Russian blues reveal the limits of language influencing colour discrimination

Jasna Martinovic[1], Galina V. Paramei[2], W. Joseph MacInnes[3]

[1] School of Psychology, University of Aberdeen, William Guild Building, AB24 3FX, Aberdeen, UK;
[2] Department of Psychology, Liverpool Hope University, Hope Park, L16 9JD, UK;
[3] School of Psychology, National Research University: Higher School of Economics, Slavyanskaya Square, 4/2, 109074, Moscow, Russia

Tables: 4; Figures: 8; Supplementary Materials: 3

Corresponding author:

Dr Jasna Martinovic
School of Psychology
University of Aberdeen
William Guild Building
Aberdeen
AB24 3UB
UK
phone: +44 1224 272240
email: j.martinovic@abdn.ac.uk

Keywords: colour categorisation, cross-linguistic, Russian blues, Whorfian effect, lightness, hue

Abstract

Chromatic stimuli across a boundary of basic colour categories (BCCs; e.g. blue and green) are discriminated faster than colorimetrically equidistant colours within a given category. Russian has two BCCs for blue, *sinij* 'dark blue' and *goluboj* 'light blue'. These language-specific BCCs were reported to enable native Russian speakers to discriminate cross-boundary dark and light blues faster than English speakers (Winawer et al., 2007, PNAS, 4, 7780-7785). We re-evaluated this finding in two experiments that employed identical tasks as in the cited study. In Experiment 1, Russian and English speakers categorised colours as *sinij/goluboj* or *dark blue/light blue* respectively; this was followed by a colour discrimination task. In Experiment 2, Russian speakers initially performed the discrimination task on *sinij*/*goluboj* and *goluboj/zelënyj* 'green' sets. They then categorised these colours in three frequency contexts with each stimulus presented: (i) an equal number of times (unbiased); more frequent (ii) either *sinij* or *goluboj*; (iii) either *goluboj* or *zelënyj*. We observed a boundary response speed advantage for *goluboj*/ *zelënyj* but not for *sinij/goluboj*. The frequency bias affected only the *sinij/goluboj* boundary such that in a lighter context, the boundary shifted towards lighter shades, and vice versa. Contrary to previous research, our results show that in Russian, stimulus discrimination at the lightness-defined blue BCC boundary is not reflected in processing speed. The *sinij/goluboj* boundary did have a sharper categorical transition than the *dark blue/light blue* boundary, but it was also affected by frequency and order biases, demonstrating that "Russian blues" are less well-structured than previously thought.

1. Introduction

The *universalist view* of colour categorisation, broadly accepted since the seminal work of Berlin and Kay (1969/1991), and in a later revision termed the Universality and Evolution model (Kay, 2015; Kay & Maffi, 1999), posits that pan-human basic colour categories (BCCs) recur and evolve from a minimum of two to a maximum of 11 in a partially fixed order across languages. This view is complemented by the *relativist view*, or the Whorfian hypothesis of linguistic relativity: according to it, different languages divide the colour continuum in an arbitrary way (Saunders & vanBrakel, 1997). In the last two decades, the discourse between these two theoretical views has led to the emergence of the *weak relativity hypothesis*, a framework which reconciles some of the perceived contradictions between universalism and relativism by acknowledging that perceptual, linguistic, social and pragmatic factors all play a role in cognitive processing of colour (e.g. Roberson, 2005). Notably, the *weak relativism* embraces the possibility of emergence of new BCCs and their corresponding basic colour terms (BCTs), specific to a given language, beyond the established 11. Indeed, Berlin and Kay (Berlin & Kay, 1969/1991) were not doctrinaire on the limit and noted the possibility of more than 11 BCCs in Russian, with two basic categories/terms for 'blue': *sinij* 'dark blue' and *goluboj* 'light blue'. The possibility of BCCs exceeding the original "ceiling" has seen further empirical findings in relation to the BLUE area of colour space. In addition to Russian, two basic "blues" are also established in other Eastern Slavonic languages (Ukrainian, Belarusian), several languages in circum-Mediterranean area (Italian, Turkish, Greek, Maltese) and in Japanese and Thai (for reviews see Davidoff, 2015; Paramei & Bimler, 2020).

In an *identification* task colours that fall near the centre of a BCC (category prototypes) have processing advantages compared to colours that fall near a category boundary. As an

example, prototype colours of blue or green categories are named and categorised faster and more accurately than a blue-green/turquoise (Agrillo & Roberson, 2009; Bornstein & Monroe, 1980; Huette & McMurray, 2010; Jraissati, Wakui, Decock, & Douven, 2012). In comparison, in a *same-different discrimination* task, responses are more accurate and faster for stimuli at the category boundary than for within-category stimuli. These phenomena are known as categorical perception (CP) effects (Goldstone & Hendrickson, 2010; Hanley, 2016; Harnad, 1987). More recently, Witzel and Gegenfurtner (2016) proposed a weaker version of CP effects, which they labelled "categorical facilitation", suggesting that categories facilitate only the identification of perceptual differences at the boundary.

"Categorical facilitation" was introduced in light of contradictory evidence for categorical perception, with sources of these contradictions rooted in variation among the studies of a *category-probing metric*, i.e. a measure of the categorical character of perception (performance measure), and a *reference-metric*, a measure of stimulus differences specified in a continuous way (see Christoph Witzel, 2018, for a thorough review). More specifically, recent studies have revisited categorical effects on colour discrimination using reference metrics other than the Munsell colour space, with more precisely controlled stimuli and rigorously defined differences between them – either by a number of just noticeable differences (JNDs; Witzel and Gegenfurtner, 2013, 2015) or perceptual distances estimated in terms of changes in the ratios of cone excitations (MacLeod-Boynton space; Danilova & Mollon, 2014), or the cone-opponent mechanisms (Derrington-Krauskopf-Lennie space; Cropper, Kvansakul, & Little, 2013; Witzel & Gegenfurtner, 2013, 2015), or as ΔE (CIELUV colour space; Jraissati et al., 2012; Witzel & Gegenfurtner, 2016). Importantly, these studies reported a range of nuanced observations, not conforming to a strong categorical effect on colour discrimination:

– no reduction of discrimination thresholds was found at the category boundaries, but reaction times (RTs) appeared to be sensitive to the transition of colour categories when colour differences were suprathreshold (C. Witzel & Gegenfurtner, 2013, 2015, 2016);
– absence of CP impact on threshold discrimination was confirmed for a language-specific category boundary (exemplified by the boundary between Korean *chorok* 'green', *cheongnok* 'blue-green' and *parang* 'blue'); a behavioural advantage was found, however, for suprathreshold cross-boundary discriminations (Roberson, Hanley, & Pak, 2009);
– no enhancement of objectively measured discrimination was found for unique green, i.e. at the categorical boundary between bluish and yellowish hues (Danilova & Mollon, 2014);
– the categorical advantage did not occur uniformly across all category boundaries, with contradictory effects at the green/blue boundary attributed to the difficulty of controlling effects of sensory mechanisms (C. Witzel & Gegenfurtner, 2015);
– lateralisation of the CP effect to the right visual field, which was taken to be the landmark evidence for the influence of language lateralised in the left hemisphere (e.g., Drivonikou et al., 2007; Gilbert, Regier, Kay, & Ivry, 2006), could not be replicated (Brown, Lindsey, & Guckes, 2011; Suegami, Aminihajibashi, & Laeng, 2014; Webster & Kay, 2012; C. Witzel & Gegenfurtner, 2011).
– when identical JND-calibrated pairs of stimuli were presented with either the discrimination or categorisation instruction ("same-different"), colour discrimination (estimated by a *d'*-measure) was shown not to be affected by colour categorisation (Cropper et al., 2013).

The findings listed above provide accumulated evidence in favour of Witzel and Gegenfurtner's (2011; also Christoph Witzel, 2018; C. Witzel & Gegenfurtner, 2011) argument that most previous investigations of CP of colour were based on incorrect assumptions of

objective distances in the CIE colour space, whereby colour stimulus pairs claimed to be psychophysically equally distant may not have been so.

In view of these recent findings, we endeavoured to scrutinise the processing-speed advantage at the *sinij/goluboj* boundary from the "categorical facilitation" perspective, in order to re-evaluate the limits of language's influence on colour perception in the example of Russian blues. As in Winawer et al. (2007), the present study focuses on the two Russian BCTs for blue – *sinij* 'dark blue' and *goluboj* 'light blue'. As these English glosses prompt, the distinction between the "Russian blues" is mainly driven by lightness (Davies & Corbett, 1994; Laws, Davies, & Andrews, 1995). Furthermore, psycholinguistic studies that employed colour stimuli varying in saturation revealed that the two "Russian blues" are also distinct along the saturation dimension, with *goluboj* having lower chromaticness than *sinij* (for reviews see Paramei, 2005, 2007; Paramei, Griber, & Mylonas, 2017; Safuanova & Korzh, 2007). The Whorfian hypothesis predicts that language-specific BCCs would manifest a behavioural advantage at their additional cross-category boundary – as a more accurate and speedier discrimination of colours straddling the boundary – compared to languages that do not differentiate categorically that area of colour space. In line with this prediction, native Russian speakers were reported to discriminate dark and light blues faster than English speakers in a speeded matching-to-target, triad discrimination task (Winawer et al., 2007).

In Winawer et al.'s study, Russian speakers showed a clear cross-category discrimination advantage for "near" colour pairs that disappeared with verbal interference (simultaneous silent rehearsal of a string of digits). However, data for individual colours were not presented: instead, RTs were collapsed across colours, specifically, over three discrimination colour pairs spanning or involving the boundary for the cross-category condition, and, for the within-category

condition, over three discrimination pairs on either side of the boundary. Since there were only four trials per colour pair, this implied 12 trials for the cross-category and 24 trials for the within-category conditions. This number of trials might be satisfactory if the sampled area of colour space was uniform in perceptual terms and if this uniformity translated to RTs. However, as discussed by Mollon and Cavonius (1986), CIE colour spaces were built on the basis of threshold experiments in which participants discriminated nearby colours without any pressure to respond quickly. Thus, equal differences in a perceptually uniform space such as CIELab or CIELUV are not a guarantee that RTs to such stimuli will also be equal. In fact, low-level, cone-opponent and cone-additive mechanisms have a strong effect on RTs in a variety of tasks (Lindsey et al., 2010; Martinovic, Mordal, & Wuerger, 2011).

The basic RT differences are further complicated if stimuli are to be discriminated based on luminance. According to the Weber-Fechner law, the just noticeable change in luminance is a fixed fraction of base luminance (Cornsweet & Pinsker, 1965). Presented on a light background, the darker colour pairs would thus be harder to discriminate, eliciting less accurate and slower responses. As an example, let us consider discriminating 6 cd/m^2 vs. 10 cd/m^2 on a 43 cd/m^2 background as opposed to 38 cd/m^2 vs. 42 cd/m^2 on that same background – the Weber fraction for brightness discrimination being ~0.11-0.14 in humans (Griebel & Schmid, 1997), the first pair would be around the limit of perception, with 0.12 (a difference of 0.093 on a pedestal of 0.767 Weber contrast). Meanwhile the second pair would be easily discriminable. In a between-subjects design, where two groups have a very large basic difference in response speed (in Winawer et al. ca. 180 ms in the “no interference” condition), such non-uniformity of RTs can pose a genuine problem, since between-subjects effects of independent variables on RTs are particularly pronounced for slow responders, as demonstrated by Kliegl, Masson, and Richter

(2010) who recently revisited these fairly well-known methodological concerns. These issues are relevant for between-subjects studies of bilingualism, wherein the experimental group (bilinguals) is significantly slower than the control group (monolinguals). This between-group RT difference is argued to reflect parallel activation of bilinguals' two languages and, hence, slower lexical access to the target language due to temporal costs of inhibiting candidates in the non-target language (Kroll, Bobb, Misra, & Guo, 2008; Yu & Schwieter, 2018).

In Winawer et al.'s study, the boundary RT advantage could have been due to slower responding to darker blues by slower, bilingual Russian speakers, compared to English speakers, inflating RTs in the within-category condition which would contain darker (and, thus, slower) discriminations. Thus, rather than being faster for cross-boundary pairs due to lexical access to *sinij/goluboj* terms, Russian participants in this alternative interpretation would have been much slower for darker within-boundary pairs than for lighter cross-boundary pairs. In light of these concerns, Winawer et al.'s RT evidence in favour of the Whorfian effect is not as robust as it may have initially seemed.

In the present study we reassessed Winawer et al.'s RT advantage across the *sinij*/*goluboj* boundary in two experiments that employed the same categorisation and discrimination tasks as in the study in question. We used a greater number of trials to be able to explore performance for different colour pairs. In Experiment 1, we used a between-subjects design with native Russian speakers and English speakers, as per Winawer et al. In Experiment 2, we compared RTs for the *sinij/goluboj* boundary to RTs for the *goluboj/zelënyj* 'green' boundary in native Russian speakers. In addition, by using categorisation task variations, we examined whether the *sinij*/*goluboj* boundary depends on the presentation context, in particular, the stimulus frequency (cf. Parducci & Wedell, 1986), since such dependence would speak against a firm categorical

nature of the *sinij/goluboj* distinction. By re-evaluating the current evidence of language-driven effects observed for the "Russian blues", our study provides an important contribution to the debate on the penetrability of perception by cognition, which remains a hotly contested topic (e.g. Firestone & Scholl, 2016; O'Callaghan, Kveraga, Shine, Adams, & Bar, 2017).

2. Experiment 1: NATIVE RUSSIAN SPEAKERS VS. ENGLISH-SPEAKING CONTROLS

In Experiment 1, we reassessed the no-interference condition of the Winawer et al.'s study using a between-participants design. From 20 colours in their study, we used 15, though with more trials per colour discrimination pair. Unfortunately, an exact replication was not feasible. First, Winawer et al. do not report the coordinates of the background. Second, it was not possible to produce the four lightest colours used in their experiment while having our monitor calibrated for stable output (i.e., with a maximum of 100 cd/m^2 and a correlated colour temperature of 6500K). We also could not produce the darkest colour, probably due to the limitations of our monitor. Third, we aimed to provide additional detail on RTs for individual dark and light blue colour pairs and an exact replication would be unable to fulfil this objective since Winawer et al. had only 4 trials per colour. The reduced range of colours (from 20 to 15) is a relatively minor change, however, and the language-specific RT advantage for *sinij/goluboj* should still be obtained when the colour range contains the boundary.

Moreover, in addition to a categorisation task, in which each colour is displayed an equal number of times, we employed two versions of the categorisation task with manipulation of stimulus presentation context to probe stability of the *sinij/goluboj* (Russian) and *dark blue/light*

blue (English) boundaries. Specifically, we introduced two biases by increasing the relative frequency of either exemplars in the upper luminance part of the stimulus range ("light bias") or exemplars in the lower luminance part of the stimulus range ("dark bias") during categorisation judgements (cf. Parducci & Perrett, 1971).

2.1. Materials and Methods

2.1.1. Participants

Russian speakers (N=18) and English-speaking controls (N=20) were recruited from the University of Aberdeen student population. Each participant reported normal or corrected-to-normal visual acuity and had normal colour vision as assessed by the City University Colour Vision Test (Fletcher, 1975). Participants gave written informed consent and were reimbursed for their time and effort. Data failed to be recorded for one Russian speaker due to technical failure; data for two controls were removed as no stable category boundary could be estimated (i.e. their responses for dark or light colours alternated rather than showing a relatively sharp switch from "dark" to "light"). Thus, the final sample included 17 native Russian speakers (age range: 21–23 years) and 18 controls (age range: 19–30 years). The study was approved by the ethics committee of the School of Psychology, University of Aberdeen, and was conducted in line with the Declaration of Helsinki.

All Russian speakers were early bilinguals (i.e. with age of acquisition of the second language (L2) between 2−7 years old), with 11 with L2 Estonian (3 reported intermediate and 8 high fluency in Estonian); 4 with L2 Lithuanian (all fluent); 1 with L2 Latvian (fluent); and 1 with L2 English (fluent). Apart from the latter participant, 16 of the native Russian participants were also fluent in English (L3), acquired during their school and/or university studies, so can be considered trilinguals. As can be seen, the Russian speakers were mainly from the Baltic states,

reflecting the fact that European Union nationals are eligible for free university tuition at Scottish universities.

In the control sample, participants were all English speaking, with English being the first language for 11 participants. The remaining participants had following languages as L1: French, German, Pashto, Dutch, Slovenian, Serbo-Croatian and Slovakian. Again, this reflects the multi-cultural population of the University of Aberdeen's student body. Among the control participants, 7 were monolinguals, 6 bilinguals, 2 trilinguals and 3 quadrilinguals.

2.1.2. Materials

The experiment was conducted using a Windows 7 Dell Precision PC with a ViSaGe MKII system [Cambridge Research Systems Ltd. (CRS), Rochester, UK]. The CRS toolbox for Matlab was used to display the stimuli and record the responses. Ilyama VisionMaster Pro 450 cathode ray tube monitor was calibrated using a ColorCal 2 (CRS, Rochester, UK). The maximum luminance output was approximately 100 cd/m^2. The purpose was to ensure that the monitor output was accurate and stable (Metha, Vingrys, & Badcock, 1993). Spectroradiometric measurements (Spectrocal, CRS, Rochester, UK) were used to generate colour stimuli using the CRS Colour Toolbox, as specified in Westland, Ripamonti, and Chung (2012). A Cedrus RB530 response box (Cedrus, San Pedro, CA, USA) was used to collect responses.

The background was set to CIE 1931 coordinates x=0.2962, y=0.3076, Y=48.88 cd/m^2. The darkest blue in the Winawer et al. study had the CIE 1931 coordinates of x=0.1607, y=0.1085, Y= 6.69 cd/m^2 while their lightest blue was x=0.2077, y=0.2377, Y=55.94 cd/m^2. To achieve a stimulus array of equidistant 20 colours, we linearly interpolated 18 colours between the darkest and lightest blues from Winawer et al. in CIELUV space, which resulted in 20

equidistant colours with ΔE=4.25 between neighbouring colours. As shown in Fig. 1a,c, on our setup we could reproduce 15 out of the 20 colours used by Winawer et al., labelled colour 2 to colour 16 (C2–C16); as mentioned above, we could not reproduce their darkest colour and four lightest colours. **Table 1** provides coordinates of the stimulus set in the CIE XYZ and CIE LCh spaces. Note from **Fig. 1c** and **Table 1** that darker blues differ in chromatic properties as well as in lightness, while lighter blues differ mainly in lightness.

Participants were seated approximately 90 cm from the screen in an otherwise dark room. All colour squares subtended 4.8° of visual angle. In the categorisation task, a single square was presented in the centre of the screen. In the discrimination task, three squares were presented in a trial (target, match and alternate), with the target colour centred horizontally and placed 2.9° above the discrimination pair (the match and the alternate squares). The discrimination pair were placed 4.4° right or left from the vertical line (**Fig. 1b**). The location of the match (left or right) was randomised.

Table 1

Coordinates of colour stimuli and the background used in Experiment 1.

Colour No.	***X***	***Y***	***Z***	***L***	***c***	***H***
2	9.88	6.67	44.93	31.04	75.11	295.50
3	11.02	8.25	46.88	34.51	69.07	292.24
4	12.44	10.07	49.83	37.97	64.63	289.13
5	14.09	12.14	53.54	41.44	61.24	286.14
6	15.98	14.48	57.89	44.91	58.56	283.24
7	18.10	17.09	62.80	48.38	56.42	280.39
8	20.46	20.01	68.26	51.84	54.68	277.60
9	23.06	23.23	74.23	55.31	53.26	274.84
10	25.92	26.79	80.71	58.78	52.11	272.12
11	29.04	30.69	87.71	62.24	51.19	269.44

12	32.43	34.95	95.22	65.71	50.45	266.79
13	36.10	39.59	103.25	69.18	49.89	264.18
14	40.06	44.63	111.81	72.65	49.47	261.61
15	44.32	50.07	120.91	76.11	49.18	259.08
16	48.89	55.94	130.56	79.58	49.02	256.59
Bgnd	47.07	48.88	62.96	75.38	9.24	280.70

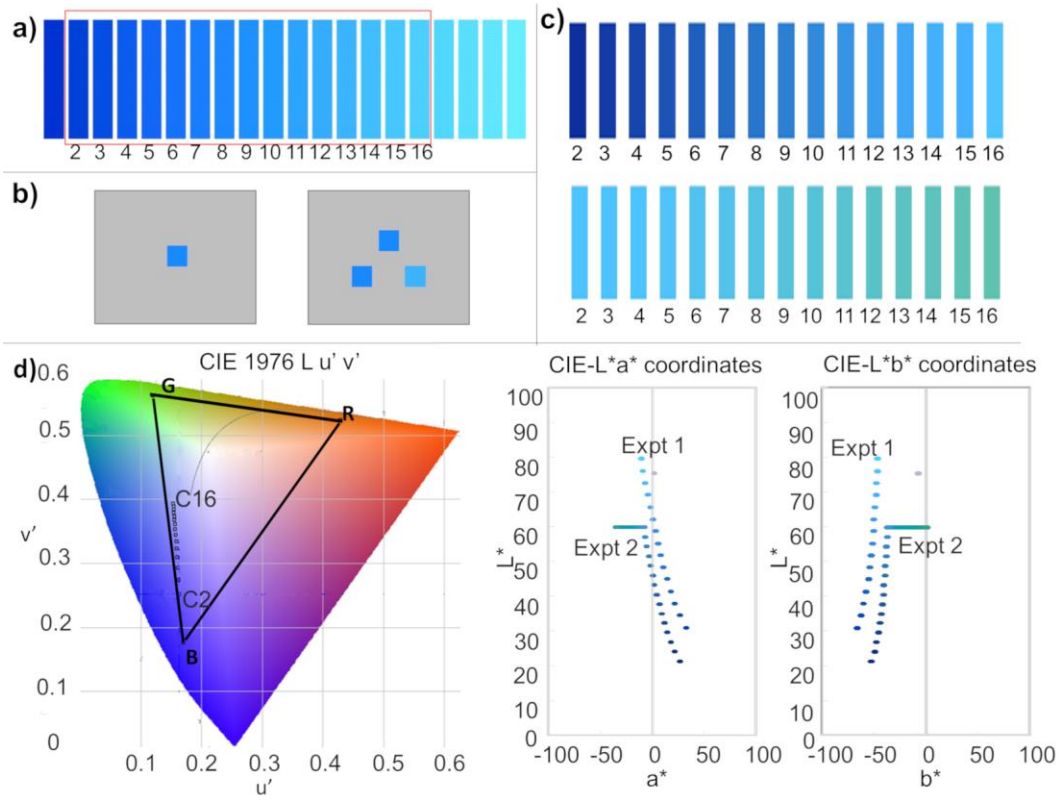


Figure 1. Stimuli and procedure. (a) Stimuli used by Winawer et al. (2007). We were able to reproduce 15 of the colours on our monitor for use in Experiment 1 (Russian vs. English speakers). We have labelled the colours from C2 to C16. (b) The categorisation task involved presentation of a single square. The discrimination, or matching-to-target task involved presentation of a triad of squares, with either the bottom left or bottom right square matching in colour the target on the top. (c) Stimuli used in our Experiment 2 (*sinij/goluboj* and *goluboj/ zelënyj*). Colour appearance on a standard screen roughly approximates their appearance in Experiment 2. Note that colours in the *sinij/goluboj* array vary mainly in lightness while colours in the *goluboj/zelënyj* array change only in hue. Colour 16 (*goluboj*) in the top array is

approximately the same as colour 2 in the bottom array. (d) CIELAB coordinates of stimuli used in Experiment 1 and Experiment 2. The grey dot indicates the background in Experiment 1.

2.1.3. Procedure

***Categorisation task*:** The 15 colours were presented in a random order as single squares and categorised by Russian participants as *sinij/goluboj* and by controls as *dark blue/light blue,* while pressing the left button for *sinij*/*dark blue* and right button for *goluboj*/*light blue*. Following this "no bias" condition, the categorisation task was repeated with four trials for each of the eight lightest colours (C9–C16) and two trials for the seven darkest colours (C2–C8; "light bias") or with four trials for the eight darkest colours (C2–C9) and two trials for the seven lightest colours (C10–C16; "dark bias"). Thus, there were twice as many trials in the biased section of the stimulus range. The trials were randomly intermixed for each participant. The two bias conditions were assigned between participants ("light bias": even participants, "dark bias": odd participants).

Discrimination task (xAB): Participants were instructed to choose, which of the two bottom squares (*A, B*) matched in colour the top square (*x*) using the left and right buttons on the response box. On each trial, one of the bottom squares was an exact match of the top square (target); the colour of the alternate square was two colour steps (ΔE = 8.5) away from the target (e.g. C2 vs. C4), following the colour difference in Winawer et al. (2007), for which the RT facilitation was reported for cross-category colours. The stimuli were displayed until response followed by an empty grey screen for a 2 s inter-trial interval (ITI).

The discrimination task started with a 20-trial practice block to familiarise participants with the task. The discrimination task proper included 13 discrimination colour pairs with two colour steps between the bottom *A* and *B* colours (C2 vs. C4; C3 vs. C5; …; C14 vs. C16). The

presentation of the discrimination pairs was randomised for each participant, with each pair presented 40 times for a total of 520 trials per participant, distributed across eight blocks (65 trials/block).

2.2. Results

All analyses were performed in *R* (R_Core_Team, 2016), using packages *gtools* (Warnes, Bolker, & Lumley, 2015), *Rmisc* (Hope, 2013), *dplyr* (Wickham, Francois, Henry, & Mueller, 2017), *export* (Wenseleers, 2016), *ggplot2* (Wickham, 2009), *lme4* (Bates, Maechler, Bolker, & Walker, 2015) and *emmeans* (Lenth, Singmann, Love, Buerkner, & Herve, 2019).

Categorisation task. First, each participant's boundary was identified as the colour of the transition point from *sinij* to *goluboj* (Russians) or from *dark blue* to *light blue* (controls). As in Winawer et al.'s study, longer RTs were used to disambiguate the boundary since colours closest to boundaries tend to be categorised more slowly (Bornstein & Korda, 1984; Bornstein & Monroe, 1980; see Supplementary **Fig. S1** for examples). The non-biased categorisation task yielded the following *sinij/goluboj* and *dark blue/light blue* (English) boundaries respectively (in terms of the stimulus number, **Fig.1a**): Russians C8.64 ± 1.58 (mean ± SD; range: C7–C11), controls C7.94 ± 1.59 (range: C5–C11). These estimates did not differ significantly ($t(33) = 1.31$, $p = .20$). The boundary plots for individual participants in our study are presented in Supplementary **Fig. S1**.

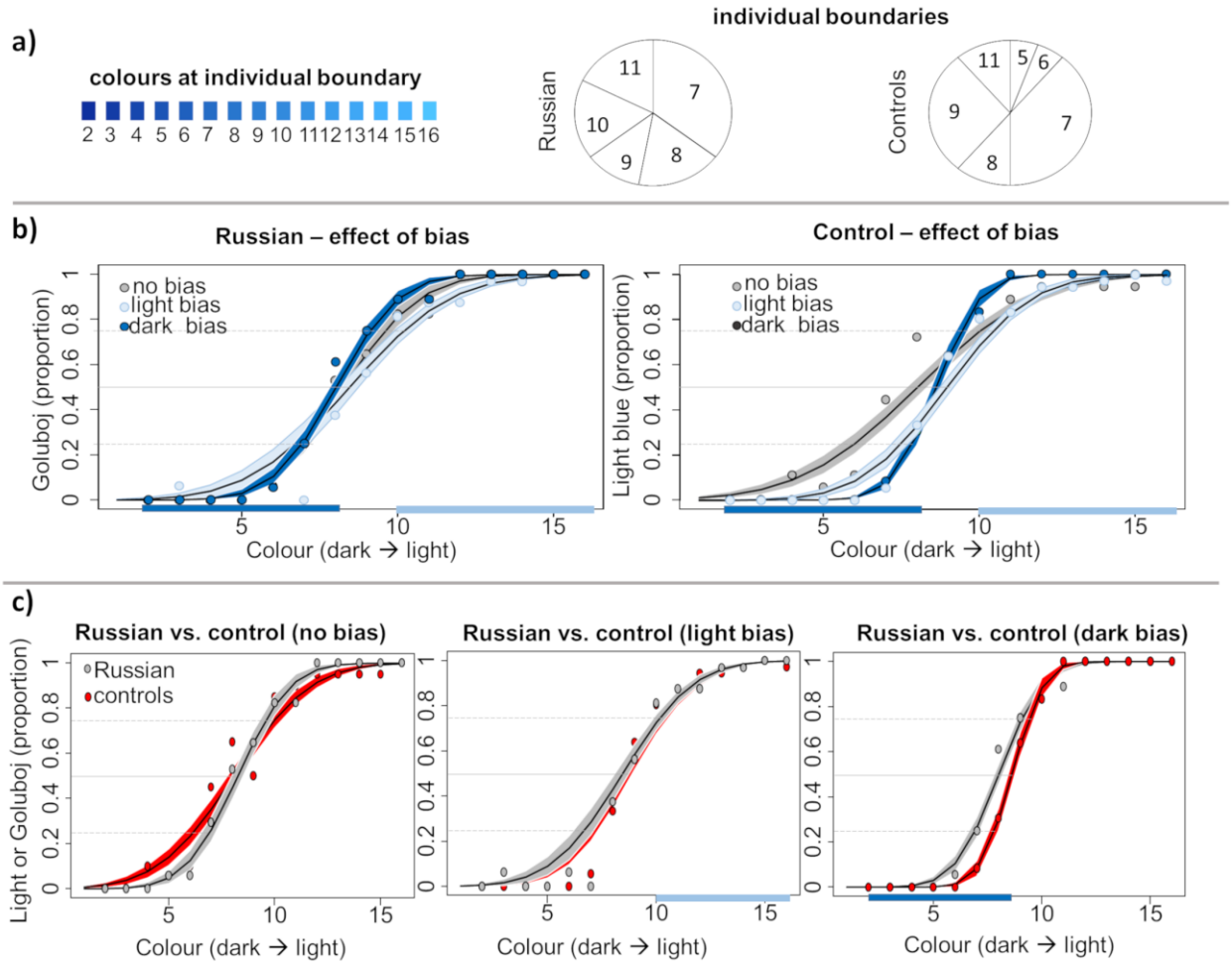


Figure 2. Categorisation of blue colours for Russian speakers and controls. (a) Pie charts of the individual boundary choices, number-coded (colour 9 isrepresented by number 9 etc.). The area of the pie chart's wedges corresponds to the frequency with which each colour represented the boundary in our sample. (b) Panels depict categorisation with and without the lightness-frequency bias, with Russian speakers to the left and controls to the right. Averaged data is depicted by circles. Best-fitting functions are superimposed; the coloured stripes around the black lines correspond to 1 SE. Horizontal grey line corresponds to 50% categorisation, i.e. PSE; dashed lines correspond to 25% and 75% categorisation, i.e. JNDs. (c) Differences in categorisation between Russian and control speakers. The leftmost graph represents unbiased categorisation, the middle graph shows categorisation with light bias and the rightmost graph

with dark bias. Note: the dark and light blue stripes on the *x*-axes of both (a) and (b) highlight the colours that were shown twice as frequently within a biased context (C1-C8 for dark; C10-C16 for light bias). Therefore, in panel (c), these are only highlighted on two of the graphs, as dark and light bias are plotted separately.

We further visualised and statistically tested possible shifts in the boundary due to the dark/light bias manipulations (**Fig. 2**). For each group, data were collapsed across participants as the number of observations for each participant was too small for meaningful within-subjects analyses. Such aggregated boundary data can reveal overall trends in categorisation whilst removing noise inherent to datasets with limited number of observations, as demonstrated by Witzel and Gegenfurtner (2015) who found the same category effects on response times as well as on error rates for both individual and aggregated stimulus pairs.

Further, categorisation at the boundary is inconsistent across repeated measurements which means that, depending on the size of the dataset, there may also be variation in boundary across measurements (see Figs. 6-7 in Witzel & Gegenfurtner, 2013). For Russian speakers, the numbers data points were as follows: for “no bias”, 255 data points (17 participants x 15 categorisations each); for “dark bias”, 414 data points (9 participants x 46 categorisations each); for “light bias”, 368 data points (8 participants x 46 categorisations each; due to failure to record responses for one participant). For controls, for the “no bias” condition, there were 270 data points (18 participants x 15 categorisations each); for both “dark bias” and “light bias”, there were 414 data points (9 participants x 46 categorisations each). From these data, we computed points of subjective equality (PSEs; 50% categorisation) and just noticeable differences (JNDs; difference between 50% and 75% categorisation; cf. Knoblauch & Maloney, 2012).

We tested the influence of stimulus bias by comparing two generalised linear *(glm)* binomial probit models (Knoblauch & Maloney, 2012): a simple model which fits a single psychometric function to all the data (i.e. one intercept and slope) vs. a model which fits separate psychometric functions to the three bias types ("no bias", "light bias" and "dark bias"). We found that Russian speakers' responses were considerably influenced by the stimulus bias ($\chi^2(4) = 16.83$, $p = .002$). This was not due to differences between the "no bias" and "light bias" ($\chi^2(2) = 3.97$, $p = .14$), or "no bias" and "dark bias" ($\chi^2(2) = 2.95$, $p = .23$), but rather due to the difference between the two biased conditions ($\chi^2(2) = 16.19$, $p < .001$; Bonferroni-corrected p value .016). Figure 2 and Table 2 summarise these data.

Table 2

Colour categorisation (expressed as colour number in the stimulus set) PSEs and JNDs for Russian and control participants. Categorisation data was collected without stimulus bias, with a bias created by presenting more exemplars from the upper part of the stimulus range ("light bias") and with a bias created by presenting more exemplars from the lower part of the stimulus range ("dark bias").

	Russian speakers		Controls	
Presentation condition	PSE	JND	PSE	JND
No bias	8.25	1.32	8.02	2.01
Light bias	8.45	1.73	8.62	1.70
Dark dias	8.01	1.08	8.61	0.78

As shown in **Fig. 2** (top right), stimulus bias also affected the control group ($\chi^2(4) = 41.17$, $p < .001$). Bonferroni-corrected post hoc tests between two of the three bias conditions revealed significant differences between "no bias" vs. "dark bias" ($\chi^2(2) = 38.77$, $p < .001$), as well as "light bias" vs. "dark bias" ($\chi^2(2) = 19.58$ $p < .001$), "no bias" vs. "light bias" ($\chi^2(2) = 7.67$, $p = .022$) did not survive correction for multiple comparisons.

We used the same approach to test whether categorisation differed between Russian and English speakers. Notably, the non-biased categorisation for Russian speakers was significantly different from that for English speakers ($\chi^2(2) = 6.57$, $p = .038$): inspection of **Fig. 2** reveals that this is mainly due to a sharper function slope. Differences between the two language groups were even more pronounced at the "dark bias" ($\chi^2(2) = 11.87$, $p = .003$), but disappeared when the "light bias" was introduced ($\chi^2(2) = 0.09$, $p = .96$). For both English and Russian speakers, the slope at the "dark bias" appears to be steeper, as signified by a smaller JND, with the differences appearing to be especially dramatic for English speakers: from a slope shallower than that for Russians with "no bias", they change to a steeper slope when the "dark bias" is implemented, with near-boundary C7 and C8 being categorised as darker than they were with "no bias".

***Discrimination task*.** From the original 18,200 RTs (520 trials x 35 participants), we first excluded all responses in which an incorrect match was selected (3% trials); we then excluded, as outliers, any RTs shorter than 250 ms (only 1 trial) or longer than 1,500 ms (5% of correct trials). For the analysis of RT data in the match-to-target task we followed the approach employed by Winawer et al. Specifically, for the cross-category condition, we took the RTs from discrimination pairs that either crossed the boundary or involved the boundary (there were three such pairs), while selecting three discrimination pairs below and three above that boundary to represent the within-boundary condition. As in Winawer et al.'s study, this resulted in a

maximum of nine discrimination pairs per participant being utilised for the RT analysis. If the boundary was too close to the darkest or lightest end of the stimulus array, RT calculations for the cross- and within-category conditions were based on data for fewer colours, but the cross-category data were never based on fewer than two discrimination pairs, and within-category data were never based on fewer than three discrimination pairs. After this data selection step, we were left with a total of 11,597 trials (69% of the post-exclusion total).

We analysed the RT data using linear mixed effects (LME) models with the following factors: boundary (discriminations at or across the boundary vs. within-category discriminations) and language (Russian or non-Russian). We also included variability of RT across the discrimination pairs (C2 vs. C4, C3 vs. C5, C4 vs. C6 and so on), as well as by-participant variability in intercept and slopes as random effects. LME models are an ideal way of analysis for this type of data, since they can deal with missing data within a factorial model and, also, account for the variability that stems from the discrimination pairs, for which the cross- and within-boundary data was obtained. To determine the best fitting model, we first fitted the maximal model described above and then proceeded to remove, first, the random effects, and then one fixed effect/interaction at a time, while assessing if their inclusion failed to affect the model's fit as measured by a chi-square test.

For assessing this best fitting model, zero level for fixed effects was set for the Russian language group. Mean intercept for the final model at zero level, for Russian speakers, was 734 ms (SE = 38 ms). The random slope effect of variability of RTs across discrimination pairs was kept as it contributed significantly to the model ($\chi^2(1) = 526.76$, $p < .001$). Removal of the language group effect also reduced the fit ($\chi^2(1) = 5.04$, $p = .025$; $\Omega^2=0.40$) with control group (English) faster than Russian speakers (–112 ms, SE = 48 ms). We removed the boundary-by-

language group interaction ($\chi^2(1) = 1.19$, $p = .28$), and the effect of the boundary ($\chi^2(1) = 1.10$, $p = .29$), as their removal did not affect the model fit. Hence, the best fitting model only included the fixed effect of the language group and the random slopes of discrimination pairs. Thus, we found neither an effect of the boundary nor an interaction of the boundary with the language group. **Fig. 3a** shows mean RTs for cross- and within-boundary conditions for Russian speakers and controls; Figure 3b shows the same RTs with within-boundary conditions separated into those that fall below the boundary (i.e. darker blue) and those that fall above the boundary (i.e. lighter blue), while **Fig. 3c** depicts mean RTs for the same main factors separately for all discrimination pairs.

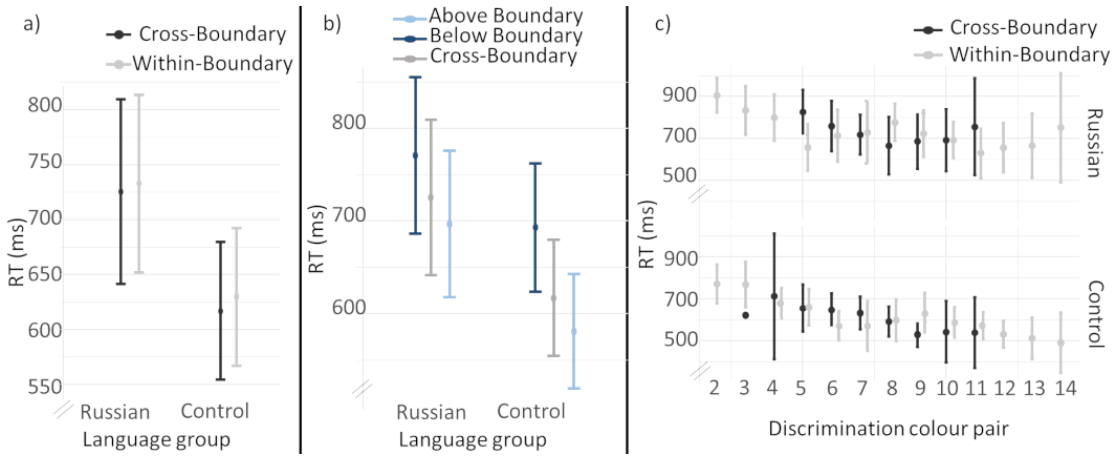


Figure 3. Reaction times in the colour discrimination (xAB) task. (a) Mean RTs for cross-boundary (black) and within-boundary (grey) discriminations for the two groups. (b) Mean RTs for cross-boundary (light grey), as well as below-boundary, darker blues (dark grey), and above-boundary, lighter blues (black), discriminations for the two groups. (c) Cross-boundary and within-boundary categorisations are shown for each colour pair for the Russian speakers (top right) and the controls (bottom right). Discrimination colour pairs are labelled by the darker colour, e.g. pair 2 is C2 vs. C4, while pair 14 is C14 vs. C16.

We then performed a *post hoc* exploratory analysis of RT distributions. Having been surprised by lack of an interaction between the language group and the boundary, we scrutinised whether our "negative case" may have emerged because our Russian speakers were much faster than in the Winawer et al. study, i.e. 734 ms vs. 1,085 ms respectively. This question was prompted by methodological concerns that group differences in RTs can be confounded by the speed of responding, since in slower responders RT effects are increased (e.g. Kliegl et al., 2010). Slower RTs for darker blues (**Fig. 3b**) imply that this may have indeed generated what was presumed to be the boundary effect in Winawer et al.'s study Density and cumulative distributions of RTs are shown in **Fig. 4**. As is apparent from **Fig. 4** (left), compared to English speakers, Russian speakers had a smaller number of very fast responses, along with a larger number of slower responses. The two distributions are not statistically different for Russian speakers, as assessed by a two-sided Kolmogorov-Smirnov test ($D = 0.02$, $p = .72$). RT 95%-confidence intervals for Russian speakers (**Fig. 3a**) varied approximately between 650 and 820 ms. Thus, in our study the longer RTs had lesser influence on the mean RT than in the Winawer et al. study, where the mean RT was around 1,000-1,100 ms.

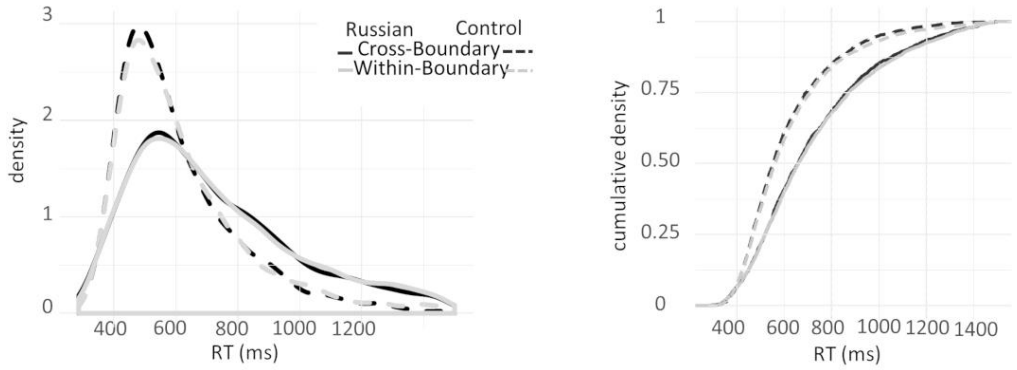


Figure 4. Densities (left) and cumulative densities (right) of RTs for Russian speakers and controls. This graph contrasts cross- and within-boundary colours, showing their distributions across the full range of possible RTs.

Further post hoc RT analyses were conducted to ascertain if the observed findings are reliable and not dependent on the selection of the boundary colours or outlier rejection procedures. First, we considered only cross-boundary discriminations and compared them to the same within-boundary pairs as above, but this did not affect the RT distribution even when outliers longer than 1,500 ms were retained in the dataset (whole sample: $D = 0.0267$, $p = .37$; Russian speakers: $D = 0.0290$, $p = .72$). However, when we considered solely the colour at the PSE (C8) as the boundary for all observers and retained outliers we did obtain a significant difference: RTs were shorter at the boundary, both for the Russian speakers ($D = 0.0575$, $p = .045$) and for the sample as a whole ($D = 0.0421$, $p = .031$). Finally, we also conducted a post hoc analysis of accuracy, using mixed logit models (Jaeger, 2008; see Supplementary Material 2 for more details): these indicate that discrimination of darker pairs resulted in lower accuracy than that of either cross-boundary or lighter pairs in English participants alone. Winawer et al. explored RT/accuracy trade-offs on the aggregate data. However, as darker and lighter pairs may give rise to different RTs, it is possible for ceiling accuracy for faster-to-respond lighter pairs to mask underlying differences in accuracy for slower-to-respond darker pairs when these conditions are combined to represent within-boundary pairs. If English participants responded less accurately and faster to darker pairs, then this could have affected the RT distribution as these would have contained fewer outlier longer RTs in their dataset than more accurate and slower Russian participants.

2.3. Interim conclusions

We did not observe a *sinij/goluboj* boundary effect for Russian speakers, although we acknowledge that our experimental setup differed from Winawer et al.'s in several aspects. In particular, we employed a narrower sampling of blue colours, with five (one darkest and four lightest) blue colours from the Winawer et al. study absent in our stimulus set (cf. colours C1 and C17-C20 in **Fig. 1a**). Note that our labelling of colours starts with the darkest colour and goes towards lighter colours (i.e. opposite to that in the Winawer et al. notation). The category boundary estimated in our study, if expressed from lightest to darkest blues (as in Winawer et al., 2007; Fig. 1), would be C12.36 ± 1.58 for Russian and C13.06 ± 1.59 for controls, thus, approximately 4-5 steps darker than in Winawer et al. (8.7 ± 2.2 for Russians, 8.6 ± 2.5 for controls). It may be that Winawer et al. used the background luminance different from that in our study (the coordinates of the background were not reported in their paper), which would noticeably affect colour appearance. It is likely that their background luminance was greater than that of their lightest blue (Y=55.94 cd/m^2) resulting in a contrast-induced subjective darkening of the test colours (cf. D. L. Bimler, Paramei, & Izmailov, 2009). Alternatively, the observed difference in the category boundary could be attributed to the assimilation effect of the stimulus range, in Winawer et al.'s case towards the lighter end of their stimulus set (Parducci & Wedell, 1986). Indeed, our findings of the bias effects in the categorisation task confirm that the lighter or darker context can significantly affect performance, with the prevalence of lighter or darker stimuli respectively affecting the transition between light and dark blues.

A steeper slope would be indicative of a sharper boundary between the two categories (Huette & McMurray, 2010) – and, indeed, compared to controls, Russians, who possess basic *sinij* and *goluboj* categories, revealed both a steeper slope and the boundary that appeared to be

less affected by biasing. The interpretation of the bias effects is, however, limited by an uneven number of trials for different bias condition categorisations. Therefore, we further investigated the effect of bias in a second experiment, to assess whether the boundary does shift depending on the presentation context – by using the "no bias", "light bias" and "dark bias" contexts with approximately same numbers of trials.

Furthermore, the effect of the boundary on discrimination speed was absent for our sample of Russian-speaking participants. All our participants were recruited from the University of Aberdeen student population, with both Russian speakers and controls performing relatively fast. According to our results, Russian speakers were on average 112 ms slower at the *sinij/goluboj* boundary (734 ms) than were the controls at the *dark blue/light blue* boundary (622 ms). The inter-group RT difference and its sign is comparable to that in the Winawer et al. study, who found that, on average, their Russian participants were slower than English participants by 147 ms (1,085 ms vs. 938 ms respectively). Note though that in both our and Winawer et al.'s study, Russian participants were all bilingual (or, in our case, even trilingual). It is argued that both languages are activated in parallel in bilinguals whereby they experience decelerated lexical access to the target language due to temporal costs of inhibition of candidates in the nontarget language (Kroll et al., 2008). This is supported by the fact that whilst performing the categorisation task, many Russian participants had relatively long RTs at or near the boundary (see Supplementary **Fig. S1**), while the controls were generally very fast.

Finally, while the Russian sample in the Winawer et al. study were at least late bilinguals in English, our Russian speakers were all early bilinguals, predominantly from the Baltic states – Lithuania, Latvia and Estonia (i.e. formerly, 1940-1991, Soviet republics). It is worth noting that light blue term (*žydra*) is basic in Lithuanian (supposedly contact-induced, due to

communication pressure from neighbouring Russian-speaking population), but it is not basic in either Latvian or Estonian (D. Bimler & Uuskula, 2017). We cannot resolve whether (salient) 'blue' terms in the respective L2 Baltic languages of our Russian speaking participants could have interfered with their L1 Russian concepts of *sinij* and *goluboj*, preventing manifestation of their RT advantage to emerge in the match-to-target task. However, we do observe a sharper categorical transition in this sample (**Fig. 2c**) which speaks in favour of their BCCs being more distinct than *dark blue* vs. *light blue* in the controls.

Alternatively, the RT interaction in the Winawer et al. study might have emerged as an amplification of RT differences for darker blue discriminations (see **Fig. 3c**, showing longer RTs for darker discrimination pairs), which might have been more pronounced for the slower-responding Russian speakers. The obvious way to resolve this question was to test a sample of monolingual (or late bilingual) Russian speakers while contrasting categorisation of *sinij/goluboj* with a categorisation along a hue-based boundary, such as *goluboj*/*zelënyj* 'green'. This would enable (i) a direct comparison with the effects at the boundary of the established hue-defined basic colour categories, while also (ii) excluding the possibility of confounding effects of bilingualism, and (iii) avoiding the pitfalls of a between-participants design. In Experiment 2, we therefore proceeded to compare categorisation and matching-to-target solely in Russian (monolingual or late bilingual) speakers for *sinij/goluboj* and *goluboj*/ *zelënyj* 'green' colour sets.

3. Experiment 2: RUSSIANS, *SINIJ/GOLUBOJ* VS. *GOLUBOJ/ZELËNYJ* 'GREEN'

3.1. Materials and Methods

The design of this experiment was preregistered (https://osf.io/5m4h7/). The pre-registration, as well as data and the analysis code are available on the Open Science Framework (OSF).

3.1.1. Participants

Participants were 26 students or staff of the National Research University: Higher School of Economics (NRU: HSE) in Moscow, Russia. All were native Russian speakers, functional monolinguals or late bilinguals with English as L2. Their age ranged from 17–25 years. All participants reported normal or corrected-to-normal vision and self-reported normal colour vision. This was the first time they took part in any colour categorisation or discrimination experiment. Six participants were excluded from analyses (three because of missing data and three other because their category boundary fell beyond our colour range), resulting in a total of 20 participants in the final dataset. All participants gave written informed consent to participate in the experiment. Participants were reimbursed for their time and effort. The study was approved by the NRU: HSE Ethics Committee and followed the tenets of the Declaration of Helsinki.

3.1.2. Materials

The experiment was implemented using psychophysics toolbox for Matlab (Brainard, 1997) on a Windows 7 (64 bit) PC with 16 GB of RAM and Quadro K600 graphics. The monitor was a cathode ray tube HP p1230 (Hewlett-Packard, CA, USA) set to 1024 x 768 pixel resolution and 85 Hz refresh rate. Colour calibration was performed with an X-rite i1 Pro photometer (X-rite, MI, USA) in high resolution mode using displayCAL, an open source

display calibration software powered by ArgylCMS (https://displaycal.net). The room was dimly lit with artificial lighting.

Colour stimuli used in Experiment 2 are given in **Table 3** and presented in **Fig. 1c,d**. The monitor gamut was constrained, therefore the colours from Experiment 1 were halved in luminance (*Y*-values) while maintaining the *X*- and *Z*-values. Further, for the *goluboj*/*zelënyj* 'green' discrimination task, we selected a shade of *goluboj* very close to the lightest blue sample in the *sinij/goluboj* discrimination task, so that we would be able to generate equally light and equally colourful stimuli in the CIE LCh space between this exemplar of *goluboj* and a good exemplar of *zelënyj* 'green'. This ensured that the stimuli of the *goluboj/zelënyj* set varied only in hue and not in any other parameter. The background colour, at two luminance levels, was metameric to D65, with luminance of Y_W = 70 cd/m^2 (white) or Y_G = 35 cd/m^2 (grey). The white background was used for both categorisation and discrimination tasks, while the grey background was used only for the categorisation task.

Table 3. Coordinates of colour stimuli used in Experiment 2.

Colour	***X***	***Y***	***Z***	***L***	***C***	***h***
Blue 2	4.94	3.33	22.46	21.34	59.61	295.50
Blue 3	5.51	4.13	23.44	24.09	54.82	292.24
Blue 4	6.22	5.04	24.91	26.84	51.30	289.13
Blue 5	7.04	6.07	26.77	29.59	48.60	286.14
Blue 6	7.99	7.24	28.94	32.34	46.48	283.24
Blue 7	9.05	8.55	31.40	35.10	44.78	280.39
Blue 8	10.23	10.00	34.13	37.85	43.40	277.60
Blue 9	11.53	11.62	37.11	40.60	42.28	274.84
Blue 10	12.96	13.40	40.36	43.35	41.36	272.12
Blue 11	14.52	15.34	43.85	46.10	40.63	269.44
Blue 12	16.21	17.48	47.61	48.85	40.04	266.79
Blue 13	18.05	19.80	51.62	51.61	39.59	264.18
Blue 14	20.03	22.31	55.90	54.36	39.26	261.61
Blue 15	22.16	25.04	60.45	57.11	39.04	259.08

Blue 16	24.44	27.97	65.28	59.86	38.91	256.59
Green 2	24.44	28	61.67	59.89	35.82	255.19
Green 3	23.71	28	60.54	59.89	35.82	249.79
Green 4	23.01	28	59.10	59.89	35.82	244.39
Green 5	22.36	28	57.38	59.89	35.82	238.99
Green 6	21.75	28	55.40	59.89	35.82	233.59
Green 7	21.20	28	53.22	59.89	35.82	228.19
Green 8	20.70	28	50.86	59.89	35.82	222.79
Green 9	20.25	28	48.36	59.89	35.82	217.39
Green 10	19.87	28	45.77	59.89	35.82	211.99
Green 11	19.54	28	43.13	59.89	35.82	206.59
Green 12	19.27	28	40.46	59.89	35.82	201.19
Green 13	19.06	28	37.82	59.89	35.82	195.79
Green 14	18.91	28	35.23	59.89	35.82	190.39
Green 15	18.82	28	32.72	59.89	35.82	184.99
Green 16	18.79	28	30.31	59.89	35.82	179.59

3.1.2. Procedure

The experiment started with the *xAB* discrimination task, i.e. in its design identical to that in Experiment 1, but with two 390-trial blocks, *sinij/goluboj* and *goluboj/ zelënyj*. The order of discrimination pairs was randomised for each participant. Responses were given using a gamepad. A trial was over when the participant indicated the target match with a left/right selection on the gamepad.

This was followed by multiple categorisation tasks with different stimulus frequencies (lightness biases), as in Experiment 1. Categorisation blocks included variation of the colour sets (2), i.e. *sinij/goluboj* or *goluboj/zelënyj*; backgrounds (2), i.e. white or grey, and the lightness biases (3), i.e. “no bias”, “dark bias” and “light bias”. The order of the 12 categorisation blocks was randomised for each participant.

Participants were asked to make a forced-choice judgement on whether a single square was *goluboj* or *zelënyj* ‘green’ (in one block) and *sinij* or *goluboj* (in another). In the “no bias” blocks, there were 45 trials in each block (3 per colour). In the biased blocks, two thirds of the

colours were from the biased half of the array (i.e. “dark bias” or “light bias”; “blue bias” or “green bias” respectively), with 46 trials in each block (four per colour in the biased half, two per colour otherwise).

Participants were allowed as many breaks between blocks as they wished; as a rule, the whole experiment took about 1 hour 20 minutes.

3.2. Results

Data analysis was performed in the same manner as in Experiment 1. Again, we first report outcomes of the categorisation task.

Categorisation task. For the “no bias” condition, we obtained 900 data points (20 participants x 45 categorisations each); for the “dark bias” and “light bias”, there were 920 data points (20 participants x 46 categorisations each).

The *sinij/goluboj* boundary was 11.45 ± 1.99 (mean ± SD; range 8-16) for the white background and 10.55 ± 1.79 (mean ± SD; range 6-13) for the grey background. The g*oluboj/zelënyj* boundary was 9.85 ± 1.09 (mean ± SD; range 7-11) for the white background and 10.80 ± 1.96 (mean ± SD; range 6-14) for the grey background. Boundary plots for individual observers are presented in Supplementary **Table S3**.

Data were collapsed across observers to assess how the boundary shifted due to the background luminance and/or frequency bias. **Table 4** shows that the background luminance affected the boundary location both for *sinij/goluboj* ($\chi^2(2) = 44.33$, $p < .001$) and *goluboj*/*zelënyj* ($\chi^2(2) = 23.07$, $p < .001$): specifically, the grey background resulted in the boundary shift towards darker blue or greener colours respectively. The effects of frequency bias were much more selective, affecting only the boundary between the blue categories defined by lightness. As

shown in **Fig. 5** (top left), at the white background psychometric functions for *sinij/goluboj* shifted in the direction of the bias ($\chi^2(4) = 37.50$, $p < .001$). Significant differences were observed between all three bias conditions: "no bias" and "light bias" ($\chi^2(2) = 12.18$, $p = .002$); "no bias" and "dark bias" ($\chi^2(2) = 8.73$, $p = .013$); "light bias" and "dark bias" ($\chi^2(2) = 32.86$, $p < .001$; Bonferroni-corrected criterion p- = .016). At the grey background, biasing also resulted in boundary shifts ($\chi^2(4) = 33.35$, $p < .001$; **Fig. 5**, bottom left). Significant differences were observed between "no bias" and "light bias" ($\chi^2(2) = 13.15$, $p = .0014$), "no bias" and "dark bias" ($\chi^2(2) = 29.27$, p < .001), as well as "light bias" and "dark bias" ($\chi^2(2) = 9.09$, $p = .011$). For *goluboj*/*zelënyj*, in comparison (**Fig. 5**, bottom right), there was no effect of bias either for the white background ($\chi^2(4) = 5.57$, $p = .23$) or grey background ($\chi^2(4) = 8.52$, $p = .07$).

Table 4. Colour category boundary (in terms of colour numbers) PSEs and JNDs for *sinij/goluboj* and *goluboj/zelënyj*. Categorisation data was collected without stimulus bias or with biases towards the lower or the upper part of the stimulus range.

Background luminance	White Background		Grey background	
sinij/goluboj				
Presentation condition	PSE	JND	PSE	JND
No bias	11.54	1.73	10.38	2.43
Dark bias	10.88	1.80	9.86	1.93
Light bias	12.29	1.93	9.15	1.85
goluboj/*zelënyj*				
	PSE	JND	PSE	JND

No bias	9.77	1.64	10.80	1.62
Dark bias	9.67	1.61	11.03	1.73
Light bias	10.04	1.79	11.39	1.63

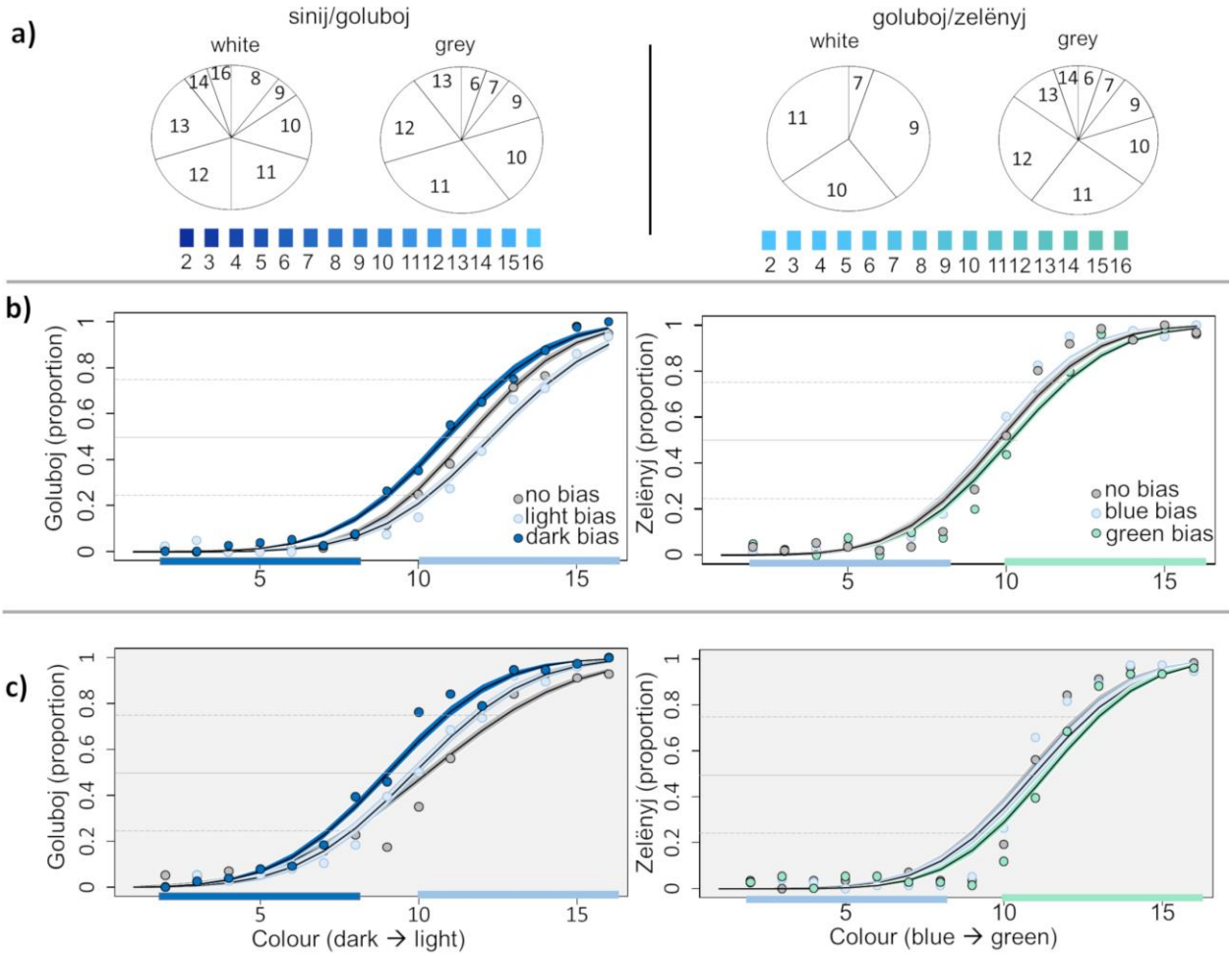


Figure 5. Categorisation of colours for *sinij/goluboj* and *goluboj/zelënyj*. (a) Pie charts of the individual boundary choices, number-coded (colour 6 isrepresented by number 6, etc.). The area of the pie chart's wedges correspond to the frequency with which each colour represented the boundary in our sample. (b) Panels depict categorisation on a white background for *sinij/golubo*j (left) and *goluboj/zelënyj* (right). c) Categorisation on a grey background for *sinij/goluboj* (left) and *goluboj/zelënyj* (right). Averaged data is depicted by circles. Best-fitting functions are superimposed: the coloured stripes around the black lines correspond to 1 SE. The horizontal

grey line corresponds to 50% categorisation, i.e. PSE; the dashed lines correspond to 25% and 75% categorisation, i.e. JNDs. Note: the dark/light blue or light blue/green strips on the x-axes of both (b) and (c) highlight the colours that were shown twice as frequently within a biased context (C2-C8 for lower bias; C10-C16 for upper bias).

Sinij/goluboj and *goluboj*/*zelënyj* boundary positions were further examined by visualising (N-1) effects in categorisation: this would reveal whether a preceding trial had affected categorisation on a subsequent trial, also known as the order bias. We conjectured that closer to the prototype colour (for an overview of prototype models of categorisation, see Hampton, 1998), the influence of the context provided by the preceding trial would be minimal as categorical membership should be unambiguous; however, near the boundary, with colours' lesser degree of category membership (cf. Douven, Wenmackers, Jraissati, & Decock, 2017), one would expect a stronger influence of context. The (N-1) plots are shown in **Fig. 6**. Colours that fall between 1 JND of the PSEs do not just have less firm category membership (categorisation scores that do not fall on the 0 and 100% lines) but also exhibit repulsive effects if the previous trial was e.g. much darker or much lighter. This demonstrates that (N-1) effects on aggregated data represent another way to investigate category boundaries, providing largely compatible results to the PSEs depicted in **Fig. 5**.

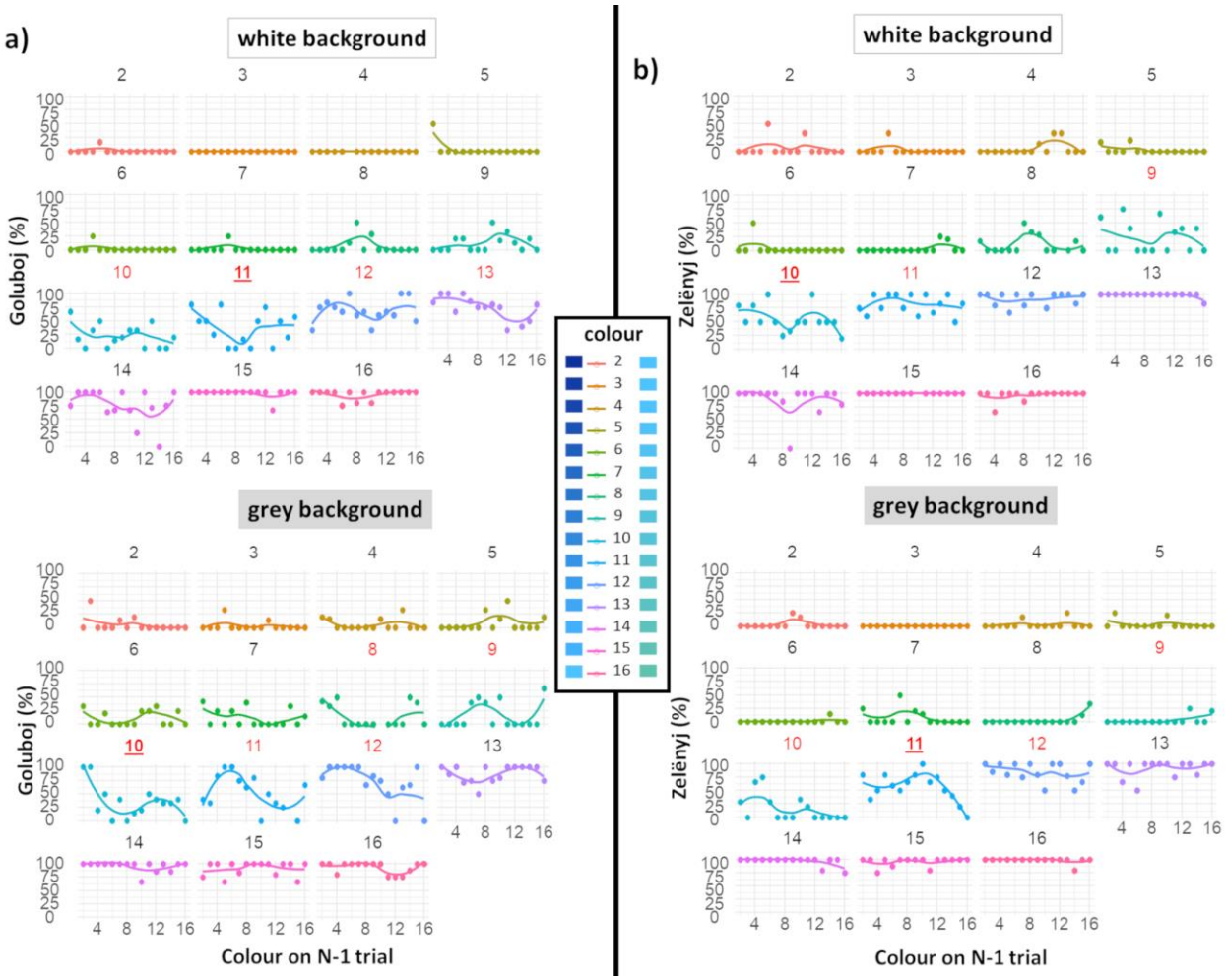


Figure 6. (N-1) effects for the categorisation into *sinij/goluboj* (a) and *goluboj/zelënyj* (b), representing the potential influence of the previous trial on categorisation (so-called order bias). Categories to which colours are assigned (*y*-axis) are depicted in relation to the colours from the previous (*N-1) trial (x*-axis) for each individual colour. Colours presented against a white background are on top and those presented against a grey background are on the bottom. Certain colours are uniformly categorised (0% or 100% on *y*-axis), however, some colours vary in terms of categorisation and this does not seem to be independent from previous trial's colour – if there were no effect of the preceding trial, the lines depicting the best-fitting curve would be roughly flat. The numbers of colours that fall within the JNDs (see **Fig. 5**) are highlighted by red, while the colour closest to the PSE is, in addition, underlined. Colours presented on the previous trial have a repulsive effect on categorisation of colours at the PSE and colours most proximal to the

PSE: e.g. when categorising into *sinij/goluboj*, if C10 on a grey background is preceded by a darker sample, it is more likely to be categorised as *goluboj*.

Discrimination xAB task. *Discrimination task data.* As in Experiment 1, we first excluded all incorrect responses (6% trials); we then excluded as outliers any RTs faster than 250 ms (only 3 trials) and longer than 1,500 ms (4% of correct trials). Once we selected only the trials that fell into the boundary or non-boundary conditions, we were left with a total of 9,423 trials (67% of the post-exclusion total). Since the discrimination pairs differed across the two colour sets (i.e. varying in lightness/chroma/hue in the *sinij/goluboj*, whilst varying in hue only in the *goluboj*/*zelënyj* condition), we fitted a separate LME model for each colour set, as outlined below. As in Experiment 1, if the boundary was too close to the darkest/lightest or bluest/greenest stimulus colour, the calculations of RTs for the cross- and within-category conditions were based on data from fewer than six colours; however, calculations for the cross-category condition were never based on data from fewer than two discrimination pairs and for the within-category condition never based on data from fewer than three discrimination pairs.

For RT data, we performed analysis using an LME model for each colour set (*sinij/goluboj* or *goluboj*/*zelënyj*) with the boundary (discriminations at or across the boundary vs. up to six nearest within-category discriminations) as a fixed effect. We also included variability of RTs across discrimination pairs (C2 vs. C4., C3 vs. C5, C4 vs. C6 and so on) and by-participant variability in intercept and slopes as random effects.

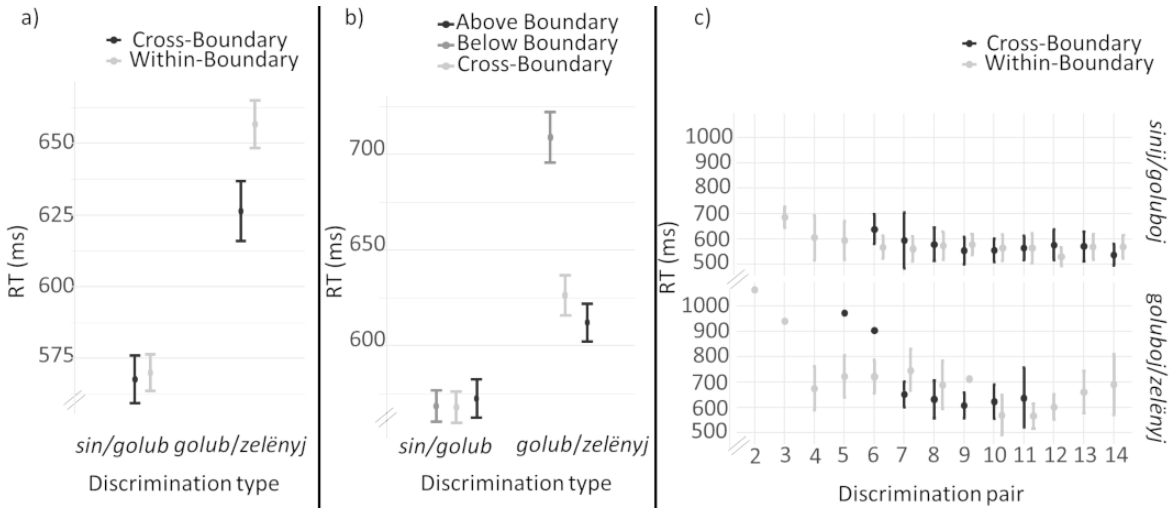


Figure 7. Reaction times in the colour discrimination (xAB) task for the two types of categorisation. (a) Mean RTs for cross-boundary (black) and within-boundary (grey) discriminations. (b) Mean RTs for cross-boundary (light grey), as well as below-boundary, darker blue/bluer (dark grey) and above-boundary, lighter blue/greener (black) discriminations. (c) Cross- and within-boundary categorisations are shown for each colour pair for *sinij/goluboj* (top right) and *goluboj/zelënyj* (bottom right) colour sets.

We found that for the *sinij/goluboj* categorisation, the model that included random slopes for RTs across discrimination pairs failed to converge, so we used a model with only by-participant random slopes and the fixed effect of the boundary. Mean intercept for this model at zero level, which was set as cross-boundary, was 567 ms (SE = 16 ms). Importantly, and contrary to previous research, removing the fixed effect of the boundary from the model did not alter the model fit ($\chi^2(1) = 0.17$, $p = .68$; with intercept estimated at 568 ms, SE = 16 ms), showing that discriminations at the boundary did not differ significantly from discriminations for nearby non-boundary colours. Inspection of **Fig. 7b-c** points at similar RTs between lighter and

darker pairs. This differs from Experiment 1 in which we observed shorter RTs for lighter pairs. The stimuli in Experiment 2 were darker (compare *Y*-values in **Tables 1** and **3**) and presented against a white rather than grey background, which implies that both darker and lighter pairs were considerably different from background luminance – unlike Experiment 1, in which lighter pairs were much closer to the luminance of the background.

For the *goluboj/zelënyj* categorisation, random slopes for RT differences across discrimination pairs ($\chi^2(1) = 267.4$, $p < .001$) did contribute significantly to the model. **Fig. 7c** shows that for *goluboj/zelënyj* there seems to be a difference between cross- and within-category conditions across discrimination pairs, with shorter RTs for cross-category discrimination pairs for a few colours (colours C7-C9). Further removing the fixed effect of the boundary from the model, however, did not improve the fit ($\chi^2(1) = 0.06$, $p = .81$). Mean intercept for the best fitting model with random effects only was 681 ms (SE = 34 ms). The classical boundary effect of processing speed which seems to be present in **Fig. 7a** is actually a by-product of longer RTs for bluer discrimination pairs, as is apparent in **Fig. 7b**. Shorter RTs in the *sinij/goluboj* categorisation compared to the *goluboj/zelënyj* categorisation are probably due to facilitation of processing because of differences along multiple dimensions in colour space, since in the *sinij/goluboj* series colours varied along lightness, saturation and hue, while in the *goluboj/zelënyj* series only hue varied (cf. Garner & Felfoldy, 1970).

We again performed a post hoc analysis of RT distributions; density and cumulative density functions are plotted in **Fig. 8**. For the *sinij/goluboj* series, the two distributions are not statistically different, as assessed by a two-sided Kolmogorov-Smirnov test ($D = 0.016$, $p = .95$), but they are significantly different for the *goluboj/zelënyj* series ($D = 0.064$, $p < .001$). Inspection

of **Fig. 8** informs us that these differences in RTs are persistent and emerge already for very fast responses.

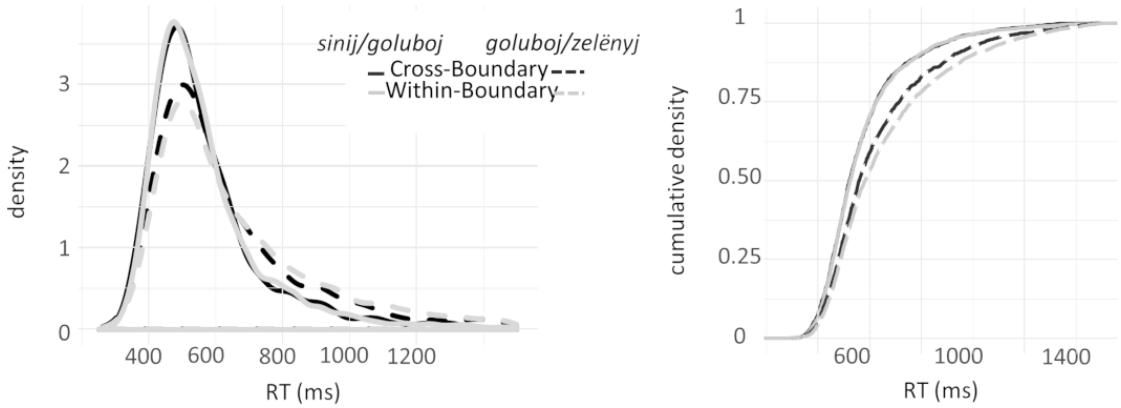


Figure 8. Densities (left) and cumulative densities (right) of RT distributions for *sinij/goluboj* and *goluboj/zelënyj* series. This graph contrasts cross- and within-boundary colours, showing their distributions across the full range of RTs.

4. Discussion

In the Experiment 1, we compared speakers of different languages on the same part of the colour space in which one language (Russian) has two BCCs, while the other language (English) has a single BCC, while in Experiment 2 we compared two categorical distinctions, a hue-based *goluboj/zelënyj* and a lightness-based *sinij/goluboj*, in a single Russian sample. We did not observe a boundary advantage in processing speed for *sinij/goluboj* in either of the two experiments – neither for monolinguals nor bi-/multilinguals. We did, however, find a difference in categorisation between *sinij/goluboj* in Russian speakers and *dark blue/light blue* in English-speaking controls: the transition of *sinij* to *goluboj* was somewhat sharper than the transition from *dark blue* to *light blue* respectively, which is characteristic of a firmer categorical

distinction. However, frequency biasing revealed a relative instability of the "Russian blues" boundary: in the "*sinij* bias" context, (medium) blue shades were more likely to be judged as *goluboj.* In some cases, similar effects also emerged in the "*goluboj* bias" context, with (medium) blue shades more likely to be judged as *sinij*. In contrast, the *goluboj*/*zelënyj* boundary remained robust against both *goluboj* and *zelënyj* bias contexts. These findings strongly suggest that although lightness-based *sinij* and *goluboj* categories function linguistically as BCCs in the Russian language, their boundary is not as firmly demarcated as between colour categories that are differentiated in terms of hue (here, *goluboj* and *zelënyj)*.

There may be several methodological sources of discrepancies between our and Winawer et al.'s results. As discussed already, first and foremost, their participants differed vastly from ours in response speed. Our RT analyses imply that the boundary effect for *sinij/goluboj* could potentially emerge for slower responses/responders but appears to be absent for faster responses/responders. In fact, Witzel and Gegenfurtner (2016) argue that red/brown RT category effects are driven by RTs for cross-boundary pairs being spread much less toward the upper end of the response time distribution than for within-category pairs. In a similar vein, Witzel and Gegenfurtner (2015) found that only an inexperienced group of participants, with mean RTs of ~750-850 ms, showed consistent categorical facilitation, unlike a trained group whose mean RTs were ~500-600 ms (see Supplementary Figs. 5, 6 in their paper).

Our Russian participants in Experiment 1 had mean RT over 700 ms but in Experiment 2, mean RT was closer to 500 ms. According to Witzel and Gegenfurtner (2015), categorical facilitation may be mediated by attention to the categorical distinction which does not occur when participants respond automatically based on sensory feed-forward information. However, neither of our experimental groups had much experience in performing colour discrimination

tasks and the mean RT of the Russian participants in Experiment 1 is closer to the inexperienced group in Witzel and Gegenfurtner's study. Thus, we doubt that overall response speed is sufficient to explain the lack of response speed facilitation in our experiments.

Winawer et al. observed a cross-category facilitation only for "near" comparisons (equivalent to those used in our study), but not for "far" comparisons, with twice as large colour distances. It is possible that in the context of "near" and "far" trials intermixed, the "near" trials were seen as subjectively more difficult which led Russian participants to increasingly attend to the *sinij/goluboj* categorical distinction. This could explain why Winawer et al. observe a RT advantage while we fail to do so. However, we do not believe that a mixture of "easy"/"far" and "hard"/"near" trials is a necessary prerequisite for observing categorical facilitation, as boundary RT effects for hue-based categories have been observed with "near" pairs alone (C. Witzel & Gegenfurtner, 2011, 2015).

Based on additional analyses of RTs in which we included outliers, there is a more parsimonious explanation: an RT difference resembling the classic boundary effect may also be generated through amplification of longer RTs for darker blue discriminations. This might have been particularly pronounced in the slower, Russian-speaking group. Our data certainly conform to this interpretation, as in Experiment 1 we observe longer RTs for darker blues, similarly to longer RTs for blues as opposed to greens, a perceptually driven RT effect that is commonly observed at the green/blue boundary (Witzel & Gegenfurtner, 2011, 2015). Comparison of RTs in Experiment 1 and Experiment 2 further emphasises the importance of low-level factors as a driver of RTs when discriminating colour pairs in terms of lightness: RTs are dependent on the difference between luminance of colour samples and luminance of the background, with

responses being faster for pairs closer to the background (as would be predicted by the Weber ratio).

Compared to the stimuli employed here, Winawer et al.'s stimulus set was slightly extended, including one additional darker and four additional lighter shades of blue (see **Fig. 1**). Extending the stimulus range can shift the mean (Parducci & Perrett, 1971), and in the present case it is likely to have shifted the *sinij/goluboj* boundary, which we show to be highly susceptible to both frequency and order biases. For the 15 colours from Winawer et al. that were used in our Experiment 1, we observe slower responses to darker blue stimuli as opposed to lighter blues, which are closer to the luminance of the background. Blue/green boundary effects observed in previous studies suffer from a similar non-categorical RT facilitation driven by RTs for bluer colours being longer than for greener colours, as observed by Witzel and Gegenfurtner (C. Witzel & Gegenfurtner, 2011, 2015) and in our own Experiment 2 (compare **Fig. 7a** and **7b**). Based on this evidence, we conclude that the emergence of an observable RT advantage for cross-category colours in Winawer et al.'s study might have been an artifact of the non-uniform RT distributions driven by discriminations of darker vs. lighter colour pairs, further impacted by considerably longer overall RTs in the Russian group.

We did, however, find a difference between *sinij/goluboj* in Russian speakers and *dark blue/light blue* in controls: as can be seen in **Fig. 2**, the transition from *sinij* to *goluboj* is somewhat sharper, which is characteristic of a firmer categorical distinction. Effects of bias also seemed to affect discrimination of *dark blue* shades in English speakers much more relative to *sinij* shades in Russian speakers. As mentioned earlier, *sinij* is the more chromatic of the two "Russian blues" categories. As demonstrated by Witzel (2016), speakers' naming consensus as well as nameability of colours significantly increase with saturation. This implies that being

more chromatic, *sinij*-colours are expected to be named with greater consensus than *goluboj*-colours. The greater consensus, in turn, implies greater stability of the category boundary (cf. Fider & Komarova, 2018). As an example of the pragmatic impact of this difference in colourfulness of the two colour terms for blue, Russian speakers are more likely to use *seryj* 'grey' when describing light blue eyes compared to English speakers (Lowry & Bryant, 2019). Conversely, *goluboj* is also a term used to describe greyness of cat fur or pigeon feathers, which are both largely achromatic. Perhaps the more achromatic nature and the more diluted hue of *goluboj* makes it less structured.

Frequency and order biasing effects (**Figs. 5, 6**) show differences between hue-based categories and "Russian blues" categories in our monolingual sample. We found both attractive and repulsive sequential effects of context on colour categorisation. In the current literature, a prominent topic concerns the degree to which, at a given moment, perception of various visual attributes is affected by the preceding stimulus. These serial dependencies can be both attractive and repulsive (Alais, Leung, & Van der Burg, 2017). Interestingly, there are pronounced attractive serial dependencies for oblique spatial orientations, but not for cardinal orientations (see Fig. 1 in Cicchini, Mikellidou, & Burr, 2017). Cardinal orientations (vertical, horizontal) represent categorical reference points when judging line orientation (Rosch, 1975). Similarly, we observe a frequency bias effect for *sinij/goluboj*, i.e. the lightness-based categorical distinction, but not for *goluboj*/*zelënyj*, a hue-based distinction. It appears that the hue-based categorical distinction between *goluboj* and *zelënyj* is firmer and more clearly delimited than the distinction between *sinij* and *goluboj*.

Based on these outcomes, we conclude that hue-based basic categories may be better structured, i.e. have higher category strength and better demarcated boundaries (cf. Fider &

Komarova, 2018) than lightness-based *sinij* and *goluboj*. Combining the frequency effects (Parducci & Perrett, 1971; Parducci & Wedell, 1986) and the recently emerging exploration of serial dependencies would be promising and productive. It is likely that a hitherto unexplored but important aspect of categorical facilitation is its ability to minimise serial dependencies and range/frequency effects by providing the observer with a template that increases the fidelity of stimulus encoding. A strong categorical distinction between sensory stimuli should lead to their perception being robust to contextual influences. In our case, this holds for the hue-based boundary but not for the lightness-based boundary. In fact, colour category structuredness, quantified as its stability in relation to contextual influence, might provide a measure of category strength and serve to establish an objective hierarchy of demarcations between basic as well as non-basic colour categories at various lightness levels.

Some might argue that our approach of collapsing categorisation data across participants when examining biases is not valid, as boundary effects are highly individual – this is certainly the approach taken by some researchers (e.g. Emery, Volbrecht, Peterzell, & Webster, 2017; Webster & Kay, 2012). Indeed, inspection of individual categorisation data, which we present in Supplementary **Figs. S1 and S3**, reveals that there is richness in individual categorisation datasets that is lost by reducing them to a single boundary value. This is particularly obvious for Experiment 2, where we record multiple responses for each colour and observe that for many participants the categorical transition is spread across several neighbouring colours. In fact, (*N-1*) effects derived from these data reveal that especially for *sinij/goluboj* but also for *goluboj*/*zelënyj*, colours at the group-derived PSE and around it seem to elicit less consistent responses, influenced by a repulsive bias from the previous trial (**Fig. 6**). We conclude that preceeding trial effects (i.e. (*N-1*) effects) seem to be a very promising vehicle to objectively

assess a boundary: they have been used for decades in studies of how templates in working memory affect attention (e.g. the dimension-weighting model; Found & Muller, 1996). If categorical facilitation is mainly driven by attentional allocation, as theorised by Witzel and Gegenfurtner (2018), this may make (*N-1*) effects very suitable to further interrogate categorical effects on perception.

Russian *sinij* and *goluboj* are not the only blue lightness-based basic colour categories: the present findings could be validated by testing the category boundary effect for other languages with two "blues" (e.g. *blu* and *azzurro* in Italian or *ao* and *mizu* in Japanese, designating, respectively, 'dark blue' and 'light blue'; for a review see Paramei & Bimler, 2020). *Teget* 'dark blue' and *bordo* 'dark red' are two frequent and salient non-BCCs in the Serbian language, segregating blue and red sub-areas of the colour space based on lightness. Jakovljev and Zdravkovic (2018) demonstrated cross-category advantages in a speeded discrimination task for *teget*/ 'blue' and *bordo*/'red'. Interestingly, the colours in that study were much darker than in experiments reported in this paper, as both *teget* and *bordo* refer to particularly dark shades of blue and red: blue stimuli ranged between 0.43–3.96 cd/m^2 while red stimuli ranged between 1.04–10.84 cd/m^2. As these are non-BCCs and Jakovljev and Zdravkovic (2018) did not test a control group whose language does not have counterparts of such colour terms, it may be that these RT advantages are specific to certain (sub-)areas of colour space irrespective of the language or irrespective of whether the colour terms that divide these areas of colour space are basic or non-basic.

Event-related potential (ERP) studies of colour categorisation (e.g. Maier & Rahman, 2018; Thierry, Athanasopoulos, Wiggett, Dering, & Kuipers, 2009; for a review see C. Witzel & Gegenfurtner, 2018) are often taken as robust evidence of categorical facilitation. The study of

Thierry et al. (2009), in particular, has been taken as firm evidence of an early neural locus of the two "Greek blues" effect, similar to the "Russian blues" effect observed by Winawer et al.

However, the evidence provided by these ERP studies is not as convincing as it might seem at first sight. Some major caveats are already discussed in the recent review by Witzel and Gegenfurtner (2018). There are also further issues which are well-known to those with ERP expertise – namely, these studies commonly report very small between-subjects effects (e.g. in Thierry et al., the crucial three-way interaction has effect size $\mu_p^2 = 0.112$, which makes its CI using non-central F distribution 0.004-0.273; calculated as per Uanhoro, 2017).

Furthermore, these low-powered, small effect-size outcomes are observed from data taken from narrow temporal windows and more-or-less arbitrarily chosen electrode sites, which introduces unnecessary researcher degrees of freedom (Gorgolewski & Poldrack, 2016; Luck & Gaspelin, 2017). The choice of ERP components and their temporal windows/electrode sites is rarely theoretically motivated beyond an argument that the perceptual vs. cognitive locus of categorical effects can be determined from whether early or late parts of the ERP waveform are modulated, whereby early parts are often taken to imply pre-attentive modulation in spite of the fact that it is well-known that attention can modulate even the earliest ERP components (e.g. Erlbeck, Kubler, Kotchoubey, & Veser, 2014; Handy & Mangun, 2000). A re-analysis of existing ERP studies using more advanced, state-of-the-art approaches that are blind to time-window or electrode-site selection (e.g. the mass-univariate-analysis approach; Pernet, Chauveau, Gaspar, & Rousselet, 2011) would be able to show if the observed ERP effects are in fact sufficiently robust to represent a meaningful contribution to the literature.

To conclude, contrary to previous reports, our results imply that in Russian the lightness-defined blue basic colour categories differ behaviourally from the hue-based basic colour

categories: *sinij/goluboj* category boundary is more context-sensitive and does not reliably manifest speeded discrimination differences. Firestone and Scholl (2016) suggest several potential pitfalls for studies that aim to assess whether cognitive factors exert a direct top-down influence on perception. Our experiments, together with previous work (Danilova & Mollon, 2014; Roberson et al., 2009; Roberson, Pak, & Hanley, 2008; C. Witzel & Gegenfurtner, 2011, 2013, 2015, 2018) demonstrate how important it is to disambiguate low-level differences and high-level effects when studying colour categorisation effects –in terms of the thorough stimulus control and the presentation context, as well as comprehensive data analysis that take into account both between-colours and between-participants differences.

Our findings have broad implications – conceptually, for studies of Whorfian effects and colour categorisation, and methodologically, for reintroducing the range/frequency effects, (*N-1*) effects, and for cumulative RT distribution analyses as highly promising tools to study categorical facilitation. The present findings provide further evidence of the transient nature of the influence of language on perception, where visual decisions may or may not be augmented by top-down modulation (Firestone & Scholl, 2016; O'Callaghan et al., 2017): effects of language on perception are stronger in the tasks that promote categorisation (e.g. the categorisation task, where we observe a sharper transition in Russians' categorisation compared to controls) and weaker or non-existent in the tasks that do not explicitly require it but rather incite a discrimination judgement (i.e. the xAB task; cf. Lupyan, 2012; Webster & Kay, 2012; Christoph Witzel, 2018). This defies a simplified model, in which basic colour categories by default enact a powerful influence on perceptual processes.

Acknowledgements

JM's work was partly supported by the DAAD re-invitation programme for former scholarship holders. The authors are grateful to Alexander Belokopytov and Galina I. Rozhkova for their generous colorimetric assistance and to Karin Tuur, Daria Tusheva and Carl Wollter for help with data collection. JM would like to thank Agnieszka Konopka and Matt Craddock for advice on data analysis and plotting in R and Christoph Witzel for suggesting the analyses of RT distributions. Finally, we would like to thank Jonathan Winawer, two other, anonymous, reviewers, and Jacob Feldman, Associate Editor, for constructive comments that greatly helped in improving our work.

Supplementary Materials

Experiment 2 was pre-registered on OSF. All the data and analysis procedures for both experiments are also publicly available on OSF (https://osf.io/5m4h7/).

We also include 3 supplementary materials: 1) Individual plots of categorisation data from Experiment 1; 2) Analysis of accuracy data from Experiment 1; 3) Individual plots of categorisation data from Experiment 2.

CrediT Authors Statement

Jasna Martinovic: conceptualisation, methodology, validation, formal analysis, investigation, resources, data curation, writing – original draft, writing – review & editing, visualisation, supervision, project administration, funding acquisition

Galina V Paramei: conceptualisation, methodology, writing – review & editing

W. Joseph MacInnes: validation, formal analysis, investigation, resources, writing – review & editing, visualisation, supervision

References

Agrillo, C., & Roberson, D. (2009). Colour language and colour cognition: Brown and Lenneberg revisited. *Visual Cognition, 17*(3), 412-430. doi:10.1080/13506280802049247

Alais, D., Leung, J., & Van der Burg, E. (2017). Linear Summation of Repulsive and Attractive Serial Dependencies: Orientation and Motion Dependencies Sum in Motion Perception. *Journal of Neuroscience, 37*(16), 4381-4390. doi:10.1523/jneurosci.4601-15.2017

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:doi:10.18637/jss.v067.i01.

Berlin, B., & Kay, P. (1969/1991). *Basic Color Terms: Their Universality and Evolution*. Berkley, CA: Univ of California Press.

Bimler, D., & Uuskula, M. (2017). A Similarity- Based Cross-Language Comparison of Basicness and Demarcation of "Blue" Terms. *Color Research and Application, 42*(3), 362-377. doi:10.1002/col.22076

Bimler, D. L., Paramei, G. V., & Izmailov, C. A. (2009). Hue and saturation shifts from spatially induced blackness. *Journal of the Optical Society of America a-Optics Image Science and Vision, 26*(1), 163-172.

Bornstein, M. H., & Korda, N. O. (1984). Discrimination and matching within and between hues measured by reaction-times - some implications for categorical perception and levels of information-processing. *Psychological Research-Psychologische Forschung, 46*(3), 207-222. doi:10.1007/bf00308884

Bornstein, M. H., & Monroe, M. D. (1980). Chromatic information-processing - rate depends on stimulus location in the category and psychological complexity. *Psychological Research-Psychologische Forschung, 42*(3), 213-225. doi:10.1007/bf00308529

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433-436. doi:10.1163/156856897x00357

Brown, A. M., Lindsey, D. T., & Guckes, K. M. (2011). Color names, color categories, and color-cued visual search: Sometimes, color perception is not categorical. *Journal of Vision, 11*(12), 21. doi:10.1167/11.12.2

Cicchini, G. M., Mikellidou, K., & Burr, D. (2017). Serial dependencies act directly on perception. *Journal of Vision, 17*(14), 9. doi:10.1167/17.14.6

Cornsweet, T. N., & Pinsker, H. M. (1965). Luminance discrimination of brief flashes under various conditions of adaptation. *The Journal of Physiology, 176*(2), 294-310. doi:10.1113/jphysiol.1965.sp007551

Cropper, S. J., Kvansakul, J. G. S., & Little, D. R. (2013). The categorisation of non-categorical colours: a novel paradigm in colour perception. *PLoS ONE, 8*(3), e59945-e59945. doi:10.1371/journal.pone.0059945

Danilova, M. V., & Mollon, J. D. (2014). Is discrimination enhanced at the boundaries of perceptual categories? A negative case. *Proceedings of the Royal Society B-Biological Sciences, 281*(1785), 8. doi:10.1098/rspb.2014.0367

Davidoff, J. (2015). *Color categorization across cultures*. Cambridge: Cambridge Univ Press.

Davies, I., & Corbett, G. (1994). The basic color terms of Russian. *Linguistics, 32*(1), 65-89. doi:10.1515/ling.1994.32.1.65

Douven, I., Wenmackers, S., Jraissati, Y., & Decock, L. (2017). Measuring Graded Membership: The Case of Color. *Cognitive Science, 41*(3), 686-722. doi:10.1111/cogs.12359

Drivonikou, G. V., Kay, P., Regier, T., Ivry, R. B., Gilbert, A. L., Franklin, A., & Davies, I. R. L. (2007). Further evidence that Whorfian effects are stronger in the right visual field than the left. *Proceedings of the National Academy of Sciences of the United States of America, 104*(3), 1097-1102. doi:10.1073/pnas.0610132104

Emery, K. J., Volbrecht, V. J., Peterzell, D. H., & Webster, M. A. (2017). Variations in normal color vision. VII. Relationships between color naming and hue scaling. *Vision Research, 141*, 66-75. doi:10.1016/j.visres.2016.12.007

Erlbeck, H., Kubler, A., Kotchoubey, B., & Veser, S. (2014). Task instructions modulate the attentional mode affecting the auditory MMN and the semantic N400. *Frontiers in Human Neuroscience, 8*, 16. doi:10.3389/fnhum.2014.00654

Fider, N., & Komarova, N. L. (2018). Quantitative study of color category boundaries. *Journal of the Optical Society of America a-Optics Image Science and Vision, 35*(4), B165-B183. doi:10.1364/josaa.35.00b165

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences, 39*, e229. doi:10.1017/s0140525x15000965

Fletcher, R. (1975). *The City University Colour Vision Test (1st edition)*. Windsor: Keeler.

Found, A., & Muller, H. J. (1996). Searching for unknown feature targets on more than one dimension: Investigating a ''dimension-weighting'' account. *Perception & Psychophysics, 58*(1), 88-101. doi:10.3758/bf03205479

Garner, W. R., & Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology, 1*(3), 225-241. doi:https://doi.org/10.1016/0010-0285(70)90016-2

Gilbert, A. L., Regier, T., Kay, P., & Ivry, R. B. (2006). Whorf hypothesis is supported in the right visual field but not the left. *Proceedings of the National Academy of Sciences of the United States of America, 103*(2), 489-494. doi:10.1073/pnas.0509868103

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *WIREs Cognitive Science, 1*(1), 69-78. doi:10.1002/wcs.26

Gorgolewski, K. J., & Poldrack, R. A. (2016). A Practical Guide for Improving Transparency and Reproducibility in Neuroimaging Research. *PLOS Biology, 14*(7), 13. doi:10.1371/journal.pbio.1002506

Griebel, U., & Schmid, A. (1997). Brightness discrimination ability in the West Indian manatee (Trichechus manatus). *Journal of Experimental Biology, 200*(Pt 11), 1587-1592.

Hampton, J. (1998). Similarity-Based Categorization and Fuzziness of Natural Categories. *Cognition, 65*, 137-165. doi:10.1016/S0010-0277(97)00042-5

Handy, T. C., & Mangun, G. R. (2000). Attention and spatial selection: Electrophysiological evidence for modulation by perceptual load. *Perception & Psychophysics, 62*(1), 175-186. doi:10.3758/bf03212070

Hanley, J. R. (2016). Color categorical perception. In M. R. Luo (Ed.), *Encyclopedia of Color Science and Technology*. New York: Springer.

Harnad, S. (1987). Psychophysical and cognitive aspects of categorical perception: A critical overview. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 1-25). Cambridge: Cambridge University Press.

Hope, R. M. (2013). Rmisc: Ryan Miscellaneous. R package version 1.5. Retrieved from https://CRAN.R-project.org/package=Rmisc

Huette, S., & McMurray, B. (2010). Continuous dynamics of color categorization. *Psychonomic Bulletin & Review, 17*(3), 348-354. doi:10.3758/pbr.17.3.348

Jaeger, T. F. (2008). Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of memory and language, 59*(4), 434-446. doi:10.1016/j.jml.2007.11.007

Jakovljev, I., & Zdravkovic, S. (2018). The colour lexicon of the Serbian language - a study of dark blue and dark red colour categories Part 2: Categorical facilitation with Serbian colour terms. *Psihologija, 51*(3), 289-308.

Jraissati, Y., Wakui, E., Decock, L., & Douven, I. (2012). Constraints on Colour Category Formation. *International Studies in the Philosophy of Science, 26*(2), 171-196. doi:10.1080/02698595.2012.703479

Kay, P. (2015). Universality of color categorization. *Handbook of Color Psychology*, 245-258.

Kay, P., & Maffi, L. (1999). Color appearance and the emergence and evolution of basic color lexicons. *American Anthropologist, 101*(4), 743-760. doi:10.1525/aa.1999.101.4.743

Kliegl, R., Masson, M. E. J., & Richter, E. M. (2010). A linear mixed model analysis of masked repetition priming. *Visual Cognition, 18*(5), 655-681. doi:10.1080/13506280902986058

Knoblauch, K., & Maloney, L. T. (2012). *Modelling Psychophysical Data in R*: Springer.

Kroll, J. F., Bobb, S. C., Misra, M., & Guo, T. (2008). Language selection in bilingual speech: evidence for inhibitory processes. *Acta Psychologica, 128*(3), 416-430. doi:10.1016/j.actpsy.2008.02.001

Laws, G., Davies, I., & Andrews, C. (1995). Linguistic structure and nonlinguistic cognition - English and Russian blues compared. *Language and Cognitive Processes, 10*(1), 59-94. doi:10.1080/01690969508407088

Lenth, R. V., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2019). emmeans: Estimated Marginal Means, aka Least Squares Means (Version 1.4.3.01). .

Lindsey, D. T., Brown, A. M., Reijnen, E., Rich, A. N., Kuzmova, Y. I., & Wolfe, J. M. (2010). Color Channels, Not Color Appearance or Color Categories, Guide Visual Search for Desaturated Color Targets. *Psychological Science, 21*(9), 1208-1214. doi:10.1177/0956797610379861

Lowry, M., & Bryant, J. (2019). Blue is in the Eye of the Beholder: A Cross-Linguistic Study on Color Perception and Memory. *Journal of Psycholinguistic Research, 48*(1), 163-179. doi:10.1007/s10936-018-9597-0

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology, 54*(1), 146-157. doi:10.1111/psyp.12639

Lupyan, G. (2012). Linguistically modulated perception label-feedback hypothesis. *Frontiers in psychology, 3*, 13. doi:10.3389/fpsyg.2012.00054

Maier, M., & Rahman, R. A. (2018). Native Language Promotes Access to Visual Consciousness. *Psychological Science, 29*(11), 1757-1772. doi:10.1177/0956797618782181

Martinovic, J., Mordal, J., & Wuerger, S. M. (2011). Event-related potentials reveal an early advantage for luminance contours in the processing of objects. *Journal of Vision, 11*(7). doi:1

10.1167/11.7.1

Metha, A. B., Vingrys, A. J., & Badcock, D. R. (1993). Calibration of a color monitor for visual psychophysics. *Behavior Research Methods Instruments & Computers, 25*(3), 371-383. doi:10.3758/bf03204528

Mollon, J. D., & Cavonius, C. R. (1986). THE DISCRIMINABILITY OF COLORS ON CRT DISPLAYS. *Journal of the Institution of Electronic and Radio Engineers, 56*(3), 107-110.

O'Callaghan, C., Kveraga, K., Shine, J. M., Adams, R. B., Jr., & Bar, M. (2017). Predictions penetrate perception: Converging insights from brain, behaviour and disorder. *Consciousness and Cognition, 47*, 63-74. doi:10.1016/j.concog.2016.05.003

Paramei, G. V. (2005). Singing the Russian blues: An argument for culturally basic color terms. *Cross-Cultural Research, 39*(1), 10-38. doi:10.1177/1069397104267888

Paramei, G. V. (2007). Russian "blues": Controversies of basicness. . In R. E. MacLaury, G. V. Paramei, & D. Dedrick (Eds.), *Anthropology of color: Interdisciplinary multilevel modeling* (pp. 75-106). Amsterdam/Philadelphia: John Benjamins.

Paramei, G. V., & Bimler, D. (2020). Language and psychology. In A. Stenvall & S. Street (Eds.), *A cultural history of colour: The Modern Age*. London: Bloomsberry.

Paramei, G. V., Griber, Y. A., & Mylonas, D. (2017). An online color naming experiment in Russian using Munsell color samples. *Color Research and Application*. doi: https://doi.org/10.1002/col.22190

Parducci, A., & Perrett, L. F. (1971). Category rating scales - Effects of relative spacing and frequency of stimulus values. *Journal of Experimental Psychology, 89*(2), 427-&. doi:10.1037/h0031258

Parducci, A., & Wedell, D. H. (1986). The category effect with rating-scales - number of categories, number of stimuli, and method of presentation. *Journal of Experimental Psychology-Human Perception and Performance, 12*(4), 496-516. doi:10.1037/0096-1523.12.4.496

Pernet, C. R., Chauveau, N., Gaspar, C., & Rousselet, G. A. (2011). LIMO EEG: A Toolbox for Hierarchical LInear MOdeling of ElectroEncephaloGraphic Data. *Computational Intelligence and Neuroscience*. doi:831409

10.1155/2011/831409

R_Core_Team. (2016). R: A language and environment for statisticall computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Roberson, D. (2005). Color categories are culturally diverse in cognition as well as in language. *Cross-Cultural Research, 39*(1), 56-71. doi:10.1177/1069397104267890

Roberson, D., Hanley, J. R., & Pak, H. (2009). Thresholds for color discrimination in English and Korean speakers. *Cognition, 112*(3), 482-487. doi:10.1016/j.cognition.2009.06.008

Roberson, D., Pak, H., & Hanley, J. R. (2008). Categorical perception of colour in the left and right visual field is verbally mediated: Evidence from Korean. *Cognition, 107*(2), 752-762. doi:10.1016/j.cognition.2007.09.001

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology, 7*(4), 532-547. doi:https://doi.org/10.1016/0010-0285(75)90021-3

Safuanova, O. V., & Korzh, N. N. (2007). *Russian color names Mapping into a perceptual color space*. Amsterdam Me: John Benjamins B V Publ.

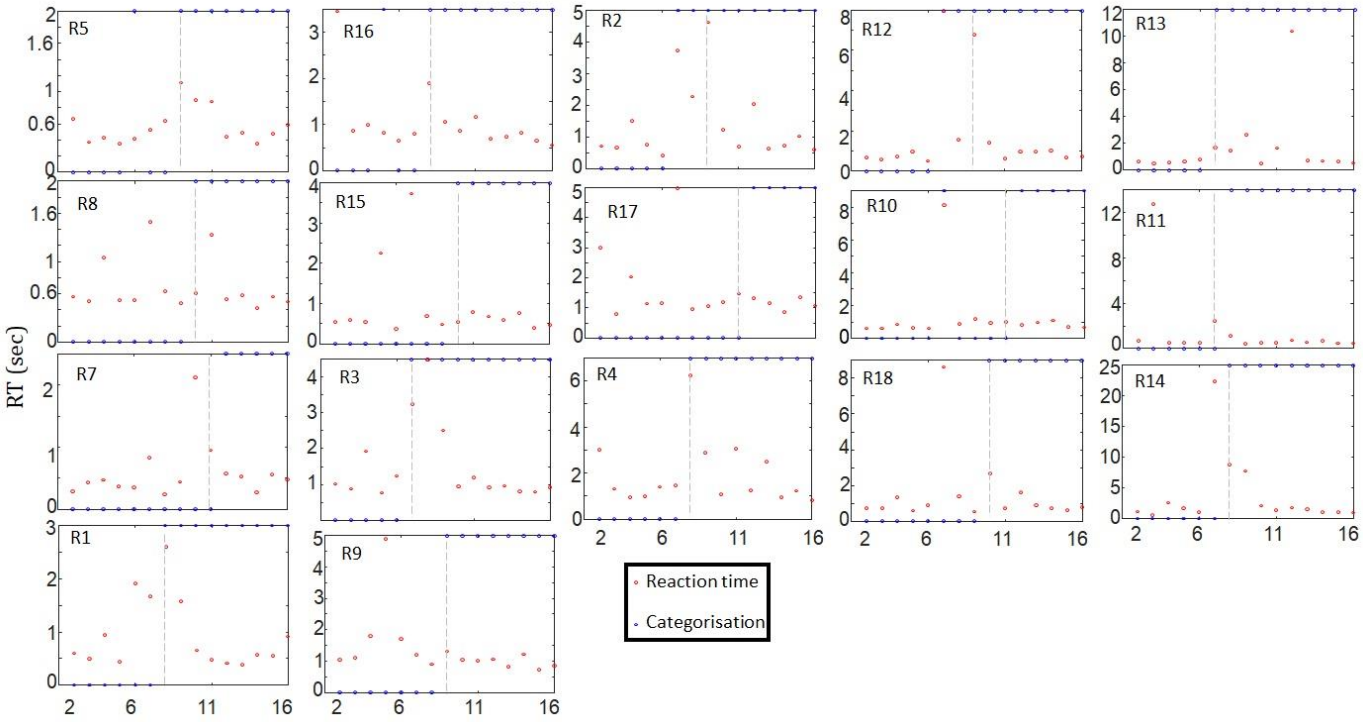
Saunders, B. A. C., & vanBrakel, J. (1997). Are there nontrivial constraints on colour categorization? *Behavioral and Brain Sciences, 20*(2), 167-&.

Suegami, T., Aminihajibashi, S., & Laeng, B. (2014). Another look at category effects on colour perception and their left hemispheric lateralisation: no evidence from a colour identification task. *Cognitive Processing, 15*(2), 217-226. doi:10.1007/s10339-013-0595-8

Thierry, G., Athanasopoulos, P., Wiggett, A., Dering, B., & Kuipers, J. R. (2009). Unconscious effects of language-specific terminology on preattentive color perception. *Proceedings of the National*

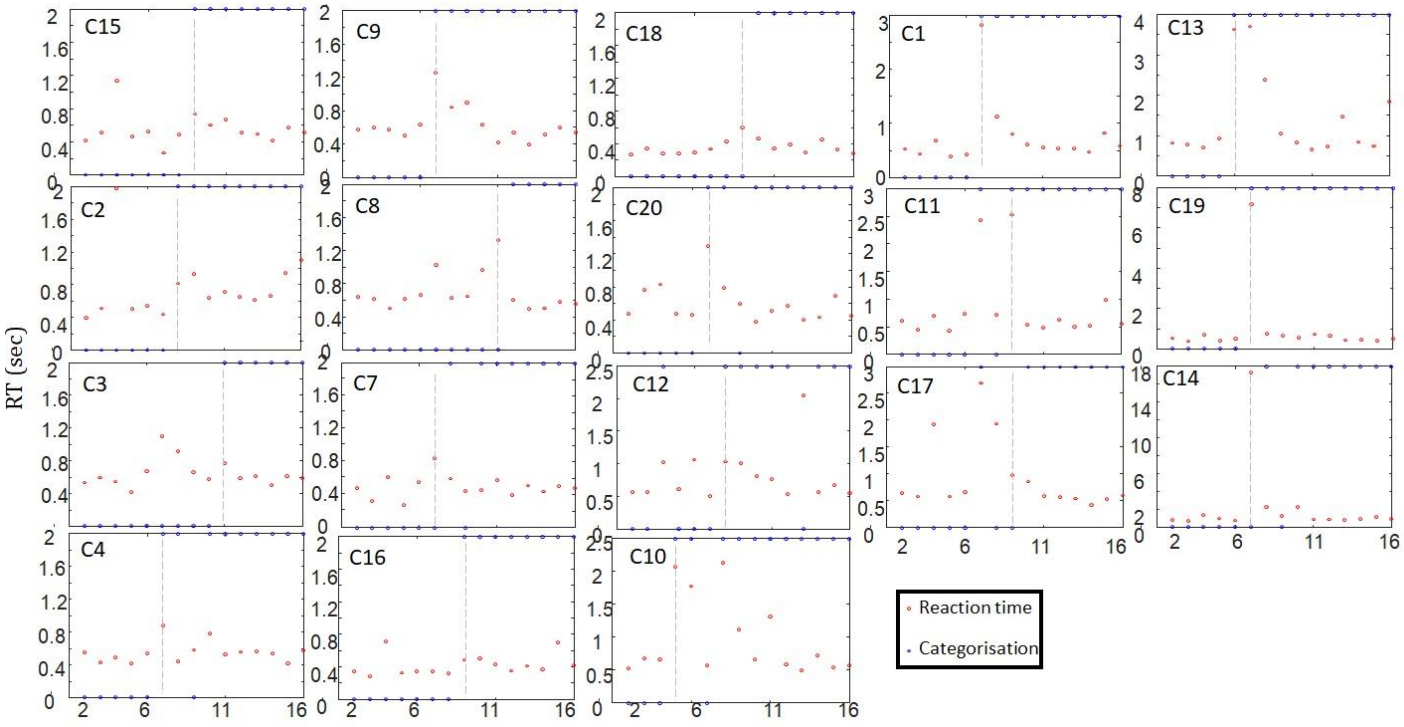
Academy of Sciences of the United States of America, 106(11), 4567-4570. doi:10.1073/pnas.0811155106
Uanhoro, J. O. (2017). Effect size calculators. Retrieved from https://effect-size-calculator.herokuapp.com/.
Warnes, G. R., Bolker, B., & Lumley, T. (2015). gtools: Various R Programming Tools. R package version 3.5.0. Retrieved from https://CRAN.R-project.org/package=gtools
Webster, M. A., & Kay, P. (2012). Color categories and color appearance. *Cognition, 122*(3), 375-392. doi:10.1016/j.cognition.2011.11.008
Wenseleers, T. (2016). export: Convert R Graphics and Statistical Output to Microsoft Office / LibreOffice, HTML and Latex. R Package version 0.2.1.
Westland, S., Ripamonti, C., & Cheung, V. (2012). *Computational Colour Science Using MATLAB 2nd Edition*: John Wiley & Sons.
Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.
Wickham, H., Francois, R., Henry, L., & Mueller, K. (2017). dplyr: A Grammar of Data Manipulation. R package version 0.7.2. Retrieved from https://CRAN.R-project.org/package=dplyr
Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences of the United States of America, 104*(19), 7780-7785. doi:10.1073/pnas.0701644104
Witzel, C. (2016). New Insights Into the Evolution of Color Terms or an Effect of Saturation? *i-Perception, 7*(5), 4. doi:10.1177/2041669516662040
Witzel, C. (2018). Misconceptions About Colour Categories. *Review of Philosophy and Psychology*. doi:10.1007/s13164-018-0404-5
Witzel, C., & Gegenfurtner, K. R. (2011). Is there a lateralized category effect for color? *Journal of Vision, 11*(12).
Witzel, C., & Gegenfurtner, K. R. (2013). Categorical sensitivity to color differences. *Journal of Vision, 13*(7), 33. doi:10.1167/13.7.1
Witzel, C., & Gegenfurtner, K. R. (2015). Categorical facilitation with equally discriminable colors. *Journal of Vision, 15*(8), 33. doi:10.1167/15.8.22
Witzel, C., & Gegenfurtner, K. R. (2016). Categorical Perception for Red and Brown. *Journal of Experimental Psychology-Human Perception and Performance, 42*(4), 540-570. doi:10.1037/xhp0000154
Witzel, C., & Gegenfurtner, K. R. (2018). Color Perception: Objects, Constancy, and Categories. In J. A. Movshon & B. A. Wandell (Eds.), *Annual Review of Vision Science, Vol 4* (Vol. 4, pp. 475-499). Palo Alto: Annual Reviews.
Yu, Z., & Schwieter, J. W. (2018). Recognizing the Effects of Language Mode on the Cognitive Advantages of Bilingualism. *Frontiers in psychology, 9*(366). doi:10.3389/fpsyg.2018.00366

Supplementary Figure 1a.



Categorisation of colours for individual Russian-speaking participants from Experiment 1. *Colours (2-16, see Fig.1) are depicted on the x axis, with sinij categorisations plotted as blue dots on the bottom and goluboj categorisations as blue dots on the top of the graphs. Note that the y axis, depicting reaction times, is differently scaled between participants, with the individual plots ordered from faster to slower overall responders. The boundary is indicated by a dashed grey line.*

Supplementary Figure 1b.



Categorisation of colours for individual non-Russian-speaking controls in Experiment 1. *Colours (2-16, see Fig.1) are depicted on the x axis, with sinij categorisations plotted as blue dots on the bottom and goluboj categorisations as blue dots on the top of the graphs. Note that the y axis, depicting reaction times, is differently scaled between participants, with the individual plots ordered from faster to slower overall responders. The boundary is indicated by a dashed grey line.*

Supplementary Material 2:

Analysis of Accuracies in Experiment 1

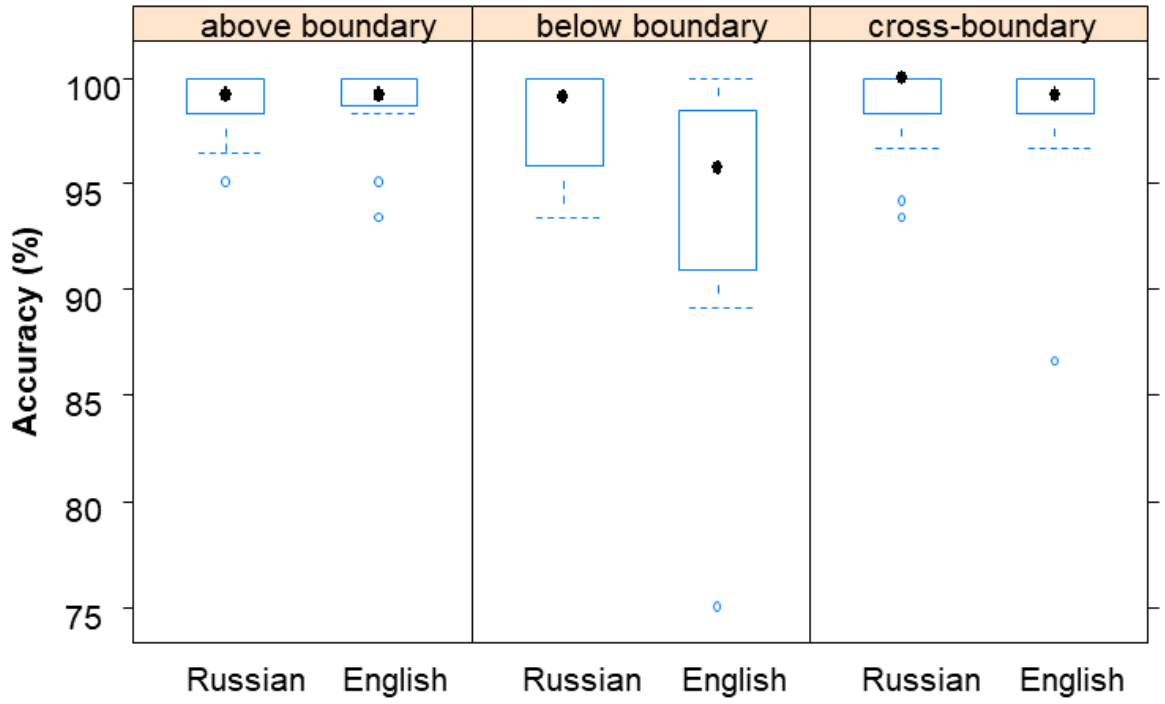
We applied a mixed logit model on our accuracy data, with Language (Russian or English) and Boundary (across or within) as fixed effect factors and random intercepts and slopes for participants and discrimination pairs. Random effects for discrimination pairs contributed significantly to the model ($\chi^2(1) = 54.40$, $p < .001$). The interaction did not contribute significantly ($\chi^2(1) = 2.13$, $p = .14$) and neither did Language ($\chi^2(1) = 4.46$, $p = .11$). To further explore accuracy data, we split the within-boundary pairs to those below and above the boundary (i.e. darker and lighter pairs). With the separation of darker and lighter within boundary pairs, the random effect of RT variability across discrimination pairs was no longer contributing to the model ($\chi^2(1) = 0.43$, $p = .51$), implying that differences in RTs were fully captured by separating the boundary variable into darker, lighter and cross-boundary pairs. The interaction of Language and Boundary was now highly significant ($\chi^2(1) = 13.67$, $p < .001$). We used emmeans package (Russel, 2019) to perform post-hoc tests on this interaction, using Tukey's HSD to correct for multiple comparisons.

These are presented in the Table below, with significant differences in bold:

Statistical comparison	Estimate of difference (log units)	SE of the estimate	z ratio	p value
darker pairs, English – cross-boundary, English	**-1.305**	**0.198**	**-6.576**	**<.001**
darker pairs, English – lighter pairs, English	**-1.798**	**0.234**	**-7.697**	**<.001**
darker pairs, English – darker pairs, Russian	-1.111	0.402	-2.762	0.06
darker pairs, English – cross-boundary, Russian	**-1.6571**	**0.421**	**-3.932**	**0.001**
darker pairs, English – lighter pairs, Russian	**-1.6203**	**0.420**	**-3.862**	**0.002**

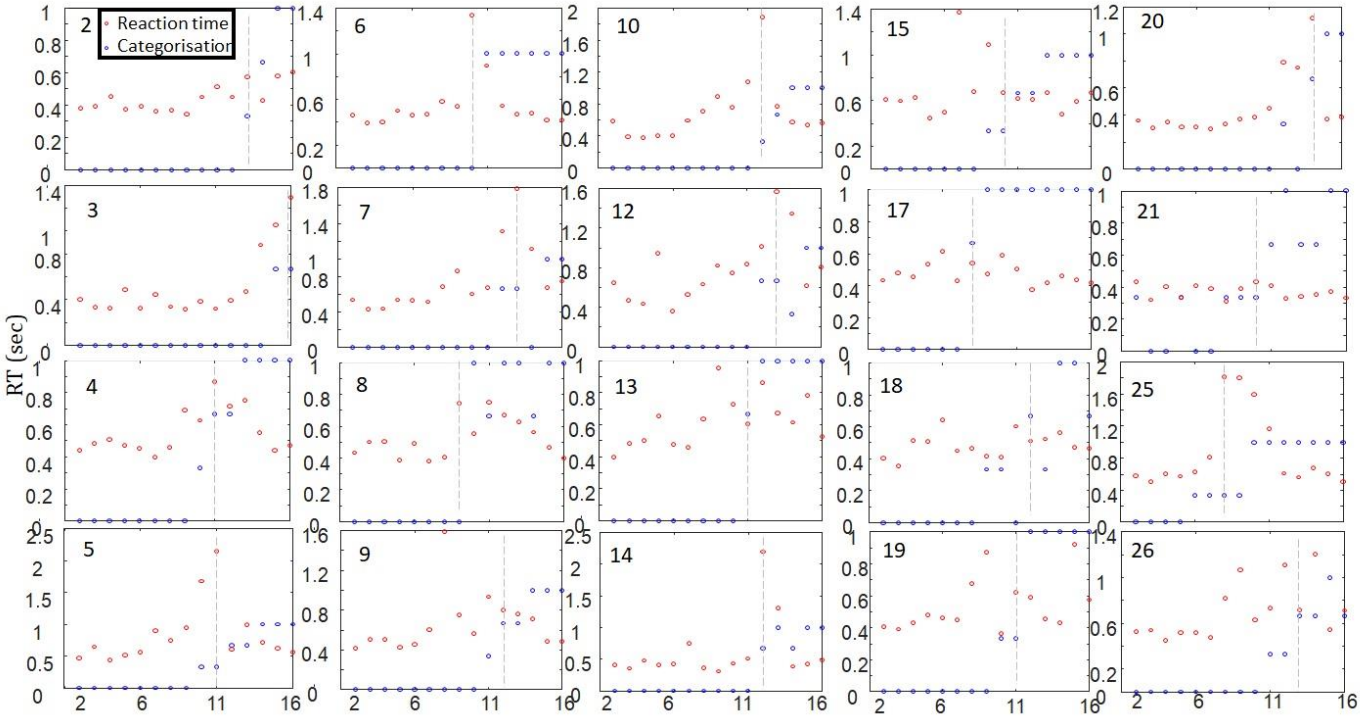
boundary, English – lighter pairs, English	-0.4924	0.261	-1.884	0.412
Cross-boundary, English – darker pairs, Russian	0.1941	0.424	0.458	0.997
Cross-boundary, English – cross-boundary, Russian	-0.3518	0.442	-0.795	0.968
Cross-boundary, English – lighter pairs, Russian	-0.3149	0.441	-0.715	0.980
lighter pairs, English – darker pairs, Russian	0.6865	0.442	1.554	0.629
lighter pairs, English – cross-boundary, Russian	0.1405	0.459	0.306	0.9996
lighter pairs, English – lighter pairs, Russian	0.1774	0.457	0.388	0.999
darker pairs, Russian – cross-boundary, Russian	-0.5460	0.250	-2.183	0.246
darker pairs, Russian – lighter pairs, Russian	-0.5091	0.247	-2.060	0.309
Cross-boundary, Russian – lighter pairs, Russian	0.0369	0.276	0.133	1.00

It can be seen that the interaction is driven by two types of differences: 1) performance is worse for darker, below boundary pairs than both cross-boundary pairs and lighter pairs in English participants only; 2) performance for these darker pairs in English participants is also worse than performance for cross-boundary and lighter pairs in Russian participants. This is also clearly visible from the box plot below.



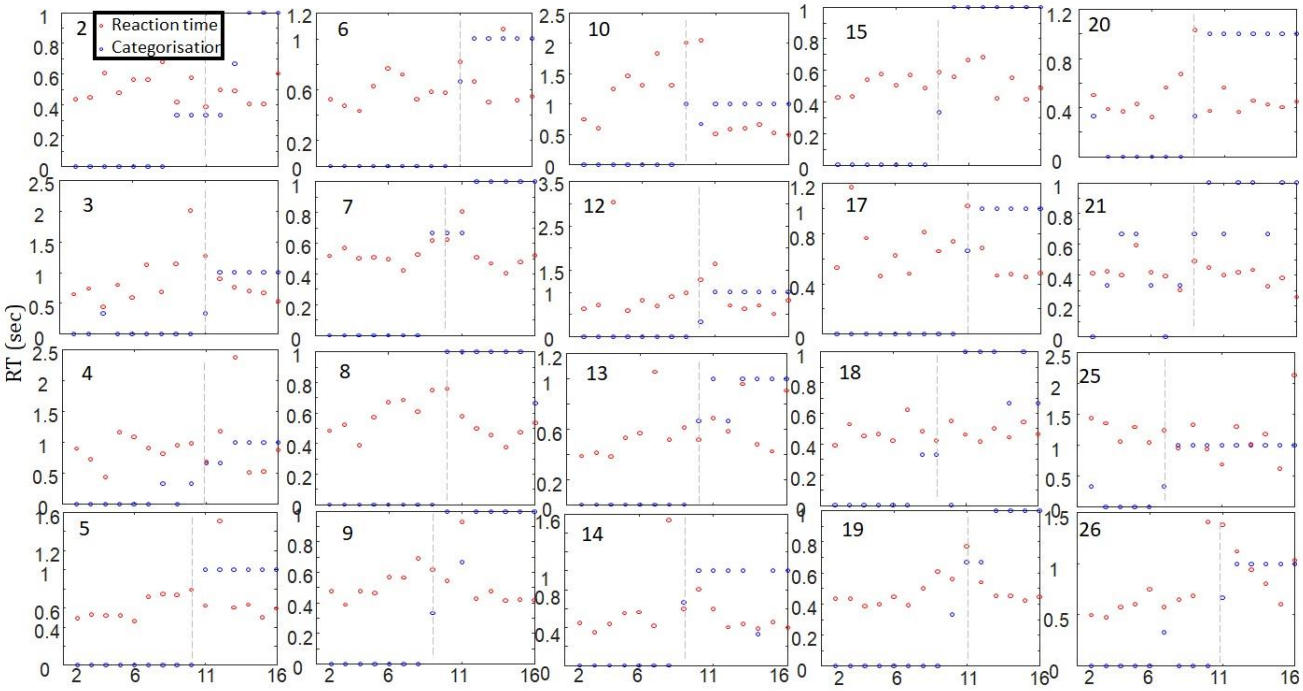
Box and whisker plot of accuracy from Experiment 1: Dot represents the mean, the top and bottom of the box represent 25th and 75th percentiles, whiskers extend to the most extreme data point which is no more than 1.5 times the length of the box away from the box. Any data points outside that range are marked as outliers.

Supplementary Figure 3a.



Categorisation of colours into sinij/goluboj for the Russian-speaking sample from Experiment 2 (white background). *Colours (2-16, see Fig.1) are depicted on the x axis. Blue dots depict the proportion of goluboj categorisations. Thus, sinij categorisations are plotted as blue dots that fall on the x axis itself while the proportion of goluboj categorisations will fall between 0 and 1. Note that the y axis, depicting reaction times (and, between 0 and 1, proportion of goluboj categorisation), is differently scaled between participants. The boundary is indicated by a dashed grey line.*

Supplementary Figure 3b.



Categorisation of colours into goluboj/green for the Russian-speaking sample from Experiment 2 (white background). *Colours (2-16, see Fig.1) are depicted on the x axis. Blue dots depict the proportion of green categorisations. Thus, goluboj categorisations are plotted as blue dots that fall on the x axis itself while the proportion of green categorisations will fall between 0 and 1. Note that the y axis, depicting reaction times (and, between 0 and 1, proportion of green categorisation), is differently scaled between participants. The boundary is indicated by a dashed grey line.*