeds showed partial rotation consistently improved performance for all languages with CNF gains from +2.3% (Himba) to +5.7% (British English), with paired  $t$ -tests confirming significant improvements (all  $t(9) > 10.44, p < 0.001$ , Cohen's  $d > 3.30$ ) that survived Bonferroni correction (Supplementary Materials Figure S3).

**Table 1: Performance comparison of ensemble methods for cross-linguistic colour naming. Models include Extra Trees (ET), Random Forest (RF), unrotated ensemble (UR), and Spin Colour Forest (SCF). Evaluation metrics: cross-validated peak accuracy, distributional similarity via Bhattacharyya coefficient  $P(W|c)$  Similarity, combined Colour Naming Fidelity score (CNF), and valid colour names identified on systematic test grid (Names). Bold values indicate highest CNF score per language. AmEn and BrEn corresponds to American English and British English, respectively.**

Language	Model	Acc.	$P(W c)$ Similarity	CNF	Names
AmEn	ET	0.66	0.632	1.292	32
	RF	0.635	0.642	1.277	35
	UR	0.61	0.649	1.259	40
	SCF	0.652	0.663	1.315	47
BrEn	ET	0.68	0.619	1.299	34
	RF	0.645	0.628	1.273	36
	UR	0.648	0.602	1.25	29
	SCF	0.68	0.629	1.309	32
French	ET	0.793	0.745	1.538	18
	RF	0.748	0.749	1.498	21
	UR	0.725	0.758	1.483	25
	SCF	0.773	0.776	1.549	27
Greek	ET	0.677	0.644	1.321	33
	RF	0.652	0.653	1.304	33
	UR	0.615	0.648	1.263	34
	SCF	0.678	0.651	1.329	32
Himba	ET	0.952	0.937	1.889	6
	RF	0.937	0.922	1.858	6
	UR	0.922	0.929	1.851	6
	SCF	0.953	0.947	1.901	7

## 5.2 Indispensable Colour Names

The systematic grid quantification reveals fundamental differences in naming granularity across languages: American English requiring 47 indispensable terms, British English 32, French 27, Greek 32, and Himba 7 to comprehensively name the same perceptual colour space (Figure 6). While Basic Colour Terms cover larger areas with higher consistency, previous saliency-based studies left substantial portions of the colour space unlabelled [32]. Our approach identifies the complete set of empirically supported colour names necessary for comprehensive spatial coverage within each linguistic community.

**5.2.1 Completeness analysis.** Table 2 summarises the characteristics of our indispensable colour vocabularies across languages, including vocabulary size of indispensable sets, their coverage of the total unique colour names, and completeness in capturing human naming patterns across different accuracy thresholds. The indispensable vocabularies range from 7 terms (Himba) to 47 terms (American English), with corresponding coverage of the systematic test grid ranging from 5.49% (French) to 29.17% (Himba). For each colour chip, we ranked colour names by probability  $P(W|c)$

**Table 2: Vocabulary characteristics and completeness analysis of indispensable colour terms across languages, showing vocabulary size, coverage percentage, and top- $n$  accuracy. AmEn and BrEn corresponds to American English and British English, respectively.**

Language	Vocab. Size	Indisp. Sets	Cov. (%)	top- $n$ completeness			
				Top 1	Top 2	Top 3	Top 4
AmEn	492	47	9.55	91.17	97.7	99.7	100
BrEn	510	32	6.27	97.00	99.8	100	100
French	356	27	5.49	98.33	100	100	100
Greek	276	32	8.99	95.50	99.5	99.8	100
Himba	24	7	29.17	99.33	100	100	100

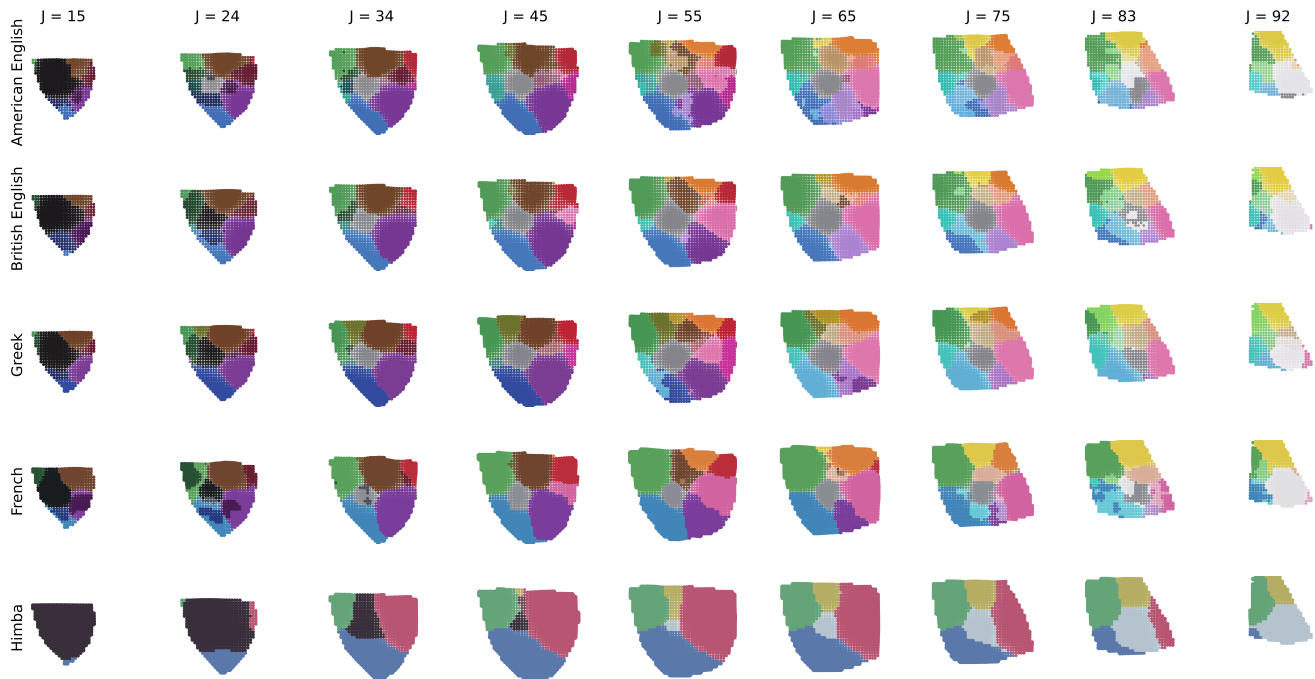
**Table 3: Performance comparison of Spin Colour Forest across languages with floor (average  $P(W|c)$  across all samples) and ceiling (bootstrapping with replacement,  $n = 1000$ ) bounds in terms of mean Bhattacharyya coefficient.**

Language	Min bound	Max bound	SCF
American English	0.26	0.90	0.66
British English	0.26	0.90	0.63
French	0.32	0.94	0.78
Greek	0.27	0.90	0.65
Himba	0.65	0.97	0.95

and calculated top- $n$  matches against the reduced vocabulary of indispensable names. Frequency-based tie-breaking was applied when  $P(W|c)$  values were equivalent, prioritizing more frequent colour names based on overall corpus frequency consistently improving performance. Near-complete coverage was achieved across all languages with top-1 accuracy exceeding 95% for all languages and with top-2 accuracy approaching or reaching 100% for most languages. All languages demonstrated 99.8%+ top-3 accuracy, indicating comprehensive capture of human colour naming behaviour.

**5.2.2 Theoretical Performance Bounds.** To contextualize our results within achievable limits, we established floor-ceiling bounds for Bhattacharyya coefficient performance. The floor represents simple averaging of conditional probabilities  $P(W|c)$  across all samples, whilst the ceiling was determined through bootstrap resampling with replacement (1,000 iterations) to estimate theoretical maximum performance given inherent variability in human responses (Table 3).

The theoretical ceiling varies systematically across languages due to vocabulary size effects (shown in Figure 4 and 6). Himba achieves a much higher ceiling (0.97) than European languages (0.90-0.94) because with only 7 colour terms, speakers show higher consensus because when multiple Himba speakers name the same colour, they are more likely to choose the same term from their limited vocabulary. In contrast, American English speakers choosing among 47 names show greater individual variation even when referring to the same perceptual colour. For example, a *mid-saturation*



**Figure 6: Cross-linguistic colour naming distributions across CAM16-UCS lightness levels. Each subplot displays colour name assignments for five languages (American English (N = 47 names), British English (N = 32), Greek (N = 32), French (N = 32) and Himba (N=7) across nine lightness slices  $J'$  = 15 to 92) spanning the range from dark to light colours under sRGB viewing conditions. Each point represents a systematic grid location in the  $a'b'$  chromatic plane, with point size indicating model prediction confidence and colours corresponding to the perceptual centroids of each colour category as determined by the Spin Colour Forest model.**

*blue* might be called *blue*, *light blue*, *sky blue*, or *navy blue* by different American English speakers (shown in Figure 6’s first row blue regions), producing lower distributional agreement. The same Himba speakers would consistently use *burou* (their single blueish term in last row of Figure 6), yielding higher theoretical maximum performance.

Our SCF method achieved 58-94% of theoretical maximum across languages, with performance scaling to vocabulary diversity: Himba (94% utilization) approached near-optimal distributional modelling in efficient vocabulary systems, whilst diverse languages (American English 63%, British English 58%, Greek 60%, French 74%) showed consistent mid-range utilization, directly reflecting the vocabulary richness-consensus trade-off described above. Crucially, all languages demonstrated substantial improvements over baseline (0.37-0.63-point gains above floor), confirming that transformation-enhanced regression captures genuine cross-cultural linguistic structure rather than methodological artifacts. The analysis validates both the practical significance of our improvements and the appropriate scaling of performance relative to theoretical constraints across diverse colour vocabulary systems.

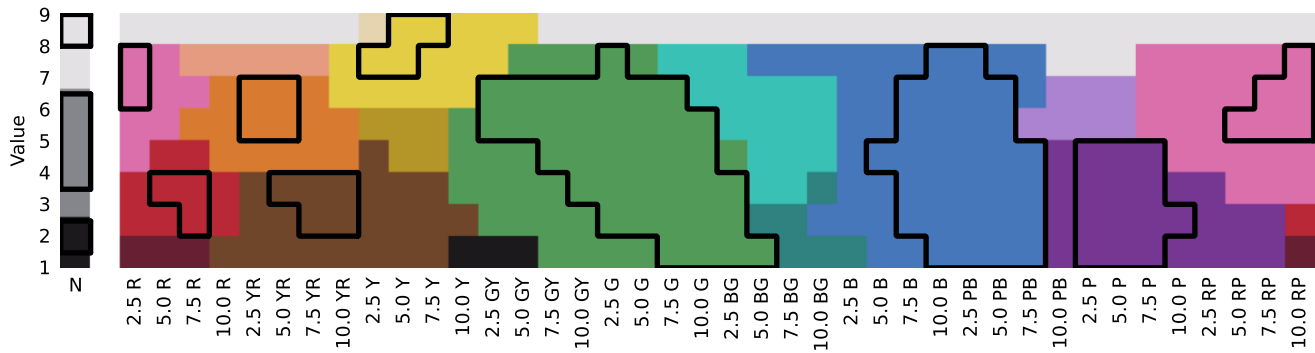
### 5.3 Comparisons Against Previous Colour Naming Models

To validate our approach against previous methods, we evaluated Spin Colour Forest on two established benchmarks. First, we tested

classification accuracy on the Sturges and Whitfield [78] Munsell array ( $n = 330$  chips). Second, we conducted perceptual comparisons using CAM16-UCS colour differences ( $\Delta E$ ) between our empirically derived foci and centroids against established probabilistic models for HCI [32, 41].

The benchmark dataset contains ground truth labels for 111 samples corresponding to the eleven Basic Colour Terms, with the remaining 219 samples unlabelled. Our model achieved 100% classification accuracy on the 111 BCT-labeled samples when excluding modifiers from our predictions to match the monolexemic ground truth, matching the performance of state-of-the-art colour naming models that are constrained to the eleven BCTs [67]. The Spin Colour Forest extended meaningful coverage to the previously unlabelled portions of the array without introducing classification errors, identifying seven additional empirically supported terms across the full array: *peach*, *cream*, *mustard*, *turquoise*, *teal*, *lilac*, and *maroon*. Figure 8 shows the performance comparison, with detailed error analysis and category counts provided in Table 4.

To compare our results against existing English-centric (Heer & Stone, 2012) and multilingual BCT-based (Kim *et al.* [41]) probabilistic models, we measured colour differences ( $\Delta E$ ) in CAM16-UCS assuming sRGB viewing conditions between our empirically derived focal and centroids colours and the output of their models (Supplementary Materials Table S4, Table S5 and Table S6)



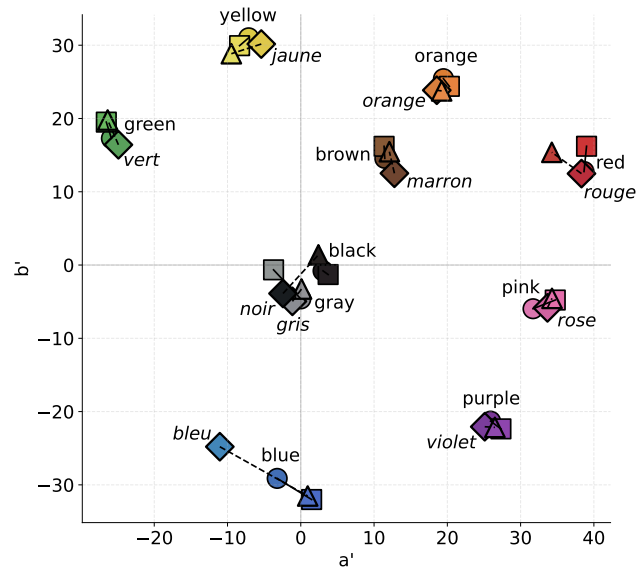
**Figure 7: Performance of Spin Colour Forest on Munsell array against Sturges & Whitfield (1995) ground truth data. Sturges and Whitfield’s mapping of BCTs in British English are drawn with black boxes. Colour regions show predicted colour name assignments, with each category displayed using the coordinates of the foci in sRGB.**

**Table 4: Performance comparison of colour naming models on Sturges & Whitfield (1995) Munsell array benchmark. Models were evaluated on 111 Basic Colour Term-labelled samples from the 320-chip array. Coincidences indicate correct classification, Errors show misclassifications, % represents error rate, and Number of names indicates the total colour vocabulary identified across the full array on the surface of the colour solid.**

Models	Coincidences	Errors	%	Number of Names
LGM	92	19	17	11
MES	107	4	4	11
TSM	108	3	3	11
SFKM	111	0	0	11
TSEM	111	0	0	11
PLSA	109	2	2	11
NICE	111	0	0	11
SCF	111	0	0	18

The comparison of our 47 American English focal colours against Heer and Stone’s colour naming thesaurus [32] revealed large perceptual differences ( $M\Delta E = 12.08, SD = 5.76$ ). Yet, our focal colour computation differs from Heer & Stone’s approach. When we compute foci by using their top-4 averaging method for a fairer comparison, our foci showed better but still moderate agreement ( $M\Delta E = 10.60, SD = 4.82$ ). Focused on the 11 BCTs the discrepancies were lower with ( $M\Delta E = 8.93, SD = 3.20$ ) with residual differences potentially attributable to adaptation variations collected against different background colours (white background in XKCD data) vs. neutral grey background ours.

The comparison of our 10 BCT centroids against Kim *et al.*’s [41] 10-cluster multilingual model revealed language-dependent perceptual alignment in the available common languages (Figure 8) American English centroids showed good agreement ( $M\Delta E = 5.66, SD = 1.99$ ). For comparison, differences between our own American and British English BCT centroids were significantly smaller ( $M\Delta E = 1.55, SD = 0.84; t(18) = 6.22, p < 0.001$ ), approaching the just



**Figure 8: Comparison of Basic Colour Term centroids in CAM16-UCS space ( $a'$ ,  $b'$  coordinates). American English (circles) and French (diamonds) centroids from the current study are compared against Kim *et al.*’s [41] English (squares) and French (triangles) multilingual model. Greater spatial proximity between markers indicates stronger perceptual alignment of colour categories across languages.**

noticeable difference threshold of  $\Delta E = 1$  [1]. French centroids exhibited larger deviations ( $M\Delta E = 7.24, SD = 3.21$ ), indicating greater perceptual misalignment.

These comparisons reveal three key findings: (1) methodological choices in focal colour calculation substantially impact results, with our maximum probability approach yielding more precise category representatives than averaging methods; (2) BCT-constrained models achieve better perceptual agreement than models attempting

full vocabulary coverage yet capture only a small fraction of indispensable colour names; and (3) experimental design critically affects data quality, with our online methodology using grey background yielding substantially better within-language consistency than parallel web surveys.

## 6 Translation Benchmark

Our data and models establish empirical baselines for how colour names map to perceptual space within each language. This grounding enables a critical question: do current large language models translate colour terms in ways that preserve perceptual meaning across cultures, or do they simply map words without respecting the categorical structure we have documented? Our translation benchmark addresses this question in two stages. First, we establish information-theoretic baselines for translation pairs (§6.1) and validate them with human participants (§6.2). Second, we evaluate LLMs against these human-validated baselines on two dimensions: lexical (§6.4) and perceptual accuracy of colour naming translations (§6.5) for BCTs and non-BCTs (§6.6). This visual-language evaluation demonstrates why empirically-grounded benchmarks are essential for assessing cross-linguistic colour translation.

### 6.1 Information-Theoretic Baseline Translation

To establish theoretical constraints for cross-linguistic colour translation, we conducted an information-theoretic analysis using Jensen-Shannon divergence (JSD) to quantify semantic closeness between colour names. Semantic closeness measures whether two colour terms—either within a language or across languages—refer to similar regions of perceptual colour space. JSD operates on the probability distributions  $P(W|c)$  of where speakers use each name across our stimulus set. For example, if English *turquoise* and Greek *τιρκουάζ* are both used primarily for the same blue-green region, they have high semantic closeness (low JSD); if they systematically refer to different hues, their JSD increases. Unlike simple lexical matching, JSD captures perceptual similarity: terms can be semantically close even with different labels, or semantically distant despite being translation equivalents in a dictionary. We computed JSD similarity matrices from empirical colour naming data across five languages, enabling systematic comparison across 10 language pairs.

Instead of relying on arbitrary cut-offs, our analysis employed empirically derived thresholds, specifically, the 75th, 90th percentiles, and a moderate mean + SD of all JSD similarity values, to define a robust baseline for translation connectivity. Analysis at the moderate threshold ( $M = 0.281$ ) revealed mean connection density of 0.136 ( $SD=0.063$ ) with source coverage of 0.954 ( $SD=0.064$ ). Translation entropy averaged 9.2 bits ( $SD=1.2$ ), indicating moderate information complexity in cross-linguistic colour mappings.

Languages with small lexicons showed systematically higher connectivity: Himba language pairs demonstrated 2.5-times higher connection density than European language pairs. This pattern is robust across all empirically derived thresholds, with a low coefficient of variation of 0.034 reflecting a genuine linguistic structure. Translation from high-diversity to low-diversity languages loses information, while the reverse introduces ambiguity (Figure 5). For example, American English distinguishes nine blueish terms (*light*

*blue, sky blue, blue, navy blue, teal, dark teal, cyan, turquoise, aqua*), all mapping to single Himba term *burou*—unavoidable many-to-one information loss. Conversely, translating Himba *burou* into English is ambiguous without perceptual context—which of the nine English terms? This asymmetry, visualised in Figure 6, explains why compact vocabularies create denser, more consolidated semantic spaces with stronger but less granular cross-linguistic mappings. Translation between languages with different categorical structures creates additional complexity. Greek distinguishes *γαλάζιο/galazio, light blue*) and *μπλε/ble, dark blue*) as basic-level terms, while English uses basic *blue* with optional modifiers. Translating English *blue* to Greek is ambiguous (*γαλάζιο* or *μπλε*?), whilst translating Greek *γαλάζιο* as *light blue* loses basic-term status.

Our chain analysis across a multi-step translation, from higher to lower diverse colour lexicons, pathway revealed a cumulative loss of translatability. As shown in Figure 9, a translation chain from American English to Himba via three intermediary languages resulted in an end-to-end coverage of 0.834 at the moderate threshold, with a cumulative loss of 0.166. This demonstrates that while cross-cultural colour translation remains largely viable, it faces theoretical constraints where a portion of the source information is inherently lost or diluted with each translation hop.

The establishment of this information-theoretic baseline provides an empirically grounded context for evaluating the performance of machine translation systems, such as LLMs. Our results show that linguistic structure fundamentally shapes translation connectivity, and any effective translation system must navigate these inherent patterns and constraints.

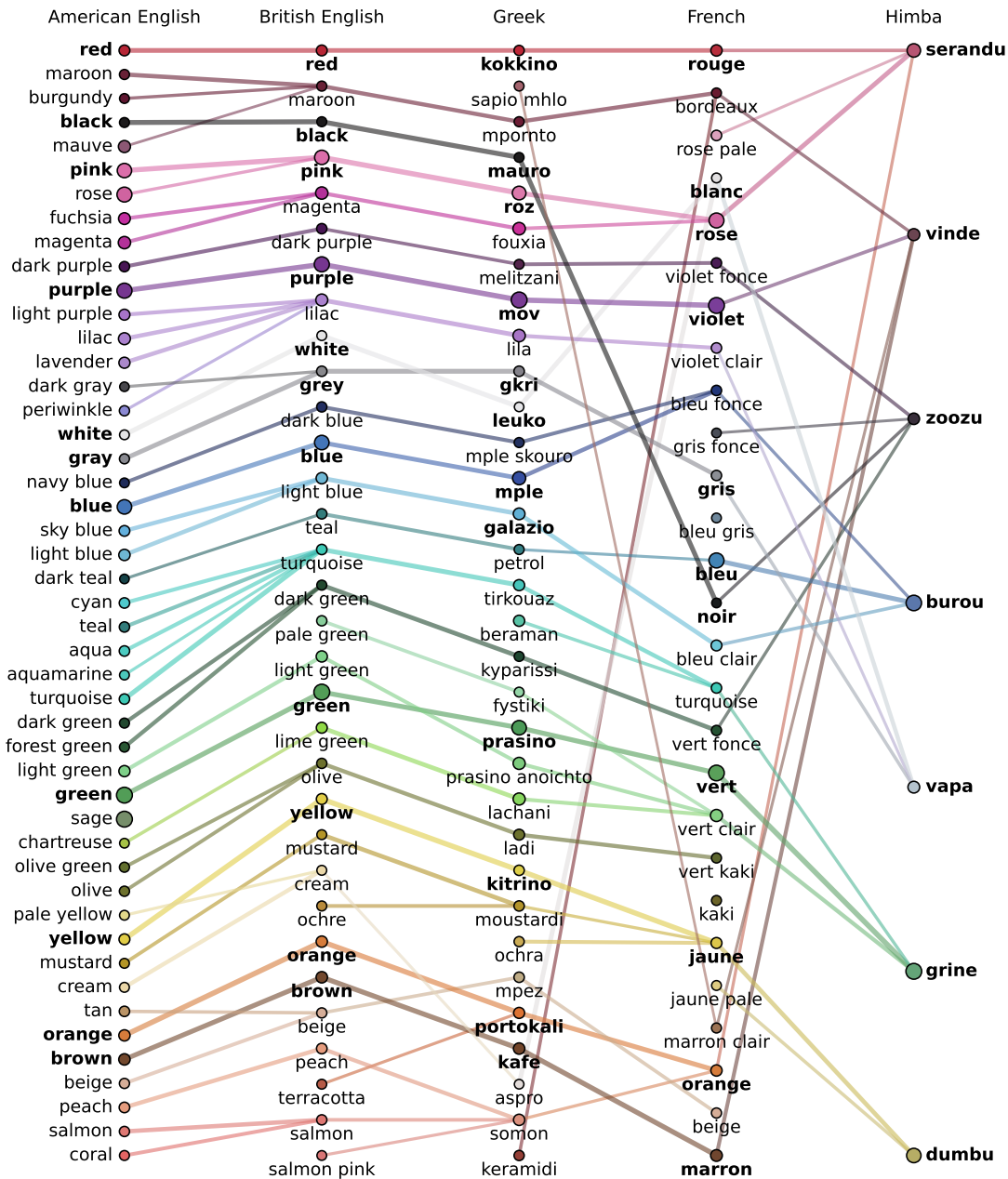
Within this information-theoretic framework, analysis of 370 optimal translation pairs reveals fundamental differences between BCTs and non-BCTs. BCT-to-BCT translations achieve significantly higher semantic similarity than non-BCT translations ( $0.5821 \pm 0.1283$  vs  $0.4728 \pm 0.0917$  JSD;  $U = 2,847$ ,  $p < 0.001$ , Cohen's  $d = 0.95$ ). This 0.1093 JSD advantage represents a large effect size, confirming that BCTs exhibit greater cross-linguistic consistency than non-BCTs [6, 41].

### 6.2 Human Validation of Baseline Translations

To validate the quality of our information-based translations used as ground truth in the translation benchmark (§6), we conducted a human evaluation study with bilingual speakers.

**6.2.1 Participants.** We recruited 19 bilingual participants across three language pairs: American English ↔ British English ( $n=7$ ,  $M$  age=27.3,  $SD=3.9$ , range: 23-35), British English ↔ French ( $n=5$ ,  $M$  age=35.2,  $SD=18.6$ , range: 24-72), and British English ↔ Greek ( $n=7$ ,  $M$  age=50.0,  $SD=11.7$ , range: 42-78). For the British English ↔ French pair, French proficiency varied (native: 3, intermediate: 1, basic: 1) while British English proficiency was consistently high (native: 2, Advanced C1-C2: 3). For the British English ↔ Greek pair, all evaluators demonstrated native Greek proficiency with British English proficiency at native ( $n=2$ ) or Advanced C1-C2 ( $n=5$ ) levels. Ethical approval was granted by Northeastern University London (No. 0001, 16th May 2023).

**6.2.2 Procedure.** Participants evaluated baseline translations via Microsoft Forms, rating each translation pair (e.g., *beige* → *μπεζ*)

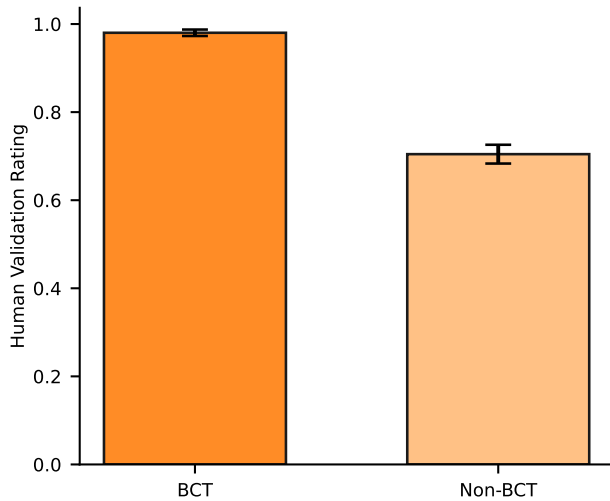


**Figure 9: Translation chain connectivity visualization across five languages. Node size reflects convex hull volume of corresponding colour category, node colour indicates empirical colour foci, and edge thickness represents JSD similarity strength above a moderate threshold of JSD = 0.281.**

on a 5-point scale: 1.00 (Perfect match), 0.75 (Very good match), 0.50 (Acceptable match), 0.25 (Poor match), and 0.00 (No match). The evaluation covered 202 unique translation pairs across all language combinations.

**6.2.3 Results.** Baseline translations received generally high ratings ( $M=0.795, SE=0.017$ ), confirming their suitability as ground truth. Translation quality did not vary significantly across source languages ( $F(3, 198) = 0.865, p = 0.460$ ). Critical differences emerged

between BCT and non-BCT translations (Figure 10). BCTs achieved significantly higher human ratings ( $M=0.980, SE=0.007$ ) compared to non-BCTs ( $M=0.705, SE=0.021; t(200) = 8.962, p < .001, Cohen's d = 1.339$ ) while non-BCT translations still received high absolute ratings. This pattern remained consistent across all source languages: American English ( $t(45) = 3.564, p < .001, d = 1.228$ ), British English ( $t(94) = 5.875, p < .001, d = 1.262$ ), French ( $t(25) = 5.107, p < .001, d = 2.000$ ), and Greek ( $t(30) = 3.622, p = 0.001, d = 1.322$ ).



**Figure 10: Human validation ratings for BCT versus non-BCT translations across language pairs. Error bars represent standard error.**

This human validation confirms that: (1) our baseline translations represent empirically grounded cross-linguistic mappings suitable for benchmarking; (2) BCTs demonstrate significantly higher translatability than non-BCTs, validating our information-theoretic analysis.

### 6.3 LLM case study: Evaluation of Claude’s Colour Naming Translations

We evaluated the colour translation capabilities of Claude Sonnet 4, Anthropic’s latest model as a case study. Data was collected between 01/08/2025 and 10/09/2025, ensuring all translations were performed on the same model version. We used a standardized, direct prompt to translate the 138 indispensable colour terms from our cross-linguistic study. The benchmark included 552 translations across 4 source languages (American English, British English, Greek, French) and 5 target languages (these four plus Himba). We notably excluded Himba as a source language because initial tests revealed that Claude consistently failed to provide translations when prompted with Himba terms. This limitation highlights a gap in the model’s knowledge for less-resourced languages and underscoring the value of our empirically grounded baseline for assessing cross-cultural AI performance.

The evaluation protocol employed standardized prompts administered through separate chat sessions for each language pair to prevent cross-contamination of responses:

- Lexical translation prompt: “Translate [colour name] to [target language]”
- Perceptual accuracy prompt: “Give best RGB examples for each colour word in [list]”

Each chat session followed a consistent two-stage protocol: first, obtain lexical translations for all target colour terms, then request RGB colour specifications for the translated terms. This sequential

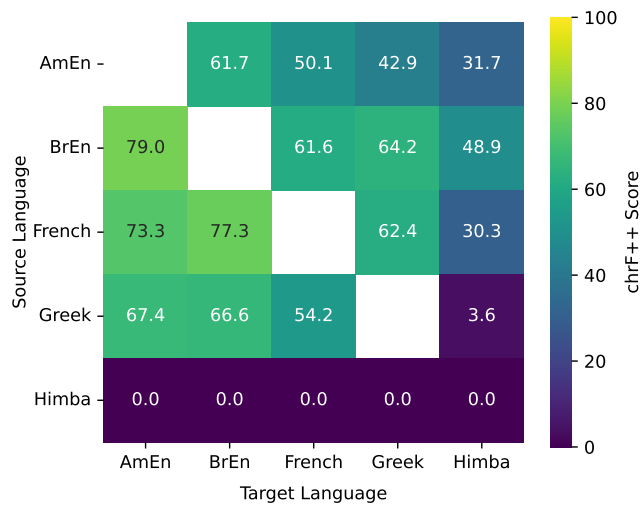
approach within single chat sessions ensured that Claude’s RGB outputs corresponded directly to its own lexical translations, enabling precise evaluation of the lexical-perceptual coupling in the model’s colour translation performance. This methodology design allows for systematic comparison of both lexical accuracy, via chrF++ character-level similarity metrics [73], and perceptual accuracy, via CAM16-UCS  $\Delta E$  colour difference measurements [90] against our baselines, while maintaining experimental control across the diverse linguistic and cultural contexts represented in our benchmark.

### 6.4 Translation Lexical Accuracy

We evaluated Claude’s lexical translation performance using the chrF++ metric [70], which quantifies character-level overlap between Claude’s output and our ground truth translations shown in Figure 11. The overall mean chrF++ score was 53.60 ( $SE=1.89$ ), suggesting a moderate degree of character-level similarity with significant differences between both target languages ( $H=73.55$ ,  $p<0.001$ ) and source languages ( $H=13.66$ ,  $p=0.003$ ). Lexical performance followed a clear hierarchy: English variants performed best (American English:  $M=73.23$ ,  $SE=4.15$ ; British English:  $M=67.17$ ,  $SE=4.15$ ), European languages showed moderate accuracy (French:  $M=54.59$ ,  $SE=4.23$ ; Greek:  $M=54.30$ ,  $SE=4.18$ ), whilst Himba translations were poorest ( $M=28.90$ ,  $SE=3.24$ ). The best lexical performance was British English to American English (chrF++=78.99), whilst Greek to Himba showed near-complete failure (chrF++=3.57). Lexical translation accuracy revealed hierarchical performance differences (chrF++ range: 3.57-78.99) with English variants outperforming Greek and French, which in turn vastly exceeded Himba translations, demonstrating systematic biases toward linguistically similar and well-represented languages in training data.

To contextualize LLM translation performance, we examined the relationship between language resource availability and translation quality. Using Wikipedia article counts as a proxy for training data availability (American/British English: 7.1M articles; French: 2.6M; Greek: 235K; Himba: 0), we found a strong positive correlation between Wikipedia article count and lexical translation accuracy (Spearman’s  $\rho = 0.97$ ,  $p = 0.005$ ). This relationship demonstrates that Claude’s lexical translation performance follows a clear hierarchy aligned with training data availability: well-resourced languages achieve substantially higher chrF++ scores, while under-resourced languages like Himba show markedly poorer lexical accuracy.

Character-level similarity metrics like chrF++ are sensitive to string-level variations such as hyphenation (*navy blue* vs. *navy-blue*) and character substitutions (*grey* vs. *gray*). Qualitative analysis of bidirectional American English↔British English translation (79 terms) revealed that 39.2% of translations received low chrF++ scores despite semantic equivalence (Supplementary Materials Table S7 and Table S8). Failures clustered into three categories: synonym mismatches (48%; e.g., *lavender* vs. *lilac*, chrF++ 7.11), modifier differences (42%; e.g., *navy blue* vs. *dark blue*, chrF++ 26.09), and colour family grouping (23%; *aqua/cyan/teal* mapped to *turquoise*, chrF++ 0–9.51) because of their different number of indispensable colour names. Complete lexical mismatches (chrF++ = 0) occurred for synonymous pairs like *cyan/turquoise* and *rose/pink*.



**Figure 11: Lexical translation accuracy for Claude across source-target language pairs. The heatmap displays  $M$  chrF++ scores with colour intensity indicating translation quality (yellow = better, purple = worse). Best results are for English variants and poorest for Himba translations.**

This demonstrates that chrF++ penalizes lexical variation in specialised colour vocabulary where cross-linguistic correspondences involve synonym selection rather than direct translation.

However, human validation revealed that chrF++ performs well overall as a translation quality metric. Claude Sonnet 4’s lexical translation accuracy measured via chrF++ showed strong positive correlation with human validation ratings of our baseline translations ( $r = 0.826$ ,  $p < .001$ ). Higher chrF++ scores reliably indicate translations that bilingual speakers rate more favourably. The strong overall correlation demonstrates that while chrF++ has systematic limitations for synonym variation, it serves as a meaningful proxy for human-judged translation quality when interpreted appropriately.

These findings justify chrF++ as our lexical evaluation metric when complemented with qualitative interpretation of low-scoring cases. Claude demonstrates strong lexical translation competence, particularly for well-resourced language pairs, achieving high human-validated accuracy. This lexical competence, however, does not guarantee perceptual accuracy, as we demonstrate next.

## 6.5 Perceptual Accuracy

We evaluated whether Claude preserves perceptual meaning when translating colour names across languages. For each translation, Claude provided both the target language term and its best RGB colour example. We compared these RGB values against empirically derived colour foci to measure perceptual accuracy using CAM16-UCS  $\Delta E$  colour difference, assuming the Internet-standard sRGB viewing conditions, where values exceeding 10  $\Delta E$  units represent very large perceptual differences [1, 52].

Analysis of 138 colour name translations across five target languages revealed uniformly poor perceptual accuracy (Figure 12). Target language significantly affected performance (Kruskal-Wallis  $H=45.98$ ,  $p<0.001$ ), with Western languages showing very large colour differences (American English:  $M=16.86$ ,  $SE=1.03$ ; British English:  $M=19.27$ ,  $SE=1.14$ ; French:  $M=16.92$ ,  $SE=0.99$  and Greek:  $M=18.56$ ,  $SE=0.91$ ). Himba translations showed substantially worse alignment ( $M=28.13$ ,  $SE=1.45$ ), significantly exceeding all European languages in pairwise comparisons with Bonferroni correction (all  $p<0.001$ ).

Notably, while Claude claimed insufficient knowledge to provide RGB examples when directly prompted for Himba colour names as a source language, it readily generated RGB values for Himba terms when presented within a translation context from other languages. Further examination revealed that Claude generated mostly identical RGB values for corresponding colour names regardless of target language. For example, Claude provided the same RGB coordinates (0,0,255) for *blue* whether translating from Greek  $\mu\pi\lambda\epsilon$ , French *bleu* or American English *blue*. Yet, our empirical data demonstrate that colour foci vary across languages—the best example of *blue* differs between Greek, French, English and Himba (*burou*) speakers.

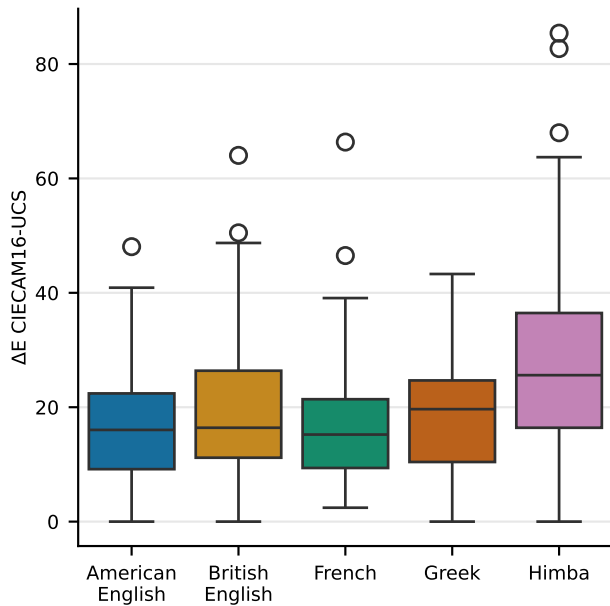
The varying perceptual errors across target languages thus mostly reflect the degree to which each language’s empirical colour naming system deviates from Claude’s assumed universals. While Himba shows the worst performance ( $M\Delta E = 28.13$ ), even well-resourced European languages exhibit very large perceptual errors ( $M\Delta E \approx 18$ ). This demonstrates that Claude’s universal mapping strategy fails across all languages—not just low-resource ones—suggesting that the problem is not merely a matter of insufficient lexical training data but a fundamental failure to ground colour terms in language-specific perceptual structures.

## 6.6 Perceptual and Lexical Evaluation for BCTs and nonBCTs

We examine the relationship between perceptual and lexical accuracy in colour translation, focusing on how Claude Sonnet 4 handles BCTs versus less frequent non-BCTs. Because BCTs are more common in multilingual training data, one might expect both lexical and perceptual accuracy to be higher for BCTs. By jointly analysing character-level similarity (chrF++) and perceptual colour differences ( $\Delta E$ ) across 552 name translation pairs from four source languages to five target languages, we test whether these two dimensions align or reveal systematic dissociations.

Claude Sonnet 4’s translation performance reveals discrepancies between lexical and perceptual accuracy when handling BCTs versus non-BCTs. Figure 13 shows that Claude achieves substantially higher lexical accuracy for BCT translations ( $75.94 \pm 39.96$  vs  $40.91 \pm 41.65$  chrF++;  $U = 49,269$ ,  $p < 0.001$ ), representing a 35-point advantage with a large effect size (Cohen’s  $d = 0.85$ ). This demonstrates that Claude’s training has effectively captured the lexical regularity of BCT translations across languages.

For perceptual translation accuracy, BCT translations showed significantly better alignment with empirical colour foci ( $18.07 \pm 12.32$  vs  $21.85 \pm 13.45$   $\Delta E$ ;  $U = 29,336$ ,  $p = 0.001$ ), with this 3.8  $\Delta E$  improvement representing a small effect size (Cohen’s  $d = 0.29$ ). Lower  $\Delta E$  values indicate better perceptual accuracy. However, both

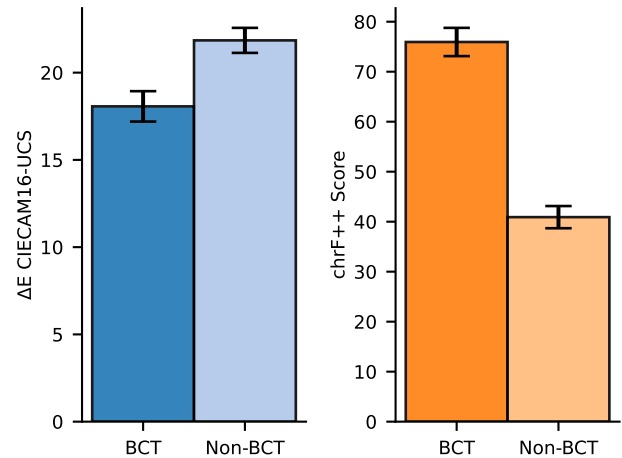


**Figure 12: Perceptual accuracy of colour naming mappings by target language measured as CAM16-UCS perceptual distances ( $\Delta E$ ) between Claude’s RGB colour outputs and empirical colour foci. Lower  $\Delta E$  values indicate better perceptual alignment between translated colour exemplars and target language colour categories.**

BCT and non-BCT translations exhibited large perceptual errors (both  $M > 18 \Delta E$ ), well exceeding the 10  $\Delta E$  threshold for very large perceptual differences. The modest perceptual advantage for BCTs contrasts sharply with the 35-point lexical advantage, revealing that Claude’s strong lexical performance for BCTs does not translate into proportionally strong perceptual accuracy.

Correlation analysis reveals a weak association between lexical and perceptual dimensions. Claude’s lexical accuracy (chrF++, higher is better) shows only a weak negative relationship with perceptual distance ( $\Delta E$ , lower is better) ( $r = -0.3123$ ,  $p < 0.001$ ,  $R^2 = 0.0975$ ), explaining less than 10% of variance. This weak predictive power is similar for both BCTs ( $r = -0.3367$ ,  $R^2 = 0.113$ ) and non-BCTs ( $r = -0.2572$ ,  $R^2 = 0.066$ ), indicating that lexical competence predicts perceptual accuracy equally poorly for both term types.

Despite BCTs’ substantially higher lexical accuracy, their lexical advantage does not create a stronger connection between producing correct words and representing correct colours. The weak correlations demonstrate that high lexical accuracy does not reliably predict perceptual accuracy, confirming the dissociation between word-level translation competence and grounded colour understanding.



**Figure 13: Translation performance for perceptual (left) and lexical accuracy (right) across BCTs and non-BCTs. Error bars represent standard error. Perceptual accuracy is measured as  $\Delta E$  CAM16-UCS colour difference from empirical target language colour foci, where lower values indicate better alignment. Lexical accuracy is measured as chrF++ character-level similarity scores, where higher values indicate better performance. Claude shows a large lexical advantage for BCTs (35.0 chrF++ points) but a small perceptual advantage (3.8  $\Delta E$  units), demonstrating dissociation between lexical and perceptual translation quality.**

## 7 Discussion

The empirical foundation of this work represents both a substantial expansion in scale and critical methodological advances over previous cross-cultural colour naming studies. Unlike earlier research examining only saturated surface colours ( $n=330$ ) with small observer groups restricted to monolexic terms [6, 8, 78] or with colours viewed against white background that compresses dark colours [32, 56], we sampled the full colour solid ( $n = 600$ ) using a large number of participants providing unconstrained naming responses against neutral grey background across five languages. This experimental advancement enables us to examine fundamental questions about cross-linguistic colour categorisation, computational colour naming modelling, and machine translation capabilities for specialised perceptual domains, providing practical guidance for human-computer interaction practitioners working across cultural contexts.

### 7.1 Beyond Basic Colour Terms: The Case for Indispensable Vocabularies

The identification of indispensable colour vocabularies, ranging from 47 in American English, 32 in British English, 32 in Greek, 27 in French to 7 in Himba, aligns with previous research demonstrating that speakers of highly diverse lexicons use ranging from 25 [50] to 50 [12, 22, 31] colour names in their native language and that the Himba have 5-7 proper colour terms [59, 72]. These vocabularies represent empirically validated minimal sets for comprehensive

colour reference within each linguistic community, departing from BCT frameworks [6] that constrain computational models to universal eleven-category systems [4, 14, 53, 67, 82, 83] and aligning with accounts that emphasise communicative need in the development of colour naming systems [26, 29, 59, 62, 72, 77].

Our  $\geq 4$ -grid-cell threshold for identifying these vocabularies is theoretically motivated by the idea that colour categories form convex sets in perceptually uniform colour spaces [25, 36]. Threshold sensitivity analysis confirmed that vocabulary sets stabilize at  $\geq 4$ : thresholds 5 and 6 showed 100% intersection with our reference threshold, while lower thresholds progressively captured spurious peripheral terms (16.1% additional names at threshold 1 for American English). This stability demonstrates that our indispensable vocabularies reflect genuine linguistic structure rather than methodological artifacts.

While BCTs show higher cross-linguistic consistency (JSD similarity 0.582 vs 0.473 for non-BCTs) and achieve better translation accuracy, they capture only 23–41% of the indispensable vocabularies we identified across the four written languages. For British English, the 11 BCTs leave substantial portions of colour space inadequately described, with 21 additional indispensable terms required for comprehensive coverage. This reveals a fundamental trade-off: BCTs achieve cross-linguistic alignment by marking major colour categories that many languages share, but this universality comes at the cost of coarse granularity. Non-BCTs enable finer perceptual discrimination—for example, introducing *turquoise* between *blue* and *green*, or distinguishing *lilac* from *purple* [63]—at the expense of reduced cross-linguistic correspondence.

## 7.2 Geometric Transformations Capture Colour Category Structure

Our Spin Colour Forest model demonstrates that geometric transformation of perceptual colour space can capture the complex boundaries of human colour naming across diverse linguistic communities. The systematic improvements across all five languages (2.7% to 4.4%) confirm that colour naming boundaries often exhibit simpler structure in transformed coordinate systems than in the standard CAM16-UCS space.

The mechanism underlying these improvements relates directly to the structure of colour spaces. While CAM16-UCS optimizes perceptual uniformity for measuring Euclidean distances, its orthogonal chromatic dimensions ( $a'$ ,  $b'$ ) do not align with unique hue axes [87, 88] or linguistic category structures [58]. The  $a'$  axis does not align with red-green, nor  $b'$  with yellow-blue. This creates systematic misalignment: Greek  $\gamma\lambda\acute{\alpha}\zeta\iota\omicron$  (light blue) and  $\mu\pi\lambda\epsilon$  (dark blue) are separated by a boundary cutting diagonally through both the lightness ( $J'$ ) and blue-yellow-related ( $b'$ ) dimensions [17], requiring multiple oblique splits in standard coordinates. This misalignment is even more striking in Himba, where category boundaries span regions that Western languages typically partition differently. For instance, *serandu* (reddish) and *grine* (greenish) overlap in desaturated regions [59]. In standard coordinates, a decision tree must learn multiple complex splits to approximate these boundaries. However, in an appropriately rotated coordinate system, these same boundaries can be captured with cleaner axis-aligned splits, reducing model complexity whilst improving generalisation.

Our fractional transformation strategy ( $\rho \approx 0.6$ – $0.8$  across languages) proves essential for balancing geometric flexibility with interpretability. Pure rotation would maximise boundary-finding capability but sacrifice the interpretable alignment between colour space dimensions and perceptual attributes. By training a proportion of trees in transformed spaces and the remainder in standard coordinates, our ensemble captures both simple categories aligned with perceptual dimensions (e.g., “dark” varying along  $J'$ ) and complex categories requiring oblique boundaries (e.g., *turquoise* spanning the ( $a'$ – $b'$ ) plane with lightness constraints). Rotation-enhanced ensemble methods have demonstrated superior performance across diverse continuous feature classification problems [3], supporting the generalisability of this approach to broader HCI applications.

## 7.3 Information-Theoretic Constraints on Colour Translation

Our translation benchmark reveals fundamental asymmetries in cross-linguistic colour communication that cannot be overcome through better algorithms or larger training sets. These constraints arise from the categorical structure of colour naming itself: when languages partition colour space differently, some information is necessarily lost or ambiguous in translation. The directional asymmetry is most evident when translating from a vocabulary-rich language to Himba—for example, American English’s nine blueish terms (*light blue*, *sky blue*, *blue*, *navy blue*, *teal*, *dark teal*, *cyan*, *turquoise*, *aqua*) all collapse a single term in Himba, *burou*. This compression is unavoidable given the vocabulary difference: no translation system can preserve distinctions that the target language does not lexicalise [92]. Conversely, translating Himba *burou* into English produces one-to-many ambiguity—which of the nine English terms is appropriate depends on perceptual context unavailable to text-only translation systems. This information-theoretic analysis explains the systematic pattern in Figure 9’s translation chain connectivity: vocabulary compression (high→low diversity) shows strong but semantically broad mappings, whilst vocabulary expansion (low→high diversity) shows weak, dispersed connections. The human validation confirms that our baseline translations represent empirically grounded cross linguistic mappings suitable for benchmarking machine translation systems, providing ground truth that reflects both perceptual alignment and human judgment of translation adequacy.

## 7.4 The Lexical-Perceptual Disconnect in LLM Translation

Our systematic translation benchmark evaluation reveals critical limitations in current large language model approaches to cross-cultural colour communication. Analysis of Claude Sonnet 4 across 552 translation instances demonstrates performance hierarchies that undermine conversational interface effectiveness. Translation accuracy follows a clear resource-based pattern [33, 85]: European language pairs achieve moderate performance (British→American English: 78.99 chrF++; intra-European: 54–67 chrF++) while cross-cultural translations fail systematically (all European→Himba: 28.90 chrF++; Greek→Himba: 3.57 chrF++). This hierarchy reveals that, even within languages, Claude attains competence mainly for the

well-resourced BCTs while systematically failing on the less frequent non-BCTs. The limitation becomes particularly severe when considering perceptual accuracy, where large colour differences exceeding on average 17  $\Delta E$  units [1] indicate substantial cross-cultural misalignment that compromises meaningful colour communication [52, 62]. The 34.2-point lexical advantage for BCTs contrasts sharply with minimal perceptual improvement (2.8 points), confirming that linguistic training captures word associations without cultural colour understanding. The benchmark establishes that current LLMs create misleading user experiences: surface-level linguistic competence of frequent words within similar language families masks systematic failures in cross-cultural colour understanding, resulting in confident-appearing translations that deliver low perceptual fidelity precisely where inclusive HCI applications require robust cross-cultural communication capabilities.

Claude's perceptual translation failures reveal a critical limitation: mapping all language-specific names onto a mostly invariant set of universal focal colours conflicts with documented cross-linguistic variation. The varying perceptual errors across languages reflect deviations from Claude's Western-centric prototypes, most dramatically in Himba ( $M\Delta E=28.13$ ) but also substantial in well-resourced European languages ( $M\Delta E\approx 18$ ). This demonstrates that the problem is not merely insufficient training data but a fundamental failure to ground colour terms in language-specific perceptual structures. Lexical accuracy (producing the correct word) does not entail perceptual accuracy (understanding what colour that word denotes in its cultural context), with significant implications for cross-cultural communication systems relying on LLM translations.

## 7.5 Implications for AI-powered HCI

Our findings challenge assumptions about general-purpose language model effectiveness for specialised cross-cultural applications. The performance hierarchy reveals systematic biases toward linguistically similar, high-resource languages, creating immediate implications for global interface design where colour-dependent applications risk systematic misinterpretation in non-Western cultural contexts.

Three critical HCI domains are particularly affected. Design software (e.g., Figma, Adobe Creative Suite) increasingly incorporates AI-powered colour suggestions and natural language colour selection, but our LLM evaluation shows that such suggestions could potentially favour English speakers while delivering perceptually inaccurate results for other languages. E-commerce platforms rely on colour name translation for product listings, yet our benchmark reveals that translations achieve lexical accuracy whilst delivering wrong colours to customers. Accessibility tools for diverse visual abilities (colour vision deficiency, age-related changes in colour perception) depend on accurate colour naming and translation, but current systems' Western-centric biases exclude non-English speakers from effective assistive technology.

Our evaluation extends Heer and Stone's [32] vision of colour naming interfaces while revealing fundamental cross-cultural limitations. While their English-centric applications—including name-based pixel selection, colour dictionaries, and palette evaluation tools—demonstrated effective natural language interaction, current LLM translation approaches cannot extend these capabilities

across cultural boundaries. The weak correlation between lexical and perceptual accuracy ( $r = 0.31$ ) means that confident-appearing AI translations systematically misrepresent colour categories as understood by different cultural communities.

Our findings align with Kim *et al.*'s [41] identification of cross-linguistic colour naming variation, which demonstrated additional nameable blueish colours in Korean and Russian beyond English BCTs. However, while Kim *et al.* focused on identifying these differences within traditional frameworks, our coverage-based approach reveals the fundamental inadequacy of BCT assumptions: the 23-34% of indispensable vocabularies that BCTs represent contrasts sharply with the systematic cross-linguistic differences documented across languages. Claude achieves competence primarily for BCTs representing only 23-34% of indispensable colour vocabularies, while systematically failing on the remaining 66-77% of names users naturally employ in conversational interfaces. This limitation directly undermines modern HCI paradigms that promise natural language interaction. Users expect design software, voice assistants, and accessibility tools to understand their full colour vocabulary from *sage green* in web design to *burgundy* in fashion applications. When systems fail to comprehend common terms like *teal*, *maroon*, or *peach*, the user experience degrades from natural interaction to frustrating constraint-based selection.

Comparison of our empirically derived focal colours against Heer and Stone's [32] model revealed moderate perceptual differences ( $M\Delta E=10.60$  for 47 American English terms using their top-4 averaging method;  $M\Delta E=8.93$  for 11 BCTs), whilst comparison of centroids against Kim *et al.*'s [41] multilingual model showed better alignment for American English ( $M\Delta E=5.66$ ) and moderate alignment for French ( $M\Delta E=7.24$ ). While these are substantially larger than our own American and British English BCT centroids showing tight agreement ( $M\Delta E=1.55$ ), the differences with existing models remain substantially smaller than Claude's translation errors ( $M\Delta E=17-28$ ). These differences likely reflect methodological variations including background colour (white vs. neutral grey) and sampling of the colour space, but establish that purpose-built colour models achieve reasonable perceptual accuracy. Claude's 2–4 $\times$  larger errors therefore represent a fundamental regression: general-purpose language models trained on text alone cannot recover the perceptual grounding that specialized models achieve through explicit colour space modelling.

Claude applies universal colour prototypes consistently across languages rather than performing genuine cross-cultural translation. The variation in translation accuracy reflects differential alignment between Claude's fixed colour mappings and target languages' empirical colour spaces. This creates misleading user experiences where linguistic competence masks systematic failures in genuine cross-cultural colour understanding, delivering poor perceptual fidelity precisely where inclusive HCI applications require robust cross-cultural communication capabilities.

Applications requiring precise colour reference—such as accessibility interfaces, cross-cultural e-commerce, or international design collaboration—need translation systems integrating empirically grounded colour-semantic mappings rather than relying solely on distributional linguistic patterns. The systematic breakdown for non-European languages illuminates broader challenges in developing culturally aware AI systems extending beyond resource

availability to architectural constraints in representing cultural diversity.

## 7.6 Limitations and Future Directions

*LLM evaluation.* Our concentrated evaluation of Claude Sonnet 4 enables deep analysis of systematic patterns in current state-of-the-art LLM translation capabilities. Preliminary studies of other LLMs support that the observed lexical-perceptual dissociation represents general architectural constraints rather than model-specific issues. Comprehensive evaluation across diverse LLM architectures will be pursued in future work; here we establish the methodological template and demonstrate it with a leading contemporary model.

Our evaluation focuses on literal colour naming and does not address contextual or figurative uses where colour terms carry cultural meanings beyond perceptual reference. Expressions like “green with envy” or Greek “πράσινα άλογα” (green horses = absurdities) require understanding both perceptual referents and culture-specific associations—linguistic competencies beyond the perceptual grounding we evaluate here.

*Language coverage.* Our five-language scope was deliberately chosen to span critical dimensions of cultural and linguistic diversity, from efficient traditional systems (7 terms in Himba) to complex industrial vocabularies (47 terms in American English), with Greek and French representing intermediate complexity. This strategic sampling reveals fundamental patterns that broader but shallower coverage could not detect. However, critical gaps remain: languages with unique categorical distinctions (e.g., Italian’s multiple blue categories) and under-represented regions (South Asia, Southeast Asia, indigenous Americas). Our methodology framework enables rapid expansion: ongoing data collection across 20+ language communities can be seamlessly integrated to extend cultural coverage while maintaining analytical depth.

*Viewing conditions and scope.* Our web-based data collection reflects ecologically valid diversity—participants used personal devices with varying display technologies and ambient lighting. We collected extensive metadata on these variables but did not systematically analyse device-specific effects on colour naming consistency. Such investigation requires focused, single-language examination of individual parameters and represents a distinct research question for future work.

*Design implications and future directions.* Our findings have direct implications for colour interface design, though realising these implications requires dedicated HCI research. Potential applications include: uncertainty visualisation when colour naming confidence is low, focal exemplar swatches illustrating cross-linguistic category centres, cross-lingual palette mediators that account for vocabulary asymmetries, and  $\Delta E$ -based warnings when translations exceed perceptual acceptability thresholds. The datasets, models, and benchmarks we provide establish the computational foundations necessary for such interfaces, whilst the design, implementation, and evaluation of these UI patterns merit future investigation combining our cross-cultural colour models with user-centred design methodologies. Promising extensions include active learning for efficient data collection in under-resourced languages, multi-modal vision-language models to address perceptual grounding failures,

and explicit cultural-semantic representations capturing meaning beyond perceptual overlap.

## 8 Conclusion

This study establishes empirical foundations for cross-linguistic colour communication through three interconnected contributions that advance both cross-cultural understanding of how people communicate about colour and human-computer interaction. First, we created comprehensive datasets of unconstrained colour naming responses in five languages to capture genuine linguistic diversity. Second, we developed the Spin Colour Forest model, which applies partial orthogonal transformations to perceptual colour space, consistently improving colour naming prediction across five languages while identifying minimal indispensable vocabularies. Third, we developed a systematic benchmark for evaluating large language model colour translation, revealing fundamental dissociations between lexical accuracy and perceptual fidelity in current translation capabilities of AI systems.

For HCI practitioners, our findings provide actionable guidance: colour interfaces should support language-specific indispensable colour vocabularies rather than assuming universal categories; AI-powered colour tools require perceptual grounding beyond text-based training; and cross-cultural design systems must account for fundamental information-theoretic constraints in colour translation. The datasets, models, and benchmarks we release enable designers and researchers to build culturally-aware colour tools grounded in empirical understanding of how different communities categorise and communicate about colours (Available on GitHub and OSF. The Colour Spinner ([colorspinner.colornaming.com](https://colorspinner.colornaming.com)) provides an interactive demonstration.

## Acknowledgments

Thanks to thousands of participants whose colour-naming contributions made this study possible; Jules Davidoff and Serge Caparos for the collection of the Himba data (British Academy/Leverhulme Grant SG171176); Mathilde Josserand for the French data collection; Jonathan Stutters for maintaining the online experiment infrastructure over many years; and Jerone Andrews for his support on an earlier version of the computational model. This work was funded by Northeastern University TIER1 FY21 and is currently supported by Leverhulme Trust RPG-2024-096.

## References

- [1] Saeedeh Abasi, Mohammad Amani Tehran, and Mark D Fairchild. 2020. Distance metrics for very large color differences. *Color Research and Application* 45, 2 (2020), 208–223.
- [2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining*. Association for Computing Machinery, Anchorage, AK, USA, 2623–2631.
- [3] Anthony Bagnall, M Flynn, J Large, J Line, A Bostrom, and G Cawley. 2018. Is rotation forest the best classifier for problems with continuous features?
- [4] Robert Benavente, Maria Vanrell, and Ramon Baldrich. 2008. Parametric fuzzy sets for automatic color naming. *Journal of the Optical Society of America A* 25, 10 (2008), 2582–2593.
- [5] Toby Berk, Lee Brownston, and Arie Kaufman. 1982. A new color-naming system for graphics languages. *IEEE Computer Graphics and Applications* 2, 03 (1982), 37–44.
- [6] Brent Berlin and Paul Kay. 1991. *Basic Color Terms: Their Universality and Evolution*. University of California Press, Berkeley, CA, USA.

- [7] Carole Patricia Biggam. 2012. *The semantics of colour: A historical approach*. Cambridge University Press, Cambridge, UK.
- [8] Robert M Boynton and Conrad X Olson. 1987. Locating basic colors in the OSA space. *Color Research and Application* 12, 2 (1987), 94–105.
- [9] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
- [10] Roger W Brown and Eric H Lenneberg. 1954. A study in language and cognition. *The Journal of Abnormal and Social Psychology* 49, 3 (1954), 454.
- [11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [12] Alphonse Chapanis. 1965. Color names for color space. *American Scientist* 53, 3 (1965), 327–346.
- [13] Zhiyi Cheng, Xiaoxiao Li, and Chen Change Loy. 2016. Pedestrian color naming via convolutional neural network. In *Asian Conference on Computer Vision*. Springer, Taipei, Taiwan, 35–51.
- [14] Jason Chuang, Maureen Stone, and Pat Hanrahan. 2008. A probabilistic model of the categorical association between colors. In *Color and Imaging Conference*, Vol. 16. Society of Imaging Science and Technology, Society of Imaging Science and Technology, Portland, OR, USA, 6–11.
- [15] Greville G. Corbett and Ian R. L. Davies. 1997. Establishing Basic Colour Terms: Measures and Techniques. In *Color Categories in Thought and Language*, C. L. Hardin and Luisa Maffi (Eds.). Cambridge University Press, Cambridge, 197–223.
- [16] Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation.
- [17] Kenny Coventry, Christos Mitsakis, Ian Davies, Julio Lillo Jover, Anna Androulaki, and Natalia Gómez-Pestaña. 2006. Basic colour terms in Modern Greek: Twelve terms including two blues. *Journal of Greek Linguistics* 7, 1 (2006), 3–47.
- [18] TD Crawford. 1982. Defining "basic color term". *Anthropological linguistics* 24, 3 (1982), 338–343.
- [19] Jules Davidoff. 1991. *Cognition Through Color*. MIT Press, Cambridge, MA, USA.
- [20] Ian Davies, Greville Corbett, Al Mtenje, and Paul Sowden. 1995. The basic colour terms of Chichewa. *Lingua* 95, 4 (1995), 259–278. doi:10.1016/0024-3841(94)00024-G
- [21] Ian R L Davies and Greville G Corbett. 1995. A practical field method for identifying probable basic colour terms. *Languages of the World* 9, 1 (1995), 25–36.
- [22] Guinilla Derefeldt and Tiina Swartling. 1995. Colour concept retrieval by free colour naming. Identification of up to 30 colours without training. *Displays* 16, 2 (1995), 69–77.
- [23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. Association for Computational Linguistics, Minneapolis, MN, USA, 4171–4186.
- [24] Maxim Enis and Mark Hopkins. 2024. From llm to nmt: Advancing low-resource machine translation with claude.
- [25] Peter Gardenfors. 2004. *Conceptual spaces: The geometry of thought*. MIT press, Cambridge, MA, USA.
- [26] Daniel J Garside, Audrey LY Chang, Hannah M Selwyn, and Bevil R Conway. 2025. The origin of color categories. *Proceedings of the National Academy of Sciences* 122, 1 (2025), e2400273121.
- [27] Karl R Gegenfurtner and Jochem Rieger. 2000. Sensory and cognitive contributions of color to the recognition of natural scenes. *Current Biology* 10, 13 (2000), 805–808.
- [28] Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine learning* 63, 1 (2006), 3–42.
- [29] Edward Gibson, Richard Futrell, Julian Jara-Ettinger, Kyle Mahowald, Leon Bergen, Sivalogeswaran Ratnasingam, Mitchell Gibson, Steven T Piantadosi, and Bevil R Conway. 2017. Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences* 114, 40 (2017), 10785–10790.
- [30] Connor C Gramazio, David H Laidlaw, and Karen B Schloss. 2016. Colorgical: Creating discriminable and preferable color palettes for information visualization. *IEEE transactions on visualization and computer graphics* 23, 1 (2016), 521–530.
- [31] Lewis D Griffin and Dimitris Mylonas. 2019. Categorical colour geometry. *PLoS one* 14, 5 (2019), e0216296.
- [32] Jeffrey Heer and Maureen Stone. 2012. Color Naming Models for Color Selection, Image Editing and Palette Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1007–1016.
- [33] Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Affy, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.
- [34] Alston S Householder. 1958. Unitary triangularization of a nonsymmetric matrix. *Journal of the ACM (JACM)* 5, 4 (1958), 339–342.
- [35] Yan Huang and Wei Liu. 2024. Evaluating the Translation Performance of Large Language Models Based on Euas-20.
- [36] Gerhard Jäger. 2010. Natural color categories are convex sets. In *Logic, Language and Meaning: 17th Amsterdam Colloquium, Amsterdam, The Netherlands, December 16-18, 2009, Revised Selected Papers*. Springer, Berlin, Heidelberg, 11–20.
- [37] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 6282–6293.
- [38] Paul Kay, Brent Berlin, Luisa Maffi, William R. Merrifield, and Richard Cook. 2011. *The World Color Survey*. Lecture Notes, Vol. 159. CSLI Publications, Stanford, CA, USA.
- [39] Paul Kay and Chad K McDaniel. 1978. The linguistic significance of the meanings of basic color terms. *Language* 54, 3 (1978), 610–646.
- [40] Kenneth Low Kelly. 1955. *The ISCC-NBS method of designating colors and a dictionary of color names*. Vol. 553. US Department of Commerce, National Bureau of Standards, Washington, DC, USA.
- [41] Younghoon Kim, Kyle Thayer, Gabriella Silva Gorsky, and Jeffrey Heer. 2019. Color Names Across Languages: Salient Colors and Term Translation in Multilingual Color Naming Models. In *EuroVis (Short Papers)*. Eurographics Association, Porto, Portugal, 31–35.
- [42] Jan Koenderink, Andrea van Doorn, and Karl Gegenfurtner. 2018. Graininess of RGB-Display Space. *i-Perception* 9, 3 (2018), 2041669518803971.
- [43] Garry Kuwanto, Afra Fezza Akyürek, Isidora Chara Tourni, Siyang Li, Alex Jones, and Derry Wijaya. 2023. Low-resource machine translation training curriculum fit for low-resource languages. In *Pacific Rim International Conference on Artificial Intelligence*. Springer, Singapore, 453–458.
- [44] Johan Maurice Gisele Lammens. 1994. *A computational model of color perception and color naming*. Ph.D. Dissertation. State University of New York at Buffalo, Buffalo, NY, USA.
- [45] DeLee Lantz and Volney Stefflre. 1964. Language and cognition revisited. *The Journal of Abnormal and Social Psychology* 69, 5 (1964), 472.
- [46] Anne Lauscher, Surangika Ranathunga, Edoardo Maria Ponti, Goran Glavaš, Roi Reichart, Ivan Vulić, and Anna Korhonen. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 4483–4499.
- [47] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.
- [48] Changjun Li, Zhiqiang Li, Zhifeng Wang, Yang Xu, Ming Ronnier Luo, Guihua Cui, Manuel Melgosa, Michael H Brill, and Michael Pointer. 2017. Comprehensive color solutions: CAM16, CAT16, and CAM16-UCS. *Color Research and Application* 42, 6 (2017), 703–718.
- [49] Helen Lin, M Ronnier Luo, Lindsay W MacDonald, and AWS Tarrant. 2001. A cross-cultural colour-naming study: Part II—Using a constrained method. *Color Research and Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur* 26, 3 (2001), 193–208.
- [50] Helen Lin, M Ronnier Luo, Lindsay W MacDonald, and Arthur WS Tarrant. 2001. A cross-cultural colour-naming study. Part I: Using an unconstrained method. *Color Research and Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur* 26, 1 (2001), 40–60.
- [51] Sharon Lin, Julie Fortuna, Chinmay Kulkarni, Maureen Stone, and Jeffrey Heer. 2013. Selecting semantically-resonant colors for data visualization. In *Computer graphics forum*, Vol. 32. Wiley Online Library, Wiley-Blackwell, Leipzig, Germany, 401–410.
- [52] M. Ronnier Luo, Qiyang Xu, Michael Pointer, Manuel Melgosa, Guihua Cui, Changjun Li, Kaida Xiao, and Minchen Huang. 2023. A comprehensive test of colour-difference formulae and uniform colour spaces using available visual datasets. *Color Research and Application* 48, 3 (2023), 267–282.
- [53] Gloria Menegaz, Arnaud Le Troter, Jean Sequeira, and Jean-Marc Boi. 2006. A discrete model for color naming. *EURASIP Journal on Advances in Signal Processing* 2007, 1 (2006), 029125.
- [54] Nathan Moroney. 2003. Unconstrained web-based color naming experiment. In *Color imaging VIII: Processing, hardcopy, and applications*, Vol. 5008. SPIE, Santa Clara, CA, USA, 36–46.
- [55] Hideto Motomura, Osamu Yamada, and Teruo Fumoto. 1997. Categorical color mapping for gamut mapping. In *Color and Imaging Conference*, Vol. 5. Society of Imaging Science and Technology, Society of Imaging Science and Technology, Scottsdale, AZ, USA, 50–55.
- [56] Randall Munroe. 2010. Color Survey Results. <http://blog.xkcd.com/2010/05/03/color-survey-results/>. Accessed: November 24, 2025.

- [57] Munsell Color. n.d.. Munsell Books of Color. <https://munsell.com/color-products/color-communications-products/munsell-books-and-sheets/>
- [58] Dimitris Mylonas. 2020. *Colour Communication Within Different Languages*. PhD thesis. University College London. <https://discovery.ucl.ac.uk/id/eprint/10089125/>
- [59] Dimitris Mylonas, Serge Caparos, and Jules Davidoff. 2022. Augmenting a colour lexicon. *Humanities and Social Sciences Communications* 9, 1 (2022), 1–12.
- [60] D Mylonas, DL Griffin, and A Stockman. 2019. Mapping colour names in cone excitation space. In *25th Symposium of the International Color Vision Society, Riga, International Color Vision Society, Riga, Latvia*, 1–10.
- [61] Dimitris Mylonas and Lindsay MacDonald. 2010. Online colour naming experiment using Munsell samples. In *Proceedings of the 5th European Conference on Colour in Graphics, Imaging, and Vision (CGIV 2010)*. Society for Imaging Science and Technology, Joensuu, Finland, 27–32.
- [62] D Mylonas and L MacDonald. 2012. Colour naming for colour communication. In *Colour Design*. Elsevier, Amsterdam, Netherlands, 254–270.
- [63] Dimitris Mylonas and Lindsay MacDonald. 2016. Augmenting basic colour terms in English. *Color Research and Application* 41, 1 (2016), 32–42.
- [64] Dimitris Mylonas, Lindsay MacDonald, and Sophie Wuerger. 2010. Towards an online color naming model. In *Color and imaging conference*, Vol. 18. Society of Imaging Science and Technology, Society of Imaging Science and Technology, San Antonio, TX, USA, 140–144.
- [65] Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, et al. 2020. Masakhane—machine translation for Africa.
- [66] Galina V Paramei, Yulia A Griber, and Dimitris Mylonas. 2018. An online color naming experiment in Russian using Munsell color samples. *Color Research and Application* 43, 3 (2018), 358–374.
- [67] C Alejandro Parraga and Arash Akbarinia. 2016. NICE: A computational solution to close the gap from colour perception to colour categorization. *PLoS one* 11, 3 (2016), e0149538.
- [68] David L Philipona and J Kevin O’regan. 2006. Color naming, unique hues, and hue cancellation predicted from singularities in reflection properties. *Visual neuroscience* 23, 3-4 (2006), 331–339.
- [69] Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning.
- [70] Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the tenth workshop on statistical machine translation*. Association for Computational Linguistics, Lisbon, Portugal, 392–395.
- [71] Terry Regier, Paul Kay, and Naveen Khetarpal. 2007. Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences* 104, 4 (2007), 1436–1441.
- [72] Debi Roberson, Jules Davidoff, Ian RL Davies, and Laura R Shapiro. 2005. Color categories: Evidence for the cultural relativity hypothesis. *Cognitive psychology* 50, 4 (2005), 378–411.
- [73] Juan José Rodríguez, Ludmila I Kuncheva, and Carlos J Alonso. 2006. Rotation forest: A new classifier ensemble method. *IEEE transactions on pattern analysis and machine intelligence* 28, 10 (2006), 1619–1630.
- [74] Barbara AC Saunders and Jaap Van Brakel. 1997. Are there nontrivial constraints on colour categorization? *Behavioral and brain sciences* 20, 2 (1997), 167–179.
- [75] Roger N Shepard. 1992. The perceptual organization of colors: an adaptation to regularities of the terrestrial world? In *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, Jerome H. Barkow, Leda Cosmides, and John Tooby (Eds.). Oxford University Press, New York, NY, USA, 495–532.
- [76] Akvile Sinkeviciute, Julien Mayor, Mila Dimitrova Vulchanova, and Natalia Kartushina. 2024. Active language modulates color perception in bilinguals. *Language Learning* 74, S1 (2024), 40–71.
- [77] Luc Steels, Tony Belpaeme, et al. 2005. Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and brain sciences* 28, 4 (2005), 469–488.
- [78] Julia Sturges and TW Allan Whitfield. 1995. Locating basic colours in the Munsell space. *Color Research and Application* 20, 6 (1995), 364–376.
- [79] Shoji Tominaga. 1985. A colour-naming method for computer color vision. In *Proceedings of the 1985 IEEE International Conference on Cybernetics and Society*, Vol. 573. IEEE, Tucson, AZ, USA, 577.
- [80] Chin Tseng, Ghulam Jilani Quadri, Zeyu Wang, and Danielle Albers Szafir. 2023. Measuring categorical perception in color-coded scatterplots. In *proceedings of the 2023 CHI conference on human factors in computing systems*. Association for Computing Machinery, Hamburg, Germany, 1–14.
- [81] Mari Uskūla. 2019. Translation of colour terms: An empirical approach toward word-translation from English into Estonian. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics* 10, 2 (2019), 69–84.
- [82] Joost Van De Weijer and Fahad Shahbaz Khan. 2015. An overview of color name applications in computer vision. In *International Workshop on Computational Color Imaging*. Springer, Saint-Etienne, France, 16–22.
- [83] Joost Van De Weijer, Cordelia Schmid, and Jakob Verbeek. 2007. Learning color names from real-world images. In *2007 IEEE conference on computer vision and pattern recognition*. IEEE, Minneapolis, MN, USA, 1–8.
- [84] Aman Kassahun Wassie, Mahdi Molaei, and Yasmin Moslem. 2024. Domain-specific translation with open-source large language models: Resource-oriented analysis.
- [85] Wikipedia contributors. 2024. List of Wikipedias. [https://en.wikipedia.org/wiki/List\\_of\\_Wikipedias](https://en.wikipedia.org/wiki/List_of_Wikipedias) Accessed: January 8, 2025.
- [86] Christoph Witzel. 2019. Misconceptions about colour categories. *Review of Philosophy and Psychology* 10, 3 (2019), 499–540.
- [87] Kaida Xiao, Chenyang Fu, Dimitris Mylonas, Dimosthenis Karatzas, and Sophie Wuerger. 2013. Unique hue data for colour appearance models. Part II: Chromatic adaptation transform. *Color Research and Application* 38, 1 (2013), 22–29. doi:10.1002/col.20725
- [88] Kaida Xiao, Sophie Wuerger, Chenyang Fu, and Dimosthenis Karatzas. 2011. Unique hue data for colour appearance models. Part I: Loci of unique hues and hue uniformity. *Color Research and Application* 36, 5 (2011), 316–323. doi:10.1002/col.20637
- [89] Danna Xue, Javier Vazquez-Corral, Luis Herranz, Yanning Zhang, and Michael S Brown. 2024. Palette-based color harmonization via color naming. *IEEE Signal Processing Letters* 31 (2024), 1474–1478.
- [90] Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking machine translation with cultural awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, Miami, FL, USA, 13078–13096.
- [91] Lu Yu, Lichao Zhang, Joost van de Weijer, Fahad Shahbaz Khan, Yongmei Cheng, and C Alejandro Parraga. 2018. Beyond eleven color names for image understanding. *Machine Vision and Applications* 29, 2 (2018), 361–373.
- [92] Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. 2018. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences* 115, 31 (2018), 7937–7942. doi:10.1073/pnas.1800521115
- [93] Semir Zeki, Samuel Cheadle, John Pepper, and Dimitris Mylonas. 2017. The Constancy of Colored After-Images. *Frontiers in Human Neuroscience* 11 (2017), 229. doi:10.3389/fnhum.2017.00229
- [94] Semir Zeki, Alvaro Javier, and Dimitris Mylonas. 2019. The biological basis of the experience and categorization of colour. *European Journal of Neuroscience* 51, 2 (2019), 670–680. doi:10.1111/ejn.14557