Original Article

QJEP

Quarterly Journal of Experimental Psychology
2024, Vol. 77(4) 694–715

DOI: 10.1177/17470218231200489
qjep.sagepub.com
Sage

Do linguistic stimuli activate experiential colour traces related to the entities they refer to and, if so, under what circumstances?

Oksana Tsaregorodtseva[1], Lyn Frazier[2], Britta Stolterfoht[3] and Barbara Kaup[1]

Abstract

The simulation view of language comprehension holds that lexical-semantic access prompts the re-enactment of sensorimotor experiences that regularly accompany word use. For the colour domain, this suggests that reading about a stop sign reactivates experiences involving the perception of the stop sign and hence experiences involving the colour red. However, it is still not clear what circumstances would limit reactivation of colour experiences during comprehension, if the activation takes place. To address this question, we varied in our study the conditions in which the target colour stimuli appeared. The experimental stimuli were individual words (Experiment (Exp.) 1, Exp. 7, 8) or sentences (Exp. 2–6) referring to objects with a typical colour of either green or red (e.g., cucumber or raspberry). Across experiments, we manipulated the presence of fillers (present or not), and whether fillers referred to objects with other colour (e.g., honey) or objects without any particular colour (e.g., car). The stimuli were presented along with two clickable "yes" and "no" buttons, one of which was red and the other green. Location and button colour varied from trial to trial. The tasks were lexical decision (Exp. 1, Exp. 7–8) and sensibility judgement (Exp. 2–6). We observed faster response times in the match vs mismatch condition in all word-based experiments, but only in those sentence-based experiments that did not have fillers. This suggests that comprehenders indeed reactivate colour experiences when processing linguistic stimuli referring to objects with a typical colour, but this activation seems to occur only under certain circumstances.

Keywords

Grounded cognition; language comprehension; perceptual simulation

Received: 20 September 2022; revised: 2 March 2023; accepted: 9 May 2023

Unlike amodal theories, which assume that language processing is based on symbolic representations (Fodor, 1975), the embodied-cognition view suggests that understanding language is based on sensory and motor systems of the mind (Barsalou, 1999; Wilson, 2002). According to this point of view, understanding language involves the reactivation of multimodal sensorimotor processes that were involved during previous daily life experiences when the recipient interacted with the referents of the words. For example, when people interact with an object such as an apple, they experience its shape, colour, taste, smell, and the associated sound is also available. According to embodied views of cognition, when the word "apple" is later encountered, these experiences are reactivated, at least partially, in the form of mental simulations to gain access to the word's meaning.

In terms of visual simulations, several empirical studies have emerged demonstrating the role of object shape in modal meaning representations (Zwaan et al., 2002; see also Huettig & Altmann, 2004, 2007; Kaup et al., 2006) or object

[1]Department of Psychology, Language and Cognition Research Group, University of Tübingen, Tubingen, Germany
[2]Department of Linguistics, University of Massachusetts Amherst, Amherst, MA, USA
[3]German Department, University of Tübingen, Tubingen, Germany

Corresponding author:
Oksana Tsaregorodtseva, Department of Psychology, Language and Cognition Research Group, University of Tübingen, Schleichstr. 4, 72076 Tübingen, Germany.
Emails: oksana.tsaregorodtseva@uni-tuebingen.de
caregrad@gmail.com

orientation (Stanfield & Zwaan, 2001). Zwaan et al. (2002) presented sentences in which the shape of the object was implied. For example, the sentence "The ranger saw the eagle in the sky" implies that the wings of the eagle are spread. In contrast, the sentence "The ranger saw the eagle in the nest" implies folded wings of the eagle. After the sentence, an image was presented, and the participant decided whether the object depicted was mentioned in the previously presented sentence or not. The picture was such that it either matched or mismatched the implied shape of the object from the sentence. For example, the picture could depict the eagle with folded or outstretched wings for the example sentence provided above. The results showed a match advantage: Participants recognised the object faster and more accurately when the shape of the image matched the shape implied by the sentence than when it did not match. A similar paradigm was applied for orientation (Stanfield & Zwaan, 2001) and demonstrated a match advantage. The match advantage can be explained by assuming that participants reactivated their perceptual experience in the form of a mental simulation during reading, which then speeded up the processing of the matching image. In addition to studies that used the sentence–picture verification task, some studies have shown that overt attention can be directed to objects with similar shape as the objects mentioned in the sentence. For example, it has been shown that participants direct their looks to a picture of a cable during the acoustic presentation of the word "snake" (Huettig & Altmann, 2004, 2007) Shape plays an essential role in the theory of affordances (Gibson, 1979; Tucker & Ellis, 1998), being a quality or property of an object that defines how to interact with the object, and this property seems to be evoked both by pictured objects and words (Borghi & Riggio, 2009; Tucker & Ellis, 2004).

Although shape has been extensively researched in the embodiment literature, colour as a visual property has not received much attention. Indeed, shape processing is deeply linked to action because shape recognition determines how we interact with an object (Scorolli & Borghi, 2015). Also, since shape engages both visual and motor systems (Smith, 2005), it's clear why it was targeted by the embodied-cognition literature. After all, one of the central statements of the theory is that perception and action are tightly connected. In addition, shape is an important property in most of the models of object recognition (Tanaka et al., 2001), whereas colour is not. Connell (2007) noticed that colour represents a so-called secondary object property, which is perceived in only one modality, contrary to primary properties (e.g., shape, size, motion), which are perceived by multiple senses. Although the role of colour in object recognition and representation is less clear, we still consider it likely that colour is a property that plays a role in the simulations created during language comprehension. Before we turn to our study investigating this question, we will discuss studies investigating the role of colour in object recognition and representation.

The role of colour in object recognition and representation

From an evolutionary perspective, recognising colour gives primates, including humans, a behavioural advantage to segregate and organise visual input in a scene: colour plays a decisive role in low-level vision as it allows distinguishing edible fruits from inedible ones, or food from the natural background (Bramão et al., 2011; Tanaka et al., 2001). In contrast, the role of colour in later stages of object recognition is currently under debate in the literature, and different studies showed mixed results. For instance, Scorolli and Borghi (2015) showed that replacing a typical colour with an atypical colour in an object (e.g., showing a blue banana) did not affect recognition of the object in various tasks. At the same time, Naor-Raz et al. (2003) demonstrated that participants were faster naming the object's colour if the object was shown in a typical colour (yellow banana) than in an atypical colour (blue banana).

Essentially, the debate about the role of colour in later stages of object recognition is a question of whether our conceptual (stored) knowledge about the colour of a particular object affects recognition of the object. In other words, the point of interest is whether a colour is a substantial part of object representation and whether it is strong enough to affect object recognition. In this way, several variables influence the occurrence of a colour effect in empirical research, targeting object representation and recognition (Bramão et al., 2011). One of them is "colour diagnosticity," defined as the degree to which an object is associated with a specific colour (Tanaka & Presnell, 1999). For instance, a banana is a highly colour diagnostic object because it is strongly associated with yellow, whereas a hammer is not.

Tanaka and Presnell (1999) ran several experiments showing that colour has a more substantial influence on recognising high diagnostic objects compared with low diagnostic objects. However, Bramão et al. (2011) showed in a meta-analysis that colour can also affect the recognition of objects with low colour diagnosticity, albeit to a more moderate extent. And here, other variables such as "semantic category" of the stimuli and the "type of task" used in experiments also played a role (for more information, see a review of Bramão et al., 2011).

It has furthermore been shown that natural objects are more sensitive to colour manipulation than are artefacts (Laws & Hunter, 2006). The explanation lies in differences concerning the looks of objects: natural objects (e.g., apple, lemon) are more similar in shape to each other than artefacts are (e.g., traffic lights, fire engine), so colour becomes more helpful when it's necessary to differentiate one similar shape from another in natural objects. Even so, Bramão et al. (2011) showed that colour can also be important for artefacts, with the task used being an important

moderator variable at least under certain conditions. For instance, Scorolli and Borghi (2015) presented pictures of fruits and animals in a semantic categorisation task (animate/non-animate), a manipulation-evaluation task (graspable or not), and a motion-evaluation task (objects move by themselves or not). No difference in performance was found when participants made decisions about objects with typical or atypical colours (yellow vs blue banana, respectively). The lack of a colour effect in these tasks can be attributed to the fact that for the targeted properties (e.g., manipulation) colour information does not provide a recognition benefit. In contrast, other tasks showed a facilitation effect when objects were displayed in a typical compared with an atypical colour or when they were not coloured at all (Tanaka & Presnell, 1999; Wurm et al., 1993). The most substantial effect was observed for a (picture) naming task, whereas the object-name verification or semantic classification task showed more moderate effects (Bramão et al., 2011). Picture-naming is a demanding task compared to semantic classification or object-name verification tasks, which suggests that a shallower task leads to a less pronounced effect. In sum, the data suggest that colour indeed is a part of object representation and can affect object recognition. However, the effect of colour is context- and task-modulated.

The role of colour in language-based meaning representations

The question of whether perceptual information about colour becomes available on the processing of a word such as "banana" has attracted some attention in language comprehension research. However, the evidence has been mixed, with some studies showing facilitation and others interference.

Most studies used words referring to colour diagnostic natural objects or artefacts (e.g., banana, strawberry, traffic lights). The most widely used task in the literature for testing whether implicit perceptual information about object colour is activated during language comprehension is the sentence–picture verification task (e.g., Connell, 2007; Mannaert et al., 2017; Zwaan & Pecher, 2012). In this paradigm, participants read a sentence that implies a particular colour (e.g., John looked at the steak on the plate), and afterwards see an object that could either match (brown steak) or mismatch (red steak) the implied colour. The results suggest that the implied colour affects the decisions about objects depicted in a coloured image. However, the direction of this language-based colour-compatibility effect differs from study to study. The results of Connell (2007) showed a mismatch advantage (faster response time in the mismatch condition), which the authors explain by assuming that the same neuronal systems are involved in perception and language processing, with the latter happening via mental simulation. When perceptual input matches perceptual simulation, the matching colour information is more difficult to ignore because visual subsystems supporting simulation are already occupied. Zwaan and Pecher (2012) replicated the experiment of Connell (2007) and demonstrated the opposite pattern: responses were faster in the match than in the mismatch condition, and Mannaert et al. (2017), aiming at resolving the discrepancy in the results, tested the effect with a different set of stimuli and again showed a match advantage in a series of experiments, supporting the results of Zwaan and Pecher (2012). The authors of the latter two studies explain their results by assuming that language comprehension reactivates perceptual experiences, which then speed up the processing of the matching image.

Interestingly, the results obtained with a different paradigm also show opposite patterns, but this time the differences are easier to explain. Naor-Raz et al. (2003) presented words referring to colour diagnostic objects, which were printed in the typical colour (e.g., banana printed in yellow) or in an atypical colour (e.g., banana printed in purple), asking participants to name the ink colour. When stimuli with the typical and atypical ink were randomised, the results revealed a mismatch advantage, whereas when typical and atypical ink colours were presented in blocks, a match advantage was observed (replicating an earlier result by Klein, 1964). The authors attribute the discrepancy in the direction of the colour-compatibility effect to strategies that participants apply in a blocked design. Specifically, when participants know that the colour is typical throughout the block, then they can use this knowledge about colour typicality to speed up their response, which is not the case for the randomised design. Thus, these results suggest that the direction of the language-based colour-compatibility effect might depend on the peculiarities of the procedure, directing attention towards the object colour in some procedures but not in others.

In addition, Connell and Lynott (2009) showed that the colour-compatibility effect is sensitive to the linguistic context. In their study, participants read sentences that implied a colour that is typical for the described object (bear in the wood brown) or atypical but possible in reality (bear on the North Pole white). Participants performed a Stroop task in which they named the ink colour of the target word (e.g., bear). The ink either matched the typical, the atypical, or an unrelated colour. Results showed that colour naming was facilitated both when the colour was typical and when it was atypical compared with the unrelated condition. Connell and Lynott (2009) argued that short-term representations elicited by linguistic context coexist with the conceptual knowledge about the object.

Short-term, context-related factors appeared to be essential also in the study by Yee et al. (2012). In their experiments, participants performed a semantic-judgement task (animal/not animal) in a priming paradigm. Each critical target (e.g., cucumber) was combined with a

colour-related prime (e.g., emerald) or with a colour-unrelated prime (e.g., pendant). Half of the participants performed a Stroop task before the semantic-judgement task and the other half performed the tasks in the reversed order. The authors reported colour priming (e.g., emerald primes cucumber) only when participants had previously completed the Stroop task. When the order of tasks was reversed, priming was eliminated. Thus, the results highlight an important issue, namely, that the colour effect appears when attention is drawn to the colour domain, which seems to suggest that colour might be an attribute that is not automatically activated during language processing. This conclusion is in line with the findings by Huettig and Altmann (2011), who showed that eye movements are driven by actually perceived surface attributes of the visual object rather than stored conceptual knowledge of the typical colour of the object. Specifically, Huettig and Altmann (2011) monitored participants' eye movements in a visual-world paradigm. Participants were listening to sentences mentioning objects that were associated with a typical colour (e.g., spinach). Overt attention mediated by "language-derived" colour representations shifted to objects that were not usually associated with the diagnostic colour but were presented in the diagnostic colour of the target concept (e.g., a green blouse). However, there were no significant shifts in attention when participants saw objects associated with diagnostic colour but presented in black and white or an atypical colour (e.g., a frog in black and white or a frog in yellow). Huettig et al. (2020) argued that the visual environment is crucial for colour effects driven by linguistic stimuli.

Taken together, the findings in this research field are in line with the main conclusions drawn in studies concerned with the role of colour in object recognition: The effect of colour is context sensitive whereby context here is understood broadly, referring either to a linguistic context (Connell & Lynott, 2009) or a visual context (Huettig & Altmann, 2011; Yee et al., 2012). However, the circumstances that affect the reactivation of colour in language comprehension are far from being resolved. There are no studies in which different types of linguistic stimuli (words or sentences) were studied within the same paradigm. For other object properties, it was shown that some embodiment-related effects are indeed responsive to the types of linguistic stimuli used. For instance, Bub and Masson (2010) investigated action affordances during sentence processing and reported clear word-based compatibility effects but no sentence-based effects during regular reading. Moreover, as mentioned above, previous research has revealed that the nature of the task may determine the extent to which colour effects are observed. Thus, we consider it helpful to have a paradigm that allows using different tasks without changing the paradigm, to enable an investigation of the level of processing that is needed for colour effects to emerge.

This study

The main goal of this study was to expand existing knowledge about the circumstances when implicit perceptual information about the colour of an object gets activated during language understanding. To investigate this issue, we used linguistic stimuli (words and sentences), which refer to objects that are typically green or red. In other words, in this study, words referring to colour diagnostic objects were employed for the primary goal's sake. Specifically, there were several aims of this study: First, providing an additional test of whether comprehenders reactivate colour experiences during comprehension. Second, finding out whether a match between implied and perceived colour leads to facilitation or interference. Third, making a step towards understanding what circumstances affect the reactivation of colour experiences during comprehension.

We developed a new paradigm taking into account results from research described above. One prerequisite for observing a language-driven colour effect lies in the corresponding visual context, namely, whether colours are actually presented which correspond to the colours implied by linguistic stimuli. The lack of such context resulted in a lack of a colour effect when participants were exposed to sentences with implied colours and saw black and white drawings of objects with a typical colour (Huettig & Altmann, 2011; Mannaert et al., 2017; see above). In our paradigm, the linguistic stimuli referring to either green or red objects (e.g., cucumber vs strawberry) were presented along with two clickable buttons, labelled "yes" and "no." The buttons were coloured in red and green. Location and button colour varied from trial to trial, so that "yes" and "no" labels appeared on the green and red buttons as well as on the left and right side. The advantage of the used paradigm is as follows: We can use different tasks in conjunction with different types of linguistic stimuli in the presence of visual context (coloured buttons) within the same paradigm, which expands the possibility of an in-depth study of whether the colour implied by the linguistic stimulus affects the responses to the coloured buttons.

In this study, we used a lexical-decision task in experiments employing words as stimuli (Experiments 1, 7–8) and a sensibility-judgement task in experiments with sentences as stimuli (Experiments 2–6).

All experiments were implemented using JsPsych (version 6.1.0.). Each experiment was available from a link, and participants could complete it at home using a regular web browser (de Leeuw & Motz, 2016). We collected data online via university emails or Prolific. At the beginning of the experiment, participants gave informed consent and were further inquired about their mother tongue. This study was approved by the Ethics Committee for Psychological Research of University of Tübingen and were carried out in accordance with the Declaration of Helsinki. After completing the experimental session,

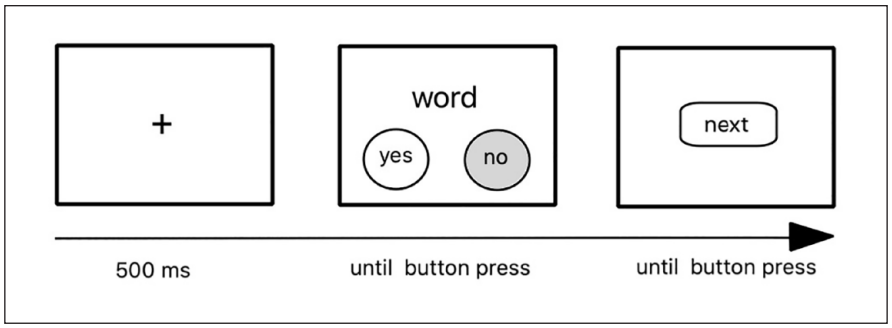

Figure 1. Schematic representation of trial procedure (not to scale).

participants were asked to take part in a brief survey about the procedure. The first question concerned people's awareness that they participated in a colour-related experiment. The answer options were "Yes," "No," and "Only later, as the experiment progressed." The second question was about the device participants used to complete the experiment. The options were: "Computer and Computer mouse," "Laptop and Touchpad," "Tablet," "Smartphone."

Experiment 1

In Experiment 1, we aimed at testing the paradigm with simple colour-compatibility effects triggered by words referring to objects with a typical colour of green or red.

Methods

Participants. Before starting the experiment, we calculated the number of participants needed to observe an interaction with a small effect size using the *simr* package (Green & MacLeod, 2016). The analysis showed that 120 participants and 60 relevant items was enough to provide power of 95%. Thus, we aimed at recruiting at least 120 participants who would meet our criteria for being included in our analyses in this experiment.

In Experiment 1, we recruited 137 volunteers from the University of Tübingen (41 males, $M_{age}=28.63$, $SD_{age}=10.32$) via e-mail who participated for course credit or for the possibility to win one of eight 20€ vouchers for an online shop in a lottery. Since one of the prerequisites for participation was German as a native language, we saved data of only 132 participants who explicitly stated that German was their mother tongue (40 males, $M_{age}=28.36$, $SD_{age}=10.23$) for analysis.

Materials. The stimuli were 60 words referring to objects that are typically green or red (e.g., cucumber vs tomato). To select those words, we conducted a pretest in which 16 participants (2 males, $M_{age}=26.38$, $SD_{age}=16.01$) saw 147 words referring to objects with different colours or with no particular colour (e.g., cucumber, honey, or car) along with coloured patches (red, yellow, green, brown, blue, as well as a white patch representing no particular colour). The patches had captions indicating the colours with a no-colour option. The participants had to mark the colour that they associated with the object. They also rated how strong their association was using a 7-point Likert-type scale placed under the previous request. Based on this rating, we selected 30 "green" and 30 "red" words referring to green or red objects having the highest strength score associated with red and green ($M_{red}=6.02$ vs $M_{green}=6.21$). The two groups of words did not differ in strength, $t(58)=1.28$, $p=.204$.

When selecting the words, we did not control for frequency or length. There was a significant difference in length between the two groups of words, $t(58)=–3.89$, $p<.001$, with longer "red" words in comparison with the group of "green" words. We will take this into account in our analyses. For testing frequency, we took frequency classes from the German corpus "Wortschatz Universität Leipzig" (http://wortschatz.uni-leipzig.de). Results showed that the frequency between "red" and "green" words did not differ, $t(58)=–1.05$, $p=.297$.

We also created 60 pseudowords using a pseudoword generator Wuggy (http://crr.ugent.be/programs-data/wuggy).

Procedure and design. The experiment consisted of eight practice trials to familiarise participants with the task, followed by 120 experimental trials. Each trial started with a fixation cross, which appeared briefly for 500 ms. After the fixation cross, the stimulus written in a black font in 16 pt size was presented along with two clickable "yes" and "no" buttons, of which one was coloured red and the other green. A schematic representation of the trial procedure is given in Figure 1. Participants were allowed to use the following devices: computer with computer mouse, computer with touchpad, tablet, or phone. Responses were given by clicking on one of the buttons by means of a mouse click, a touch on the touchpad, tablet, or phone.

Location and button colour varied from trial to trial, so that "yes" and "no" responses appeared both with a green and a red button, and both on the right and the left side of the screen in the course of the experiment. The task was to evaluate whether the presented string was a word in the German language (lexical decision). We implemented four

different lists, with each word being shown with each button configuration (colour and side) in one of the lists. After participants had responded by clicking on one of the two buttons, a clickable button "Next" was presented in the centre of the screen. This ensured that participants returned their response effector to the central position of the screen before turning to the next trial. The presentation of words/pseudowords was randomised, and the whole experiment took approximately 10–12 min to complete. The design was a 2 (referent colour: "green" vs "red") × 2 (yes-button colour: "green" vs "red") within-participants design.

Results and discussion

Prior to analysing the data, we set an accuracy threshold. When doing so, we first analysed the accuracy data visually using the *plot*, *hist*, *boxplot* functions, and realised that there were three extreme outliers lying below 80% accuracy. For comparison, the accuracy of the original set ranged from 50% to 99% ($M = 94.74$, $SD = 5.67$). The accuracy of participants after excluding the outliers was quite high: it ranged from 88% to 99% ($M = 95.52$, $SD = 2.22$).

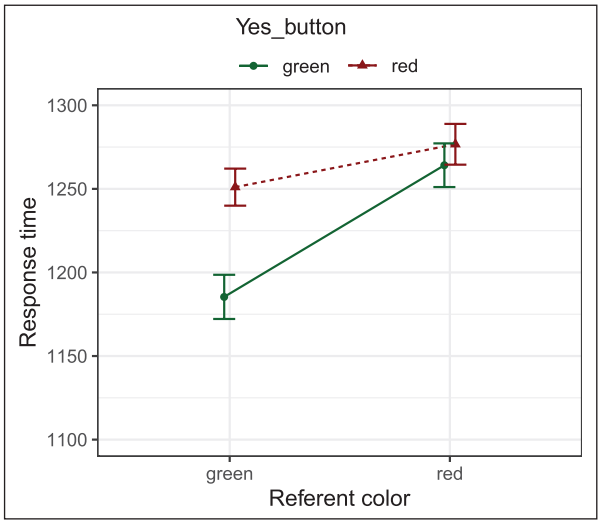

Figure 2. Response times in the lexical decision task of Experiment 1.
Error bars denote 95% within-subjects confidence intervals calculated as recommended by Morey (2008).

After that, we excluded the data of participants who did not reach an accuracy threshold of 80% of correct answers, which resulted in a selection of 129 participants (38 males, $M_{age} = 28.51$, $SD_{age} = 10.29$). Then this group was reduced to a final set consisting of 120 participants (35 males, $M_{age} = 28.66$, $SD_{age} = 10.43$) by equalising the number of participants per list. When we equalised the number of participants, we took participants randomly from each list.

The time required to read the stimulus and make a lexical decision about it served as the dependent variable. Our analyses are based on the response times in word-trials. Response times in non-word trials were excluded from the analyses, as were response times in error trials. Error trials represented 3.42% of the relevant data. Based on visual inspection of the data plot, answers faster than 100 ms and slower than 2,000 ms were considered absolute outliers and were omitted (9.51% of the data). For relative outlier elimination, we took differences between the participants and between items into account (Kaup et al., 2006, 2010; Kelter et al., 2004). More specifically, we first converted the RTs to *z*-scores by participant and then converted these *z*-scores to *z*-scores per item and condition. We then eliminated all trials for which the absolute value of this *z*-score was above 2 (3.69% of the relevant data). We use this outlier elimination procedure in all of our experiments. The means of the final set of response times are depicted in Figure 2 (see also Table 1).

We analysed the results employing a linear mixed-effects model (LMEM), using the free statistic software *R* (Version 4.0.3) and the R-package lme4 (Bates et al., 2015). We implemented contrast coding for categorical variables. Regarding the random-effects structure, we selected the most complex structure with which our models would converge after adding fixed effects of word length, referent colour, yes-button colour, and the interaction of referent colour and yes-button colour. This was the case for a model with random intercepts for participants and items.[1] The same procedure was applied for all of the subsequent experiments reported in this article.

With this random-effects structure, we then created a base model with fixed effects of word length, referent colour, and yes-button colour. We compared this base model to several reduced models, dropping the fixed effects one by one employing a likelihood-ratio test.

Table 1. Means and standard deviations per condition in Experiments 1–4.

Referent colour	Experiment 1		Experiment 2		Experiment 3		Experiment 4	
	Yes-button		Yes-button		Yes-button		Yes-button	
	Green	Red	Green	Red	Green	Red	Green	Red
Green	1,185 (175)	1,251 (177)	2,300 (486)	2,437 (547)	3,235 (911)	3,395 (1042)	3,212 (880)	3,284 (914)
Red	1,264 (179)	1,277 (181)	2,440 (547)	2,466 (557)	3,355 (966)	3,418 (1018)	3,273 (860)	3,363 (929)

Table 2. The percentage of correct responses in Experiments 1–4.

Referent colour	Experiment 1		Experiment 2		Experiment 3		Experiment 4	
	Yes-button		Yes-button		Yes-button		Yes-button	
	Green	Red	Green	Red	Green	Red	Green	Red
Green	98%	96%	95%	93%	93%	93%	93%	92%
Red	96%	96%	93%	93%	91%	92%	93%	91%

The base model was better than the reduced model without the fixed effect of word length ($\chi2(1)=34.29$, $p<.001$, $\beta=24.97$, $t=6.81$) indicating that participants were faster to respond to shorter words than to longer words. The base model also outperformed the model without the fixed effect of yes-button colour ($\chi2(1)=48.33$, $p<.001$, $\beta=–19.26$, $t=–6.83$) suggesting that participants responded faster to a green than to a red button. However, the base model did not show better results than the reduced model without the fixed effect of referent colour ($\chi2(1)=0.78$, $p=.376$, $\beta=–12.41$, $t=0.91$).

We then compared the base model with the model including the interaction between referent colour and yes-button colour. The model with the interaction revealed a better fit than the base model ($\chi2(1)=26.98$, $p<.001$, $\beta=–14.68$, $t=–5.20$). This model was taken as the best model for further comparison. We also tested possible models that included interactions between word length and other factors. However, they did not improve the fit (all values of $p>.6$).

We looked at simple effects to understand the interaction between the referent colour and the yes-button colour. Analysis of simple effects revealed that participants were 66 ms faster when pressing green yes-buttons than red yes-buttons in making decision about "green words" ($\chi2(1)=75.74$, $p<.001$, $\beta=–34.05$, $t=–8.76$). However, when they made a decision about "red" words, no effect of yes-button was obtained ($p=.293$). In addition, on "green" words, the green yes-button was pressed 79 ms faster than on "red" words ($\chi2(1)=7.33$, $p=.007$, $\beta=–44.52$, $t=–2.79$), but no difference was found for the "red" words ($p=.394$). Thus, the interaction we observed was driven mostly by "green" words, but not "red" words.

An analysis of accuracy rates showed a marginal interaction between referent colour and yes-button colour ($\chi2(1)=3.31$, $p=.069$, $\beta=0.12$, $z=1.81$), demonstrating a match advantage when participants clicked green buttons and saw "green" words (see Table 2). Also, a main effect of yes-button colour ($\chi2(1)=5.93$, $p=.015$, $\beta=0.18$, $z=2.54$), was observed, indicating that green buttons were pressed more accurately than red buttons (97% vs 96%). In addition, we obtained a main effect of length ($\chi^2(1)=4.33$, $p=.038$, $\beta=0.09$, $z=2.07$), showing that participants were more accurate with shorter words. The percentage of correct responses per condition is presented in Table 2.

Taken together, the results of Experiment 1 supported the hypothesis that the simulated colour influences responses to coloured buttons in the given paradigm, demonstrating a match advantage: responses were faster when there was a match between the implied colour and the colour of the yes-button (pressing a green button for a green word) compared with when there was a mismatch (pressing a red button for a green word). However, the analysis of simple effects showed that the match advantage was mainly driven by the "green" words. We assume that the lack of the expected matching advantage for red words can be explained by several factors, one of which is the cultural association between red and prohibition. We will discuss this in detail in the post hoc analysis section below where we argue that the our match advantage effect for the red words (pressing a red button is faster than pressing a green button after reading a word referring to a red referent) is counteracted by a response prolonging effect associated with pressing red buttons.

Experiment 2

The goal of Experiment 2 was to test whether a language-driven colour-compatibility effect would also appear when linguistic stimuli were sentences instead of words. Instead of a lexical decision task, we employed a sentence–sensibility–judgement task in this experiment. We structured the sentences in such a way that the critical word was always the last word in the sentence. Being in the last position, the critical nouns appeared very close to the point in time when participants made their response choice. The aim of this decision was to provide the strongest conditions for potentially observing compatibility effects with sentences as materials as a starting point.

Methods

Participants. We calculated the number of participants needed to obtain the interaction effect on the basis of the previous experiment using the *simr* package (Green & MacLeod, 2016). The analysis showed that 120 participants and 60 relevant items (i.e., the conditions realised in Experiment 1) was enough to provide power of $>95\%$. Thus, we aimed at recruiting at least 120 participants who would meet our criteria for being included in our analyses

in this and later parallel experiments. Due to the particular participant recruitment procedures employed, for some experiments, recruitment procedures resulted in a final number of participants that was slightly larger. Importantly, however, the minimal amount of 120 participants was obtained in all our experiments.

Participants were recruited via an e-mail to students of the University of Tübingen who did not receive the invitation mail for Experiment 1. A total of 137 volunteers completed the experiment (27 males, $M_{age}=23.77$, $SD_{age}=5.68$). They participated for course credit or for the possibility to win one of eight 20€ vouchers for an online shop in a lottery. Of these 137 volunteers, 128 participants who were native German speakers entered our analyses (25 males, $M_{age}=23.70$, $SD_{age}=5.77$).

Materials. From the words used in Experiment 1, we created 60 simple sentences with the structure "Locative" + "Predicate" + "Object" (e.g., *Auf dem Schneidebrett liegt ein Granatapfel.* / "On the cutting board, there is a pomegranate.") with the object word always being the last word in the sentence and representing a referent with a typical red or green colour. The set of object words was identical to the words used in Experiment 1. Thus, 30 sensical sentences referred to an object that is typically green and the remaining 30 to an object that is typically red. We also constructed 60 nonsensical sentences with the same structure (e.g., *Auf dem Datum ist ein Fußboden.* / "On the date, there is a floor.") for the purpose of the task ("no"-responses). The object in the nonsensical sentences referred to an object without any particular colour (e.g., *Fußboden*/ "floor").

Procedure and design. The procedure for Experiment 2 was the same as for the previous experiment, with one exception. This time, the task was a sentence–sensibility–judgement task, where participants had to read the sentence and evaluate whether the sentence made sense or not. The response time was measured from the onset of the sentence. The experiment took approximately 15 min. The design again was a 2 (referent colour: "green" vs "red") × 2 (yes-button colour: "green" vs "red") within-participants design.

Results and discussion

As in Experiment 1, we first set the accuracy threshold for the experiment with a different task and with different types of stimuli. We defined the outliers using the same procedure as in Experiment 1. The accuracy of the original set varied from 72% to 99% ($M=94.06$, $SD=4.02$). The accuracy of participants after excluding the outliers was again quite high and it ranged from 82% to 99% ($M=94.24$, $SD=3.5$). We then set the threshold for the rest of the experiments to 80%, accordingly.

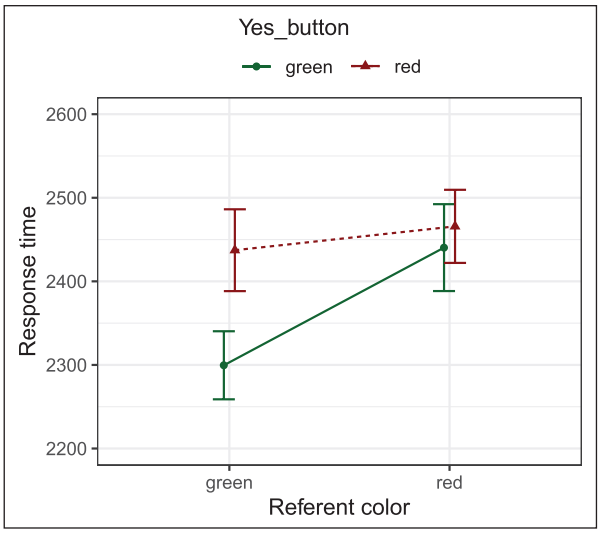

Figure 3. Response times in the sentence–sensibility–judgement task of Experiment 2.
Error bars denote 95% within-subjects confidence intervals calculated as recommended by Morey (2008).

We eliminated participants' data with an accuracy rate of less than 80%, which decreased the number of participants to 127 (24 males, $M_{age}=23.71$, $SD_{age}=5.79$). After that, we equalised the number of participants per list so that the final set consisted of 120 participants (35 males, $M_{age}=23.77$, $SD_{age}=5.89$).

We only analysed responses to sensical items and eliminated incorrect answers for our response time analysis. Errors were made in 6.56% of the trials. Visual inspection of the data led us to eliminate responses that were longer than 20,000 ms or shorter than 500 ms as absolute outliers. This resulted in the omission of 0.05% of the relevant data. It should be noted that the different time frames of absolute outliers definition occurred because longer processing times are expected when sentences are involved. For relative outlier elimination, we applied the same procedure as in Experiment 1. Based on this procedure, 4.67% of relevant trials were discarded.

The means of the final set of response times in Experiment 2 are depicted in Figure 3 (see also Table 1).

We used a base model consisting of fixed main effects for sentence length, referent colour and yes-button colour, and random intercepts for participants and items. The models with higher complexities did not converge. Sequential dropping of main effects from the base model revealed a main effect of sentence length ($\chi^2(1)=24.67$, $p<.001$, $\beta=34.24$, $t=5.53$), and a main effect of yes-button colour ($\chi^2(1)=13.98$, $p<.001$, $\beta=-39.78$, $t=-3.72$), but no effect of referent colour ($\chi^2(1)=0.01$, $p=.996$, $\beta=-0.04$, $t=-0.001$). Adding the interaction between referent colour and yes-button colour to the base model significantly improved the fit ($\chi^2(1)=5.82$, $p=.016$, $\beta=-25.79$, $t=-2.41$). The pattern of the results thus

replicated the one in Experiment 1, namely, the response in which the yes-button colour matched the colour of the referent were faster than responses in which the yes-button colour mismatched the colour of the referent. The model with interaction was considered the best model for the next analysis steps. The added interaction between sentence length and yes-button colour improved the best model ($\chi^2(1)=10.38$, $p=.001$, $\beta=-19.11$, $t=-3.37$), suggesting that length had a stronger impact for red than for green words.

Analysis of simple effects showed the same pattern as in Experiment 1 and demonstrated the same asymmetry between "green" and "red" items. We observed a button effect for "green" words where green yes-buttons were pressed 137 ms faster than red yes-buttons ($\chi^2(1)=21.46$, $p<.001$, $\beta=-65.58$, $t=-4.64$). No effect was observed for "red" words ($p=.427$). Also, green yes-buttons were pressed marginally faster after "green" words ($\chi^2(1)=2.95$, $p=.086$, $\beta=-79.02$, $t=-1.74$), but no difference was obtained for red yes-buttons ($p=.554$).

An analysis of accuracy rates showed a marginal interaction between referent colour and yes-button colour ($\chi^2(1)=3.68$, $p=.055$, $\beta=0.10$, $z=1.99$), demonstrating a match advantage for "green" words and green yes-buttons. In addition, there was a tendency for shorter sentences to be processed more accurately than longer sentences, resulting in a marginal main effect of length ($\chi^2(1)=3.58$. $p=.058$, $\beta=-0.04$, $z=-1.92$). The percentage of correct responses per condition is presented in Table 2.

Overall, Experiment 2 revealed that a language-driven colour-compatibility effect can be observed in simple sentences in which the critical noun appears at the end of the sentence.

It is worth noting that in this experiment conditions were such that it would be particularly likely that a colour-compatibility effect is being observed. The critical nouns appeared at the very end of sentences, which was close to the moment when the participants made a choice. In the next experiment, we tested whether a colour compatibility effect will also be observed if the critical noun does not appear at this privileged position in the sentences.

Experiment 3

In the previous experiment, the critical noun (e.g., tomato or cucumber) appeared at the very end of the sentence. Due to this sentence structure, the critical noun appeared close to the point in time when participants made their response decision pressing one of the coloured buttons. Also, the sentence-final position has many special properties: it is highly predictable, given the presence of a period, it receives especially long processing times in reading studies (so-called "wrap-up effect"), and its occurrence is not masked by the presentation of a following word (Hirotani et al., 2006; Kuperman et al., 2010; Rayner et al., 1989). The critical nouns in Experiment 2, in addition to being close to the time of response, had these special conditions.

In Experiment 3, we aimed at investigating whether the language-driven colour-compatibility effect would also be observed for sentences in which the critical word appeared at a less privileged position. In this experiment, the critical words appeared before the end of the sentences. Furthermore, the sentences were intended to describe a vivid event, which supposedly distributed attention towards several aspects of the event without drawing attention directly to the target object, unlike in Experiment 2. Thus, in Experiment 3, we tested whether a colour compatibility effect would also be observed, when critical nouns were in less special conditions.

Methods

Participants. A total of 242 (43 males, $M_{age}=23.06$, $SD_{age}=6.32$) volunteers were recruited via the e-mail list of the University of Tübingen. A prerequisite for participation was the absence of participation in previous experiments. Our final data set included data of those 233 participants (42 males, $M_{age}=22.61$, $SD_{age}=3.81$), who were German native speakers. All volunteers could be compensated with course credit or participate in the lottery and win one of nine 20€ vouchers for an online-shop.

Materials. We created 60 sensical and 60 nonsensical sentences. All sentences, including nonsensical ones, had the following structure: "Determiner" + "Subject" + "Part of Predicate (auxiliary verb)" + "Adverbial modifier of time" + "Object" + "Predicate." The adverbial modifier of time was always *jetzt gleich*, the verb *wird/werden* was always used as an auxiliary verb (e.g., *Der freche Junge wird jetzt gleich nach dem Frosch greifen. /* "The cheeky boy will grab the frog now."). The "Object" referred to an object typically having either a green or red colour for sensical items, and to an object without any particular colour for nonsensical items (e.g., *Der traurige Großvater wird jetzt gleich die Schublade einweichen./* "The sad grandfather will soak the drawer now"). The sentences described an action about to happen, and the critical noun always appeared before the end of the sentence. The object words in the sensical sentences were the same words as used in Experiments 1 and 2. Thus, 30 of the sensical sentences referred to an object that is typically green and the remaining 30 sensical sentences referred to an object that is typically red.

Procedure and design. The procedure and the design of Experiment 3 were the same as in Experiment 2. The experiment lasted 15–20 min.

Results and discussion

Before analysis, we deleted data of 16 participants who did not reach the threshold of 80% accuracy, and we then

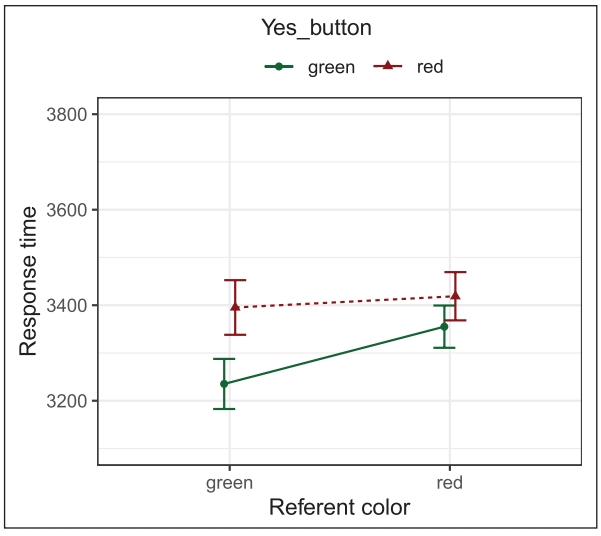

Figure 4. Response times in the sentence–sensibility–judgement task of Experiment 3.
Error bars denote 95% within-subjects confidence intervals calculated as recommended by Morey (2008).

equalised the number of data files per list, resulting in a final set of 184 participants (29 males, M_{age}=22.73, SD_{age}=3.94). As before, we only analysed responses to sensical items and discarded non-accurate responses, leading to a data loss of 7.91%. We then applied the same procedure of outlier elimination as in Experiment 2, reducing the data set by 0.08% for absolute outliers and another 4.47% for relative outliers. The means of the final set of response times is depicted in Figure 4 (see also Table 1).

The base model comprised as fixed effects sentence length, referent colour, and yes-button colour as well as by-participants and by-item random intercepts. Similar to the previous experiments, we consequently dropped each fixed main effect and compared the reduced model with the base model. The likelihood-ratio test revealed that there was main effect of sentence length ($\chi^2(1)=20.33$, $p<.001$, $\beta=29.87$, $t=4.94$), and a main effect of button colour ($\chi^2(1)=23.56$, $p<.001$, $\beta=–54.6$, $t=–4.86$), but no effect of referent colour ($\chi^2(1)=0.23$, $p=.63$, $\beta=18.73$, $t=0.48$). The model with an interaction between referent colour and yes-button colour outperformed the base model ($\chi^2(1)=4.32$, $p=.038$, $\beta=–23.36$, $t=–2.08$). The models with the interaction between sentence length and other factors did not improve the last best model (all values of $p \geqslant .5$).

Again, the analysis of simple effects revealed differences in match effects for the two types of words. A match advantage, was observed for the "green" sentences where participants pressed the green yes-buttons 160 ms faster than the red yes-buttons ($\chi^2(1)=24.01$, $p<.001$, $\beta=–78.16$, $t=–4.91$). Interestingly, in this experiment, we also observed an effect of button colour for the "red" sentences, where participants pressed the green yes-buttons 63 ms faster than the red yes-buttons ($\chi^2(1)=3.96$, $p=.046$, $\beta=–31.54$, $t=–1.99$), indicating a mismatch advantage for red sentences. No effects of word type were observed, neither for green yes-buttons ($p=.202$) nor for red yes-buttons ($p=.821$). Thus, in Experiment 3, we observed a match advantage for green sentences but a mismatch advantage for "red" sentences. Still it should be noted, that the significant interaction indicated a match advantage overall with faster response times in the match than in the mismatch conditions. This interaction mainly reflects that differences between the response times for the green and red buttons was larger for the "green" (match advantage) than for the "red" items (mismatch advantage). We will discuss the issue in more detail in the post hoc analysis section below. For now, we conclude that a match advantage was observed in this experiment as in the previous two experiments.

An analysis of accuracy rates showed a main effect of yes-button colour ($\chi^2(1)=9.75$, $p=.002$, $\beta=0.12$, $z=3.24$), with greater accuracy for green than for red buttons (93% vs 91%). Other effects and interactions were not observed for accuracy (all values of $p \geqslant .1$). The percentage of correct responses per condition is presented in Table 2.

Taken together, the three experiments we have conducted so far support the hypothesis that the colour implied by a linguistic stimulus influences responses to coloured buttons after reading this stimulus. We thus observed language-driven colour compatibility effects with words, with simple sentences in which the critical noun appeared at the end, and with sentences in which the critical noun appeared within the sentence describing a vivid event. All experiments showed a match advantage overall. Responses were faster in conditions in which the button to be pressed matched the colour implied by the linguistic stimulus compared with conditions in which there was mismatch between implied colour and button colour. Simple effect analyses revealed that the match advantage was mostly driven by the green items in all experiments. These findings are thus in line with those experiments in previous research, demonstrating facilitation of responses in matching conditions (Mannaert et al., 2017; Zwaan & Pecher, 2012).

As a next step, we aimed at exploring the stability of the language-driven colour compatibility effect. In particular, on the basis of the conclusions drawn from the previous literature that colour effects are context-sensitive (see above), we were interested in whether these effects would still be observed when the materials include a relevant amount of fillers that either refer to objects without a typical colour or to objects with typical colours that are different from the button colours. To the best of our knowledge, no previous study has directly tested the role of fillers on the presence of the colour effect in a paradigm allowing words and sentences as materials. Moreover, previous studies have mostly not included additional items in the materials if not required by the task. Thus, as of yet, the role of fillers in finding a

language-based colour compatibility effect is unclear. Considering that this issue is highly relevant for the question of whether the effects are automatic or not, we deemed it worth investigating. Investigating the role of fillers in a paradigm allowing both words and sentences as materials will further allow us to determine whether there is a difference between these two stimulus types with regard to the automaticity of effects. In principle, it seems conceivable that effects with words as materials are automatic whereas effects with sentences as materials are more context-dependent (Kaup et al., 2016). We begin by investigating the role of fillers with sentence materials.

Experiment 4

In the previous experiment, we tried to draw attention away from the target object by having the sentences describe a vivid event. We also placed the critical nouns before the sentence end to separate the points in time when participants process the target noun from those of response-selection. Nevertheless, we observed a language-driven colour compatibility effect. In Experiment 4, we added fillers to the stimulus set. These fillers referred to objects that had a colour other than green or red. Thus, relative to the previous experiment, attention was drawn away from the specific button colours red and green. If the language-based colour-compatibility effect hinges on the salience of the colour domain in general, then we should again observe such an effect in this experiment. In contrast, if it hinges on the salience of the specific colours present in the visual context, then no colour-compatibility effect should be observed in the present experiment.

Methods

Participants. We recruited 237 participants (45 males, M_{age} = 23.77, SD_{age} = 6.89) via the mailing list of the University of Tübingen. We selected data of 197 participants (38 males, M_{age} = 23.21, SD_{age} = 6.26) who indicated German as their native tongue and who had not participated in Experiments 1–3. Participation was compensated with course credit. Also, participants could take part in a lottery and win one of nine 20€ vouchers for the online shop.

Materials. The set of stimuli consisted of 120 sentences (60 sensical, 60 nonsensical items), which we used in Experiment 3. To the existing set, we added 60 fillers, which were sensical items where the "Object" position was taken by nouns referring to objects with a colour other than green or red. (e.g., *Der neugierige Wissenschaftler wird jetzt die Banane schälen./* "The curious scientist will now peel the banana."). When we selected the words for the set of fillers, we took the words from the pretest for this study. These words were rated by the pretest participants as having a particular colour (yellow or blue, or brown etc.) We took the words having the highest association with colour (M = 5.45). It is worth noting that the strength of association for filler words was slightly lower than for "green-red" words, mainly because it was difficult to find the other 60 words for which colour was a diagnostic feature.

We also created 60 additional nonsensical items with the same structure to equalise the number of sensical and nonsensical sentences. Overall, there were 120 sentences that made sense (experimental items + fillers) and 120 sentences that did not make sense.

Procedure and design. The procedure and design were the same as in Experiment 3. The experiment lasted 25–30 min.

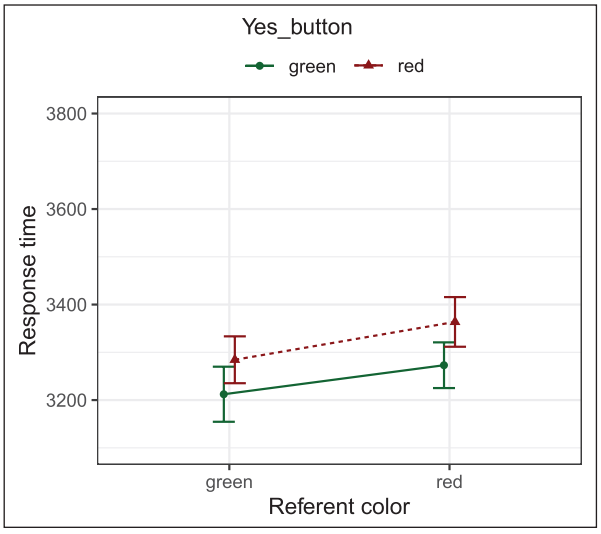

Figure 5. Response times in the sentence–sensibility–judgement task of Experiment 4.
Error bars denote 95% within-subjects confidence intervals calculated as recommended by Morey (2008).

Results and discussion

The data of 22 participants who did not pass a threshold in 80% of correct answers were excluded from further analysis. After that, we equalised the number of participants per list and arrived at a final set of 160 participants (28 males, M_{age} = 23.42, SD_{age} = 6.67).

Only correct responses to experimental items were taken into account for the response time analyses. Non-accurate responses were deleted (8.14% of relevant trials). We then followed the same procedure as in the previous experiments, namely, we discarded absolute outliers by visual inspection of the data, which led to eliminating responses below 500 ms and above 20,000 ms, reducing the data set by 0.14%. We additionally eliminated 4.36% of the data as relative outliers. The means of the final response times are depicted in Figure 5 (see also Table 1).

The base model included fixed effects of sentence length, referent colour, and yes-button colour, as well as random intercepts for participants and items. According to the likelihood-ratio test, the base model showed a better fit in comparison with the reduced model without length ($\chi^2(1)=20.63, p<.001, \beta=29.61, t=4.98$) and the reduced model without button colour ($\chi^2(1)=9.06$, $p=.002$, $\beta=-40.17$, $t=-3.01$). The results thus revealed both a main effect of word length and a main effect of yes-button colour. As previously, we did not observe a main effect of referent colour ($\chi^2(1)=0.24$, $p=.622$, $\beta=18.55$, $t=0.49$). The model with an interaction between referent colour and yes-button colour did not outperform the base model ($\chi^2(1)=0.07$, $p=.798$, $\beta=3.42$, $t=0.26$). Other models with interactions of word length with the other factors did not improve the fit (all values of $p \geqslant .4$). An analysis of accuracy rates showed a main effect of yes-button colour ($\chi^2(1)=7.27$, $p=.007$, $\beta=0.11$, $z=2.78$), with more accurate responses with green compared to red buttons (93% vs 91%). Percentage of correct responses per condition is presented in Table 2.

The results suggest no language-based compatibility effect. However, one might argue that the lack of an effect in Experiments 4 might be because it lasted longer than Experiments 2 and 3 due to the doubling of the number of stimuli. Participants most likely became faster on the second part due to practice, so the responses may have become too fast to catch the effect. To test that, we conducted Experiment 5.

Experiment 5

In Experiment 5, we replicated Experiment 4, but with fewer critical stimuli and fillers. If our assumption is correct that the lack of an effect is due to practice developing over the course of the experiment, then we should observe a language-based compatibility effect in Experiment 5 in which we reduced the duration of the experiment. If practice does not play a significant role for the effect, then the interaction between referent colour and yes-button will not be significant, as in the previous experiments with materials involving fillers.

Methods

Participants. We recruited 207 volunteers (46 males, $M_{age}=23.58$, $SD_{age}=6.02$) via the mailing list of the University of Tübingen. Volunteers participated for course credit or in exchange for the opportunity to win one of four 20€ vouchers in the lottery. We then selected 180 participants who indicated that German was their mother tongue and who had not participated in Experiment 1–4 (40 males, $M_{age}=23.15$, $SD_{age}=6.93$).

Materials. For this experiment, we selected those words from the set of words used in previous experiments with the highest strength score associated with red and green. More specifically, we selected 10 "green" and "red" words ($M_{red}=6.20$ vs $M_{green}=6.19$), which did not differ in terms of strength ($t(18)=-0.9439$, $p=.365$), length ($t(18)=1.44$, $p=.168$), or frequency ($t(18)=0.80$, $p=.434$). We then took sentences from the stimulus set of Experiment 4 in which the selected words appeared as critical nouns. These sentences made up our experimental stimuli in Experiment 5. For the remaining stimuli, we used 20 sensical sentences and 40 nonsensical sentences from Experiment 4. Thus, there were 40 sentences that made sense and 40 sentences that did not make sense.

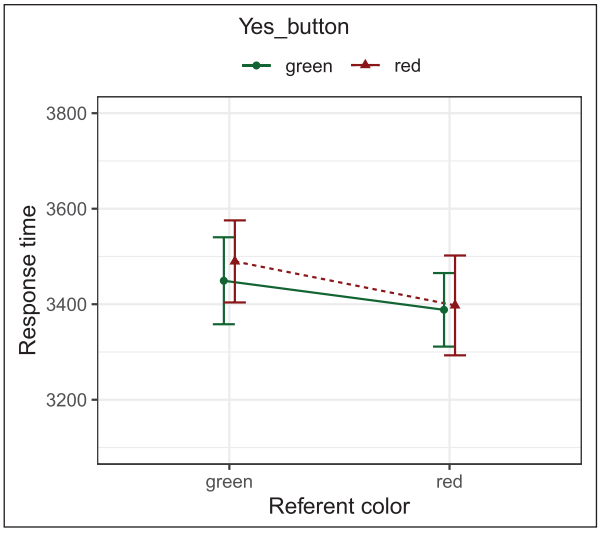

Figure 6. Response times in the sentence–sensibility–judgement task of Experiment 5.
Error bars denote 95% within-subjects confidence intervals calculated as recommended by Morey (2008).

Procedure and design. The procedure and design were the same as in previous experiments.

Results and discussion

We removed the data of 16 participants who did not pass the 80% accuracy threshold and equalised the number of participants per list, resulting in 140 participants (32 males, $M_{age}=23.09$, $SD_{age}=4.95$). For the response time analyses, we took into account only correct responses to experimental items. Non-accurate responses were eliminated (7.03% of relevant trials). After that, we followed the same procedure as before, specifically, we deleted absolute outliers by visual inspection of the data, which led to eliminating responses below 500 ms and above 20,000 ms, reducing the data set by 0.14%. We also discarded 4.93% of the data as relative outliers. The means of the final response times are depicted in Figure 6 (see also Table 3).

The base model with three fixed effects of sentence length, referent colour, and yes-button colour did not outperform the reduced model without length ($\chi^2(1)=0.83$,

Table 3. Means and standard deviations per condition in Experiments 5–7.

Referent colour	Experiment 5		Experiment 6		Experiment 7		Experiment 8	
	Yes-button		Yes-button		Yes-button		Yes-button	
	Green	Red	Green	Red	Green	Red	Green	Red
Green	3,449 (1,214)	3,388 (1,127)	3,484 (1,016)	3,598 (934)	1,166 (194)	1,224 (210)	1,131 (206)	1,182 (223)
Red	3,490 (980)	3,398 (1,002)	3,534 (924)	3,651 (1,094)	1,223 (199)	1,253 (202)	1,195 (213)	1,219 (202)

Table 4. The percentage of correct responses in Experiments 5–8.

Referent colour	Experiment 5		Experiment 6		Experiment 7		Experiment 8	
	Yes-button		Yes-button		Yes-button		Yes-button	
	Green	Red	Green	Red	Green	Red	Green	Red
Green	97%	95%	93%	91%	97%	95%	97%	96%
Red	90%	90%	91%	91%	95%	95%	96%	95%

$p = .363$, $\beta = 11.99$, $t = 0.93$), nor the reduced model without yes-button colour, $\chi2(1) = 0.14$, $p = .705$, $\beta = –8.99$, $t = –0.38$. In addition, it was no better than the model without referent colour ($\chi2(1) = 0.09$, $p = .765$, $\beta = 21.75$, $t = 0.30$). We compared the model with two fixed effects of referent colour and yes-button colour with the model with the interaction between referent colour and yes-button colour. The comparison showed that the interaction was not significant ($\chi^2(1) = 0.06$, $p = .806$, $\beta = –5.82$, $t = –0.25$). The accuracy analysis did not show any significant effect (all values of $p \geqslant .1$). The percentage of correct responses per condition is presented in Table 4.

Results of Experiment 5 showed no language-based compatibility effect, despite the reducing of the stimulus set.[2]

In the next experiment, we continue to explore the role of fillers, drawing attention away from colours in general.

Experiment 6

Experiment 6 was similar to Experiment 4, with one exception. Instead of fillers referring to objects with a typical colour other than green or red, we used fillers referring to objects with no colour. The logic of this manipulation is the following: Possibly, the colour-compatibility effect with red and green depends on the salience of this particular colour-pair and not on the salience of the colour dimension in general. Thus, using fillers referring to objects with typical colours other than green or red (similar to the previous experiment) might actually be a particularly strong measure for reducing the salience of the colour red and green. If so, and if the compatibility effect indeed depends on the salience of this particular colour pair, then it seems possible that using fillers referring to objects with no typical colours reduces salience of the colour dimension overall but restores the salience of the particular red–green pair. Thus, in this case, we would expect to see a re-emerging colour-compatibility effect in this experiment using fillers referring to objects with no particular typical colour.

Methods

Participants. We recruited 172 students (37 males, $M_{age} = 23.88$, $SD_{age} = 5.16$) from the University of Tübingen participated in the experiment for course credit. They also could participate in the lottery and win one of nine 20€ vouchers for the online shop. We sent the invitation to all students of the University of Tübingen, registered in the database. A total 134 participants (32 males, $M_{age} = 23.51$, $SD_{age} = 5.94$) reported that they were native German speakers and had not participated in Experiments 1–5 of this series. We took their data for further analyses.

Materials. The set of stimuli contained the same sentences as in Experiment 4, with one exception. We replaced the 60 sensical sentences in which we mentioned objects referring to different colours with 60 sensical sentences in which we referred to objects with no particular colour (e.g., *Die energiegeladene Nachbarin wird jetzt gleich Dinge in die Kiste legen./* "The energetic neighbour will put things in the box now."). In selecting fillers for Experiment 6, we again used the results of a pretest done for this study. Also, due to the insufficient number of resulting words, we used the results of another pretest we did for the follow-up study.[3] We took three types of words. First, we selected those words, which participants rated as not having a particular colour. Second, we took those words that had low association strength with any colour. Third, we selected the words where participants' responses had low agreement on some colour.[4]

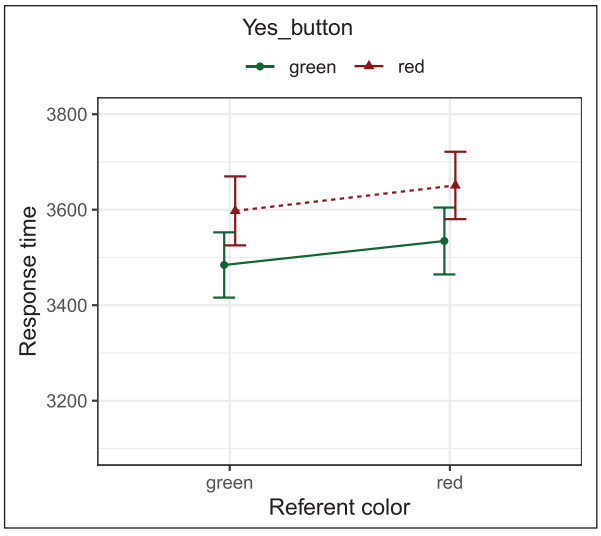

Figure 7. Response times in the sentence–sensibility–judgement task of Experiment 6.
Error bars denote 95% within-subjects confidence intervals calculated as recommended by Morey (2008).

Overall, there were 120 sentences that made sense (experimental items + fillers) and 120 sentences that did not make sense.

Procedure and design. The procedure and design were the same as in the previous experiments. There was one exception at the end of the experiment in the survey part. We added one more question to the survey about awareness and devices, namely, we asked people whether they had participated in a similar experiment before. We did that to separate people who already participated in Experiments 1–5 from those who participated the first time (see above). The experiment took approximately 25–30 min.

Results and discussion

As in the previous experiments, preparing the data for the final analyses, we first deleted the data of 10 participants who did not pass the threshold of 80% accuracy. Also, the number of data was equalised per list, so in the end the set for analyses comprised data of 120 participants (29 males, $M_{age}=23.44$, $SD_{age}=6.09$).

We only analysed responses to sensical items referring to typical green or red objects. Erroneous responses were omitted (8.74% of relevant trials). The outlier-elimination procedures from above eliminated responses above 20,000 ms and below 500 ms as absolute outliers (0.18% of relevant data) as well as 4.89% of the data as relative outliers. The means of the final set of response times are depicted in Figure 7 (see also Table 3).

The base model we used for further comparison consisted of three fixed effects: sentence length, referent colour, and yes-button colour, as well as random by-item and by-participants intercepts. We compared the base model to reduced models, where we consequently dropped one main effect after the other. Likelihood-ratio test showed that there was a main effect of yes-button colour ($\chi^2(1)=11.43$, $p<.001$, $\beta=-56.32$, $t=-3.38$), length ($\chi^2(1)=13.8$, $p<.001$, $\beta=27.87$, $t=3.96$), but no main effect of referent colour ($\chi^2(1)=0$ .29, $p=.588$, $\beta=24.18$, $t=.54$). Adding the interaction between referent colour and yes-button colour to the base model did not improve the fit ($\chi^2(1)=0.01$, $p=.936$, $\beta=1.33$, $t=0.08$). Thus, the pattern of the response time's results replicated the pattern obtained in Experiment 4.[5] Other models with interactions between length and the other factors were not better than the base model (all values of $p \geqslant .3$). The accuracy analysis showed only a marginal main effect of length ($\chi^2(1)=3.64$, $p=.057$, $\beta=0.04$, $z=1.96$), with more accurate responses for longer sentences. Other effects and interactions were not obtained (all values of $p \geqslant .1$). The percentage of correct responses per condition is shown in Table 4.

To summarise, the three sentence-based experiments involving additional filler sentences did not reveal any language-based colour compatibility effects. In addition, experiment length does not seem to be crucial. We can conclude that with sentences as materials the effects are fairly unstable, depending to a high extent on the linguistic context in which the sentences are processed. If the context makes salient not only the colour domain in general but the specific colours that are targeted in the compatibility effect, then a colour-compatibility effect is observed, otherwise not. This clearly suggests that the sentence-based colour-compatibility effect is not automatic and highly sensitive to context factors. We conducted two further experiments to investigate the context-sensitivity of the effect when words instead of sentences are used as materials.

Experiment 7

In Experiment 7, we did the same as in Experiment 4, but with words as material. In other words, we added filler words referring to objects with a typical colour other than red or green. We thus made the colour domain salient, but relative to Experiment 1, we reduced attention to the specific colour targeted by the compatibility effect.

Methods

Participants. We recruited 173 volunteers (31 males, $M_{age}=25.05$, $SD_{age}=7.32$) via the mailing list of the University of Tübingen and Prolific. Students from the University of Tübingen were compensated with course credit or could win one of seven 30€ vouchers in the lottery. Participation from Prolific was restricted to those who had not participated in the previous experiment and was compensated by 4 pounds. From the data set, 128 participants

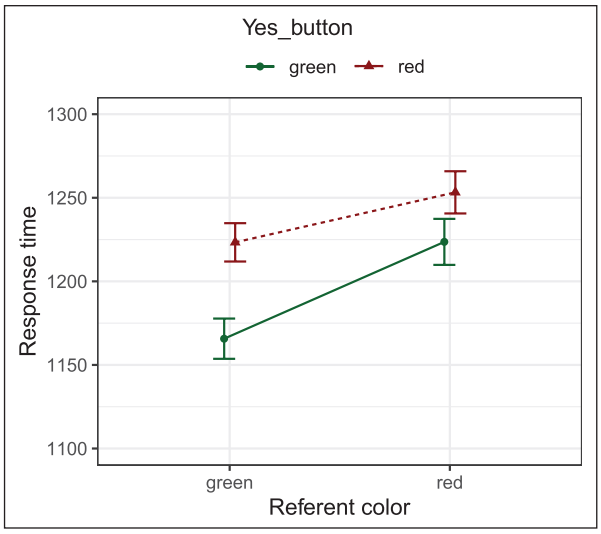

Figure 8. Response times in the lexical decision task of Experiment 7.
Error bars denote 95% within-subjects confidence intervals calculated as recommended by Morey (2008).

indicated that German was their mother tongue and that they had not participated in the previous experiments (26 males, M_{age}=24.71, SD_{age}=7.46).

Materials. The stimuli set comprised 120 words and 120 pseudowords. When creating the stimulus set, we first took the stimuli form Experiment 1 (60 green and red words and 60 pseudowords). As fillers, we used 60 words referring to objects with a typical colour other than green or red (e.g., honey or beer) from Experiment 4. Also, we added 60 pseudowords to balance the number of words and pseudowords. Pseudowords were created with the Wuggy software (http://crr.ugent.be/programs-data/wuggy).

Procedure and design. The procedure and design were the same as in Experiment 1. As in Experiment 1, a lexical decision task was used.

Results and discussion

We analysed the data in the same way as in the other word-based experiments. Elimination of the data of one non-accurate subjects and equalising the number of participants per list led us to a final data set of 120 participants (24 males, M_{age}=24.72, SD_{age}=7.56). Non-accurate responses and absolute outlier elimination (above 2,000 ms and below 500 ms) reduced the data set by 4.42% and 8.17%, respectively. About 3.76% of the trials were discarded as relative outliers. The means of the final response times are depicted in Figure 8 (see also Table 3).

The base model was better than the reduced model without word length ($\chi^2(1)=29.84$, $p<.001$, $\beta=21.01$, $t=6.22$), the reduced model without yes-button colour ($\chi2(1)=55.42$, $p<.001$, $\beta=–21.42$, $t=–7.41$), but not better than the model without referent colour ($\chi^2(1)=0.76$, $p=.383$, $\beta=11.14$, $t=0.88$). The model with an interaction between referent colour and yes-button colour significantly outperformed the base model ($\chi^2(1)=7.72$, $p=.013$, $\beta=–7.22$, $t=–2.49$), and was considered the last best model for further analysis. The interactions between referent colour, yes-button colour, and word length were not significant (all values of $p \geq .1$).

Analysis of simple effects showed the advantage of green yes-buttons over red yes-buttons for both "green" ($\chi^2(1)=50.41$, $p<.001$, $\beta=28.87$, $t=–7.13$), and "red" words ($\chi^2(1)=12.38$, $p<.001$, $\beta=–14.56$, $t=–3.52$). As in Experiment 3, the difference between the response times for the two button colours was larger for the "green" than for the "red" words. Also, the green yes-buttons were pressed faster in the presence of the "green" words than the presence of "red" words ($\chi^2(1)=4.51$, $p=.034$, $\beta=–31.01$, $t=–2.16$), which was not true for the red yes-buttons ($p=.255$). Overall, the interaction thus again shows a match advantage, which is mainly driven by the green words.

When we tested accuracy, we ran into convergence issues in the model taking into account random effects of participants and items at the same time. However, simpler models that took into account one of the random effects converged. Thus, we performed the accuracy analysis twice, once with the random effect of participants and another time with the random effect of items. Both analyses showed similar results. The accuracy analysis revealed a main effect of yes-button colour when participants were taken as a random effect ($\chi2(1)=7.71$, $p=.005$, $\beta=0.16$, $z=2.93$), and when items were taken as a random effect ($\chi2(1)=7.85$, $p=.005$, $\beta=0.16$, $z=2.83$). As before, participants were more accurate when they had to press a green yes-button than when they had to press a red yes-button (96% vs 95%). Other effects and interactions were not significant (all values of $p \geq .1$). The percentage of correct responses per condition is presented in Table 4.

In summary, we observed a language-based compatibility effect in this experiment, even though we added fillers referring to objects with different colours. In the next experiment, we investigate whether a full-strength word-based effect is observed when filler words refer to objects with no particular colour.

Experiment 8

In Experiment 8, we aimed at replicating the conditions from Experiment 6, but with words as material. In other words, we presented participants with the critical words as in Experiment 1, but added words referring to objects with no typical colour. We thus made the colour domain less salient compared with the word-based paradigm employed in Experiment 1.

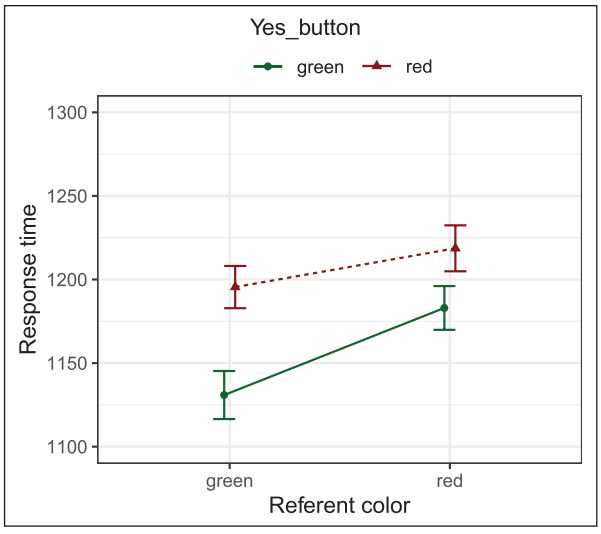

Figure 9. Response times in the lexical decision task of Experiment 8.
Error bars denote 95% within-subjects confidence intervals calculated as recommended by Morey (2008).

Methods

Participants. We recruited 148 volunteers (60 males, M_{age} = 26.82, SD_{age} = 10.39) via the mailing list of the University of Tübingen or from the Prolific platform. Volunteers recruited via the University mailing list participated for course credit or in exchange for the opportunity to win one of seven 30€ vouchers in the lottery. Participants from the Prolific online platform were compensated with 4 pounds. We selected 125 participants who indicated that German was their mother tongue and who had not participated in any of the previous studies (56 males, M_{age} = 26.96, SD_{age} = 8.92).

Materials. We used the stimuli from Experiment 7, except for the fillers. As fillers, we used the filler object words from Experiment 6 referring to objects with no particular colour (e.g., car or jacket).

Procedure and design. The procedure and design were the same as in previous experiments. A lexical decision task was employed.

Results and discussion

We applied the same procedure for data selection and trim as before. A total of 124 participants (55 males, M_{age} = 27.01, SD_{age} = 8.94) passed the threshold of 80% accuracy. After equalising the number of participants per list, we arrived at the final data set of 120 participants (52 males, M_{age} = 26.88, SD_{age} = 8.99).

Filler trials were discarded, as well as erroneous responses for the response time analysis. Errors were made in 4.22% of the trials. Visual inspection of the data resulted in eliminating responses above 2,000 ms and below 150 ms as absolute outliers. This reduced the data set by 7.43%. We additionally eliminated 4.07% of the trials as relative outliers. The means of the final response times are depicted in Figure 9 (see also Table 3).

The base model with three fixed effects of word length, referent colour, and yes-button colour was better than the reduced model without length ($\chi 2(1) = 27.7$, $p < .001$, $\beta = 20.66$, $t = 5.94$), and the reduced model without yes-button colour ($\chi^2(1) = 78.98$, $p < .001$, $\beta = –25.83$, $t = –8.87$), but it was not better than the model without referent colour ($\chi^2(1) = 1.16$, $p = .283$, $\beta = 14.07$, $t = 1.08$). The direction of the main effects was the same as in Experiment 1. The interaction between referent colour and yes-button colour was significant ($\chi 2(1) = 5.09$, $p = .024$, $\beta = –6.58$, $t = –2.26$). Other models with the interaction between word length and other factors were not better than the previous best model (all values of $p \geqslant .3$).

Simple effects analysis revealed the same pattern as Experiment 7. Participants showed faster response times when pressing green yes-buttons than when pressing red yes-buttons on reading both green ($\chi^2(1) = 62.94$, $p < .001$, $\beta = –32.12$, $t = –7.98$, $d = –0.29$), and red words ($\chi 2(1) = 20.68$, $p < .1$, $\beta = –19.18$, $t = –4.56$, $d = –0.17$). In addition, participants were faster when they pressed green yes-buttons on "green" words than on "red" words ($\chi 2(1) = 3.59$, $p = .058$, $\beta = –27.41$, $t = –1.92$). The same effect was not found for the red yes-buttons ($p = .370$).

The accuracy analysis showed only an effect of yes-button colour ($\chi^2(1) = 4.13$, $p = .042$, $\beta = 0.12$, $t = 2.09$). No other main effects and interactions reached significance (all values of $p \geqslant .2$). The percentage of correct responses per condition is presented in Table 4.

The results of Experiments 7 and 8 replicated the results of Experiment 1, even though filler words were added referring to objects with no specific colour or with colours other than green or red. Importantly, unlike in the experiments with sentences as materials, adding the fillers did not seem to eliminate the compatibility effect with words, in line with the hypothesis that word-based effects might be less context-dependent than sentence-based effects.

Post hoc analysis

The results of our eight experiments showed that a language-based compatibility effect is observed with different language stimuli (words and sentences) and persists even if the critical nouns do not appear at the end of the sentences such that the processing of the critical nouns and response selection are separated in time. However, the effect seems to be influenced by the proportionality of items in the experiment that refer to entities with a typical colour of green or red. Across our experiments, adding stimuli referring to entities without a typical colour of red

Table 5. The effect-size (Cohen-*d*) of the critical effect in Experiments 1–8.

Experiments							
1	2	3	4	5	6	7	8
–0.14*	0.06*	–0.04*	0.01	–0.01	0.002	–0.07*	–0.06*

Asterisks denote significant interaction in experiments.

Table 6. Means and standard deviations per condition for the pooled data in the post hoc analysis.

Compatibility	Stimulus type			
	Single_words		Sentences	
	Presence of fillers			
	Yes	No	Yes	No
Yes	1,190 (193)	1,229 (170)	3,417 (973)	2,909 (897)
No	1,206 (205)	1,257 (171)	3,415 (918)	2,963 (927)

or green reduced the effect, especially in the experiments with sentences (see Table 5).

To gather more information concerning the influence of this factor (presence of fillers: yes vs no), we combined all the data into one data set to summarise the results. We then collapsed the (critical) referent colour and yes-button colour factors into a compatibility factor (compatible vs non-compatible) and defined the stimulus-type factor (word vs sentence) and the presence-of-fillers factor (yes vs no). All of these were used as fixed effects in subsequent modelling. Unlike we expected, a step-by-step analysis of the models showed no three-way interaction of compatibility, stimulus type and presence of fillers ($\chi2(1)=1.30, p=.254$, $\beta=–5.15$, $t=–1.14$). Our analyses revealed that the following model described the pooled data better than others:

Response time ~ Compatibility × Presence-of-Fillers + Presence-of-Fillers × stimulus-type + Length + (1|Participant) + (1|item).

Thus, the compatibility effect seems to depend on the presence of fillers ($\chi2(1)=6.29$, $p=.012$, $\beta=–10.73$, $t=–2.51$), in the sense that the language-based colour-compatibility effect is reduced when fillers are included. In addition, the model revealed an interaction between stimulus type and the presence of fillers ($\chi2(1)=26.18$, $p<.001$, $\beta=–128.51$, $t=–5.16$), suggesting that the presence of fillers affects response times to sentences more than response times to words, in the sense that the difference between experiments with fillers and without fillers in word-based experiments is smaller than in sentence-based experiments. The means of the final response times are presented in Table 6. We ran two additional analyses, which analysed word-based and sentence-based experiments separately. The results showed an interaction between compatibility and presence of fillers both for words ($\chi2(1)=4.82$, $p=.028$, $\beta=–3.90$, $t=–2.20$), and for sentences ($\chi2(1)=4.72$, $p=.029$, $\beta=–14.13$, $t=–2.17$), as materials. Thus, post hoc analyses confirmed the previous results, indicating that the presence of fillers affected both types of stimuli, in particular, either reducing or eliminating the language-based colour compatibility effect.

We asked participants if they were aware that they had taken part in an experiment investigating colour processing in each experiment. Also, we asked them about the device they had been using. We extended the best model we used for the pooled analysis[6] with device and awareness as fixed effects to understand whether these two factors had any influence. The results showed no effect of awareness ($p>.1$). The effect of the device was only marginally significant ($\chi2(3)=7.11$, $p=.068$), indicating that participants worked faster with a computer and computer mouse than with a smartphone ($p=.041$).[7] Other models with the interaction between the device and other factors did not improve the fit (all values of $p>.7$). Thus, a post hoc analysis showed that devices only marginally affect the overall response time, but the device does not significantly influence the strength of the colour effect for different types of stimuli, regardless of whether fillers were present or not.

In all of our experiments, which showed a language-driven colour compatibility effect, the effect was driven by the differences in the green items whereas the red items showed a much smaller and sometimes even reversed compatibility effect. To follow up on this observation, we analysed red and green items separately in a combined analysis of Experiments 1–3, 7, 8. As expected, there was a highly significant main effect of compatibility for green items ($\chi2(1)=85.51$, $p<.001$, $\beta=–51.33$ $t=–9.26$, $d=0.15$). In contrast, for the red items, there was a reversed compatibility effect, which was smaller in terms of the size ($\chi2(1)=8.14$, $p=.004$, $\beta=17.15$, $t=2.85$, $d=0.06$). We think there is an obvious explanation for this asymmetry, namely that an effect of faster response times for green items counteracts the effect of compatibility. The reason is as follows.

Culturally, red colour is associated with prohibition and "stop" whereas green colour is associated with permission and "go." In addition, previous research has shown that green and "yes" are associated as are red and "no" (Dudschig et al., 2023). Considering that in our experiments all experimental trials required yes responses, this is another reason why pressing the green button was faster than pressing the red button, as reflected in the main effect of button colour reported above. Obviously, this main effect supports the colour-compatibility effect for green items (green button presses should be faster than red button presses), but counteracts the colour-compatibility effect for red items (red button presses should be faster than green button presses), so that it overcomes the effect of colour compatibility in red items. It should be noted,

however, that the pattern of results we obtained in our experiments suggests that in addition to an association between green and "yes," the typical colour of the referent also had an effect. Otherwise, it would be difficult to explain why we observed interactions between button colour and referent colour in Experiments 1–3, 7–8, instead of just a main effect of the button colour.

We also conducted a combined analysis with the 'no'-data included in addition to the 'yes'-data to further explore the relationship between button colour and response polarity ("yes" vs "no"). We combined the data of all experiments in one analysis, collapsing across referent colour. Thus, as fixed effects in this analysis, we included button colour and response polarity. We observed the significant interaction between button colour and response polarity ($\chi 2(1) = 21.56$, $p < .001$, $\beta = 40.78$, $t = 4.67$). The interaction indicates that the polarity of the response affects the colour button being pressed: green buttons are pressed faster than red buttons in the yes-button condition, but the other way around in the no-button condition, which is in line with the assumption of a cultural association between green and yes and red and no. Also, these results replicate earlier results mentioned above by (Dudschig et al., 2023).

General discussion

We conducted eight experiments in which participants made decisions about words or sentences referring to objects with a typical colour using coloured response buttons. Previous research has indicated that colour is distinct from visual properties such as shape or orientation, and is highly context sensitive.

This study was aimed at investigating the conditions under which colour information gets activated during language comprehension. We developed a paradigm that allows testing language-based colour-compatibility effects with different stimuli and tasks. In this paradigm, participants respond to linguistic stimuli referring to objects with a typical colour by means of clicking on coloured buttons that may or may not match the colour of the referent. One advantage of this paradigm is that objects' shape recognition is not involved in the decision-making process, as participants only see the coloured patch, but not the picture of the object as they do in a sentence-picture verification paradigm. Thus, with this paradigm, we are able to directly investigate the activation of colour information during language processing.

In Experiment 1, we used individual words as stimuli and a lexical decision task. The results showed that even when the task does not require deep semantic processing, the colour implied by the words affected responses given by clicking on a coloured button. Responses were faster when the button colour matched the implied colour compared with when there was a mismatch in colour. Thus, we observed a match advantage in this language-based colour compatibility effect. In Experiment 2, the same green/red words were presented in the context of simple sentences, where the critical noun was always mentioned at the end of the sentences, when participants made a choice. The task this time was a sensibility judgement task. The results showed the same pattern as in Experiment 1, namely, a match advantage in the response times to the sentences. We observed the same results in Experiment 3, in which the critical noun was placed before the sentence ended and the sentences described a vivid event presumably drawing attention away from the target object.

It is worth noting that the difference between Experiment 2 and Experiment 3 could be expressed from a pragmatical point of view. In Experiment 2, the critical noun was placed at the very end of the sentence, which tended to be more salient from a pragmatic perspective due to the so-called end focus, the tendency to put the important semantic element at the end (Flowerdew, 1992; Leech, 2016). Thus, in the sentences such as *Auf dem Tisch steht eine Gurke/On the table, there is a cucumber* the focus is on the object cucumber, whereas in the sentences such as *Der große Hilfskoch wird jetzt gleich die Gurke schälen*/*The tall assistant cook is about to peel the cucumber* the focus is likely to be the action—peel the cucumber. In other words, one could argue that the critical nouns were more salient in Experiment 2 than in Experiment 3 from a pragmatic perspective. We consider it noteworthy that we still observed the colour-compatibility effect in Experiment 3 even under these circumstances. To the best of our knowledge, there are no studies investigating the effect of topic-focus structure specifically on embodiment effects. In our view, this would be a promising line of future research.

The results of the first three experiments thus clearly support the idea that colour information gets reactivated when words are processed that refer to objects with a typical colour. The results are thus in line with predictions from the embodied-cognition view according to which the processing of words leads to a reactivation of experiential traces stemming from our interactions with the referents of these words. According to this view, the reactivated experiential traces include colour traces, which in turn affect the responses to the coloured buttons in the experimental setup. Our results further show that colour traces are not only reactivated when words are presented in isolation, but also when these words appear in sentences. This, of course, does not mean that comprehenders experientially simulated sentence meanings during comprehension, as the effects observed with sentences as materials may still be due to word-based effects. In other words, the words in the sentence may have activated the corresponding traces independent of sentence meaning. With the materials employed in our experiments, we cannot provide further information concerning this issue, as all of our sentences referred to objects with the same colour as the individual

words in the sentences did (Kaup et al., 2016). What we can conclude, however, from the first three experiments is that processing words referring to objects with a typical colour may lead to the activation of the respective colour information independent of whether the words are presented in isolation or in the context of sentences that describe events for which the colour of the target entity is not in the focus. Thus, sentence meaning does not seem to override any effects based on the words alone. Moreover, our results clearly show a match advantage and thus replicate the effects of Zwaan and Pecher (2012) and Mannaert et al. (2017).

Let us next turn to the question how stable these language-based compatibilities are. As we noted in the introduction, most previous studies did not include fillers beyond the necessities of the task. However, we know from other meaning dimensions in studies conducted in the context of the embodied-cognition framework that the presence of fillers can sometimes be decisive (Dudschig & Kaup, 2017). We, therefore, replicated Experiments 1 and 3 with the addition of different types of fillers in Experiments 4 to 8. More specifically, in Experiments 4 and 6, we added fillers to the experimental set that we used in Experiment 3. Both types of fillers—referring to objects of other colours or objects without a specific colour—resulted in the disappearance of the language-based colour-compatibility effect. In Experiment 5, we replicated the Experiment 4 but with fewer stimuli to see whether the duration of the experiments affects the language-based colour-compatibility effect. However, the results of Experiment 5 also showed no effect. We, therefore, are fairly confident that it is the presence of fillers and not the length of the experiments that is the decisive factor.

Interestingly, in Experiments 7 and 8, when we used individual words (as in Experiment 1) and added fillers, the effect reappeared. This seems to suggest that language-based colour-compatibility effects are more stable and less context-dependent for tasks involving individual words than for tasks involving sentences. However, our post hoc analysis combining the results of all eight experiments did not show a significant three-way interaction of compatibility, stimulus type (words vs sentences) and the presence of fillers. Rather what we found was a two-way interaction of compatibility and presence of fillers. We, therefore, must conclude that language-based colour-compatibility effects are context dependent for words and sentences as materials in the sense that the effects diminish or disappear when a relevant number of fillers are included in the material set. The results of the individual experiments moreover suggest that reducing the salience of the colour domain by including fillers referring to objects without a typical colour does not differ in preventing the colour compatibility effects from reducing the salience of the colour domain by taking attention away from the specific colours targeted in the compatibility effects (red and green in our case). In any case, the data do not support the idea that words would automatically activate experiential colour traces during comprehension.

In the introduction, we discussed that colour diagnosticity is one of the factors affecting how much colour is involved in object recognition and representation. In our experiments, we used words related to colour diagnostic objects, which were established using a pretest in which participants were asked to match the object to the associated colour and provide information on the strength of the association. However, this measurement of colour diagnostics might not be enough. For instance, for some objects colour is diagnostic in the sense that a specific colour even appears as a feature in a feature listing task in which participants must freely list the features of an object (Huettig & Altmann, 2011; Tanaka & Presnell, 1999). In an attempt to test the idea that a lack of diagnosticity partly explains the absence of the colour effect in Experiments 4 and 6, we reduced the set of stimuli by using only words referring to objects that have a more strictly defined colour diagnosticity. For this purpose, we selected only those words from our stimuli set that have a green or red colour associated with these words according to Nelson's association norms (Nelson et al., 2004). However, this also resulted in no interaction between referent colour and yes-button colour, neither in Experiment 4 ($\chi2(1)=0.90$, $p=.343$, $\beta=–21.52$, $t=–0.95$), nor in Experiment 6 ($\chi2(1)=0.34$, $p=.557$, $\beta=–17.35$, $t=–0.59$).

Up to now, we have framed our experiments and the observed results from the perspective of the simulation view on language comprehension. Our results indicated a strong context dependency of the language-based colour compatibility effect and thus speak against an automatic activation of colour experiences during the processing of linguistic stimuli referring to entities with a typical colour. As such, our results are not in line with strong versions of the simulation view of language comprehension according to which experiential simulations are a prerequisite of understanding language and are automatically activated during language comprehension (see Kaup et al., 2016 for a review of different versions of the simulation view of language comprehension; see also Huettig et al., 2020; Ostarek & Huettig, 2019). One interpretation of our results, therefore, is that language comprehension often but not always involves experiential simulations of the referent situation, in line with hybrid views of language comprehension (Binder & Desai, 2011; Dove, 2009). Another interpretation would be that comprehenders always build experiential simulations of the referent situation but that colour information is only included under certain circumstances, for instance, when it is made particularly salient by the choice of the linguistic materials. In this case, the results and their interpretation would be in line with the previous research about role of colour for object recognition (Bramão et al., 2011; Tanaka et al., 2001) or role of

colour representation in language comprehension (Huettig et al., 2020), which showed that colour information is context- and task-sensitive. The question arises whether our results can also be explained on the basis of purely amodal meaning representations (Fodor, 1975; Kintsch, 1988; McKoon & Ratcliff, 1992). We think such an account is possible as well. An amodal account of our results would need to assume that comprehenders activate amodal colour concepts when reading words referring to entities with a typical colour (not implausible), as well as when seeing coloured response buttons on the screen (less plausible but of course possible). In case of incompatible conditions, the two activated colour concepts would then interfere and slow down response times. Again, one would need to assume that colour concepts are only activated during language processing under certain circumstances. One way to distinguish between the two accounts would be to prevent participants from internally verbalising the colour of the response buttons by means of articulatory suppression. Future research will be needed to clarify this aspect of our results. We would like to note, however, that there is a slight asymmetry between accounts that attribute the language-based colour compatibility effect to experiential vs amodal meaning representations: If comprehenders indeed have reactive experiences of the referents during language comprehension and these experiences include typical colour information, then compatibility effects with coloured response buttons are clearly predicted. If comprehenders build amodal meaning representations and these include typical colour information, then compatibility effects with coloured response buttons are only predicted on the basis of the additional assumption that comprehenders internally verbalise the colour of the response button or activate amodal colour concepts corresponding to the colour of the response button. Thus, in our view, a purely amodal account is possible in principle but less parsimonious than an account that attributes the colour-compatibility effect to reactivated experiences.

To conclude, our findings suggest that comprehenders indeed activate colour information when processing different types of linguistic stimuli referring to objects with a typical colour. We consider it likely that colour information becomes available through experiential simulations of experiences during comprehension. However, our results also allow for an account in terms of amodal meaning representations. What our results clearly indicate, however, is that colour information is not activated routinely during comprehension but only under certain circumstances. The relevant factor seems to be the salience of the colour domain. If a relevant amount of the materials is not related to the particular colours targeted by the compatibility effect, then the compatibility effects diminish for words as materials and disappear for sentences as materials. What other circumstances modulate the appearance of language-mediated colour compatibility effects is a matter for future research.

Acknowledgements

We thank Chuck Clifton for comments and suggestions on the article.

Authors' note

The data sets generated for this study and the stimulus material can be found in the Open Science Framework (OSF) at [https://osf.io/n7m9u/] (doi 10.17605/OSF.IO/N7M9U).

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Deutsche Forschungsgemeinschaft (DFG) [Project ID 75650358, 419433647].

ORCID iD

Oksana Tsaregorodtseva https://orcid.org/0000-0002-0427-3992

Notes

1. To avoid the convergence problem that appeared with *lmer* models, we also run Bayesian regression models (using the *brms* package) with the maximal random effect structure justified by the design. Overall, with *brms*, we replicated the results obtained with the *lmer* models: we observed the same pattern of results in terms of interactions and main effects. We present a brief summary of the Bayesian analyses in the online folder together with the materials of these studies.
2. It could of course be argued that the lack of an effect in Experiment 5 is due to the decrease in statistical power following the decrease in the number of stimuli. To gain more information concerning the role of practice in this experiment, we conducted a post hoc analysis of the data from Experiment 4. For this analysis, we divided our data into 20 time intervals (bins). A 2 (reference colour) × 2 (yes-button colour) × 20 (time bins) analysis showed no interaction between factors ($p > .3$). The analysis showed only main effect of time bins ($\chi2(1) = 19.81$, $p < .001$), suggesting that participants indeed became faster over time. Thus, this post hoc analysis does not provide evidence that practice systematically modulates the interaction between reference colour and yes-button colour.
3. Another pretest was done for the follow-up study, where we were planning to use "yellow" words instead of "red" words to avoid the asymmetry we observed in this research. The pretest was performed in the same way as the pretest for this study (Experiment 1: *Materials*). Ten participants (8 males, $M = 26.6$, $SD = 5.82$) evaluated 184 words, referring to objects having a typical colour or not having a typical colour.
4. The results of the selection are in the online folder with the materials for all experiments.

5. As in the post hoc analysis for Experiment 4, we ran a post hoc analysis with the factors of reference colour, yes-button colour and time bins for the data of Experiment 6. As in Experiment 4, the results suggest that practice did not influence the effect: the interaction of reference colour, yes-button colour, and time bins was not significant ($p > .7$).
6. Response time ~ Compatibility × Presence-of-Fillers + Presence-of-Fillers × stimulus-type + Length + (1|Participant) + (1|item)
7. For calculating *t*-tests, we used the function ls_means of the lmerTest package.

References

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–660.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv*:1506.04967.

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, *15*(11), 527–536.

Borghi, A. M., & Riggio, L. (2009). Sentence comprehension and simulation of object temporary, canonical and stable affordances. *Brain Research*, *1253*, 117–128.

Bramão, I., Reis, A., Petersson, K. M., & Faísca, L. (2011). The role of colour information on object recognition: A review and meta-analysis. *Acta Psychologica*, *138*(1), 244–253.

Bub, D. N., & Masson, M. E. (2010). On the nature of hand-action representations evoked during written sentence comprehension. *Cognition*, *116*(3), 394–408.

Connell, L. (2007). Representing object colour in language comprehension. *Cognition*, *102*(3), 476–485.

Connell, L., & Lynott, D. (2009). Is a bear white in the woods? Parallel representation of implied object colour during language comprehension. *Psychonomic Bulletin & Review*, 16(3), 573–577.

de Leeuw, J. R., & Motz, B. A. (2016). Psychophysics in a Web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task. *Behavior Research Methods*, *48*(1), 1–12.

Dove, G. (2009). Beyond perceptual symbols: A call for representational pluralism. *Cognition*, *110*(3), 412–431.

Dudschig, C., & Kaup, B. (2017). Is it all task-specific? The role of binary responses, verbal mediation, and saliency for eliciting language-space associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(2), 259–270.

Dudschig, C., Kaup, B., & Mackenzie, I. G. (2023). The grounding of logical operations: The role of color, shape, and emotional faces for "yes" or "no" decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *49*(3), 477–492.

Flowerdew, J. L. (1992). Salience in the performance of one speech act: The case of definitions. *Discourse Processes*, *15*(2), 165–181.

Fodor, J. A. (1975). *The language of thought* (Vol. 5). Harvard University Press.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin.

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498.

Hirotani, M., Frazier, L., & Rayner, K. (2006). Punctuation and intonation effects on clause and sentence wrap-up: Evidence from eye movements. *Journal of Memory and Language*, *54*(3), 425–443.

Huettig, F., & Altmann, G. T. (2004). The on-line processing of ambiguous and unambiguous words in context: Evidence from head-mounted eyetracking. In M. Carreiras & C. Clifton (eds.), *The On-line Study of Sentence Comprehension* (pp. 187–208). Psychology Press.

Huettig, F., & Altmann, G. T. (2007). Visual-shape competition during language-mediated attention is based on lexical input and not modulated by contextual appropriateness. *Visual Cognition*, *15*(8), 985–1018.

Huettig, F., & Altmann, G. T. (2011). Looking at anything that is green when hearing "frog": How object surface colour and stored object colour knowledge influence language-mediated overt attention. *Quarterly Journal of Experimental Psychology*, *64*(1), 122–145.

Huettig, F., Guerra, E., & Helo, A. (2020). Towards understanding the task dependency of embodied language processing: The influence of colour during language-vision interactions. *Journal of Cognition*, *3*(1), Article 41.

Kaup, B., de la Vega, I., Strozyk, J., & Dudschig, C. (2016). The role of sensorimotor processes in meaning composition. In M. H. Fischer & Y. Coello (Eds.), *Conceptual and interactive embodiment* (pp. 58-82). Routledge.

Kaup, B., Lüdtke, J., & Maienborn, C. (2010). "The drawer is still closed": Simulating past and future actions when processing sentences that describe a state. *Brain and Language*, *112*, 159–166.

Kaup, B., Lüdtke, J., & Zwaan, R. A. (2006). Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? *Journal of Pragmatics*, *38*(7), 1033–1050.

Kelter, S., Kaup, B., & Claus, B. (2004). Representing a described sequence of events: A dynamic view of narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 451–464.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163–182.

Klein, G. S. (1964). Semantic power measured through the interference of words with colour-naming. *The American Journal of Psychology*, *77*(4), 576–588.

Kuperman, V., Dambacher, M., Nuthmann, A., & Kliegl, R. (2010). The effect of word position on eye-movements in sentence and paragraph reading. *Quarterly Journal of Experimental Psychology*, *63*(9), 1838–1857.

Laws, K. R., & Hunter, M. Z. (2006). The impact of colour, spatial resolution, and presentation speed on category naming. *Brain and Cognition*, *62*(2), 89–97.

Leech, G. N. (2016). *Principles of pragmatics*. Routledge.

Mannaert, L. N. H., Dijkstra, K., & Zwaan, R. A. (2017). Is colour an integral part of a rich mental simulation? *Memory & Cognition*, *45*(6), 974–982.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, *4*(2), 61–64.

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review*, *99*(3), 440–466.

Naor-Raz, G., Tarr, M. J., & Kersten, D. (2003). Is colour an intrinsic property of object representation? *Perception*, *32*(6), 667–680.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407.

Ostarek, M., & Huettig, F. (2019). Six challenges for embodiment research. *Current Directions in Psychological Science*, *28*(6), 593–599.

Rayner, K., Sereno, S. C., Morris, R. K., Schmauder, A. R., & Clifton, C., Jr. (1989). Eye movements and on-line language comprehension processes. *Language and Cognitive Processes*, *4*(3–4), SI21–SI49.

Scorolli, C., & Borghi, A. M. (2015). Square bananas, blue horses: The relative weight of shape and colour in concept recognition and representation. *Frontiers in Psychology*, 6, Article 1542.

Smith, L. B. (2005). Action alters shape categories. *Cognitive Science*, *29*(4), 665–679.

Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science*, *12*(2), 153–156.

Tanaka, J., Weiskopf, D., & Williams, P. (2001). The role of colour in high-level vision. *Trends in Cognitive Sciences*, *5*(5), 211–215.

Tanaka, J. W., & Presnell, L. M. (1999). Colour diagnosticity in object recognition. *Perception & Psychophysics*, *61*(6), 1140–1153.

Tucker, M., & Ellis, R. (1998). On the relations between seen objects and components of potential actions. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 830–846.

Tucker, M., & Ellis, R. (2004). Action priming by briefly presented objects. *Acta Psychologica*, *116*(2), 185–203.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625–636.

Wurm, L. H., Legge, G. E., Isenberg, L. M., & Luebker, A. (1993). Colour improves object recognition in normal and low vision. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(4), 899–911.

Yee, E., Ahmed, S. Z., & Thompson-Schill, S. L. (2012). Colourless green ideas (can) prime furiously. *Psychological Science*, *23*(4), 364–369.

Zwaan, R. A., & Pecher, D. (2012). Revisiting mental simulation in language comprehension: Six replication attempts. *PLOS ONE*, *7*(12), Article e51382.

Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. *Psychological Science*, *13*(2), 168–171.