Linguistic and Perceptual Categories in Colour Vision:  
A Critical Review*

David Bimler  

*Acknowledgements: Willett Kempton, for kind permission to quote from private remarks in an internet user-group; Valérie Bonnardel, Anna Franklin, Kimberly Jameson, Galina Paramei, and Debi Roberson for helpful discussions and feedback.

Part of the visual processing of colour stimuli is their classification into a small number of categories, with an effect on the pattern of subjective inter-stimulus similarities. That is, colour perception is categorical. The relationship between the categories of perception and those of the colour lexicon can be probed in various ways. “Natural experiments” exist: verbal categories vary between languages, and are not available at all to sufficiently young children. In addition, verbal responses can be discouraged while colour similarity is assessed. Here I review a number of recent studies of categorical perception. I argue that the majority are consistent with a moderate form of linguistic relativity, in which the acquisition of a colour lexicon does influence the perception of similarities. Subjects continue to manifest categorical effects under conditions of speeded search, which might be expected to preclude verbal responses. I interpret this as evidence about the early, pre-attentive stages of verbal processing.

Keywords: colour, categorical perception, color categories, visual processing

1. Introduction

The goal of this review is to explore the origins of the categories of colour experience, and their relationship to the categories of colour language. These have been the subject of considerable research. Any attempt to cover the field


Part of the visual processing of colour stimuli is their classification into a small number of categories, with an effect on the pattern of subjective inter-stimulus similarities. That is, colour perception is categorical. The relationship between the categories of perception and those of the colour lexicon can be probed in various ways. “Natural experiments” exist: verbal categories vary between languages, and are not available at all to sufficiently young children. In addition, verbal responses can be discouraged while colour similarity is assessed. Here I review a number of recent studies of categorical perception. I argue that the majority are consistent with a moderate form of linguistic relativity, in which the acquisition of a colour lexicon does influence the perception of similarities. Subjects continue to manifest categorical effects under conditions of speeded search, which might be expected to preclude verbal responses. I interpret this as evidence about the early, pre-attentive stages of verbal processing.

Keywords: colour, categorical perception, color categories, visual processing

1. Introduction

The goal of this review is to explore the origins of the categories of colour experience, and their relationship to the categories of colour language. These have been the subject of considerable research. Any attempt to cover the field
comprehensively would require a far longer review, since long-standing debates are at issue. The intention here is to summarize part of this body of research, with an emphasis on recent studies, and to address contradictions among them by promoting alternative explanations, with which the researchers themselves might not always agree. The hope is to learn something about the interplay between cognition and perception.

It is convenient to begin in the middle, by highlighting some counter-intuitive results obtained from a methodology of speeded visual search. Daoutis, Pilling and Davies (2006b) timed subjects while they decided whether a sample of some target colour was present somewhere in an array of samples of two distractor colours. A typical session might use three shades of green: Gr1, Gr2, Gr3. The distractors and target were chosen so that the three points representing them in a colour-coordinate scheme such as CIELUV, are collinear. Daoutis et al. found that if the target is at the end of the row of points, so that it can be distinguished from the distractors by applying a simple decision rule (e.g. a sample is the target if it has a value of 0.54 or more on the u* coordinate), then the task is easy. The presence of a target can be detected rapidly, as if it ‘pops out’; increasing the number of distractors in the array increases the detection time, but slowly. Conversely, if the target is the central stimulus in terms of colour space, flanked by the distractors (so it is not ‘linearly separable’ by a single decision rule), then detection time is longer, and increases more rapidly as distractors are added. The search is less efficient, seemingly requiring conscious attention to each stimulus in turn.1

None of this is problematic. However, a different picture emerges if the three hues are still collinear, but are chosen to lie in different categories (for instance, a blue, a purple and a pink). The purple is the central hue here, yet it can still be detected rapidly when it serves as the target, despite the lack of a simple colour-space criterion to distinguish it from the blue and pink distractors. It would seem that the observer can use a higher-order criterion for rapid detection: namely, membership of the colour category PURPLE, possessed by the target but not by its two neighbours in colour space.

1 Wolfe (1998) argues that the terms ‘efficient’ and ‘inefficient’ are more accurate than the older distinction between ‘parallel’ and ‘serial’ search, and avoids imposing premature assumptions about the underlying architecture.
Thus there are categorical differences among the pink, purple and blue stimuli (but not among Gr1, Gr2 and Gr3), allowing rapid search without the aid of linear separability. One explanation would be to interpret colour categories as basic perceptual qualities. That is, we could postulate that the qualities of blueness, pinkness \textit{et cetera} correspond to discrete neural pathways within visual processing; these \textit{allow} the production of a verbal label for each stimulus, but do not \textit{require} any labelling.

Alternatively, it could be that the observer tacitly assigns a label of ‘pink’ or ‘blue’ to each stimulus in the array, thereby providing the selection criterion in the form of the ‘purple’ label. Yet this implies that multiple stimuli can receive labels rapidly and in parallel (given the low contribution to search time from the number of stimuli in the array). This does not match our introspective impression that words reach awareness and are articulated in a serial, linear order. Even so, several lines of evidence link the verbal labelling of stimuli with rapid search, as well as with ‘categorical perception’, another characteristic phenomenon of colour categories. These will be explored at more length below, but they include the following:

Categorical properties are stronger in or exclusive to the right-hand half of the visual field, which projects to the left cerebral hemisphere where most linguistic functions are localized;

When category boundaries in the observers’ language differ from their locations in English, the perceptual effects vary accordingly;

Categorical properties are weakened or obviated by loss of language functions as a result of brain injury, or by the distraction of those functions by a verbal task competing with the primary visual task.

Another situation where language functions cannot contribute to categorical phenomena is when the observers are infants; here the evidence is still in flux.

Other visual-search procedures yield the same outcome: that the visibility of a target is enhanced when the distractors are in a different colour category (relative to target / distractor combinations that are equally distant in terms of colour space, but occupy a single category). In Pilling and Davies (2004), several target samples were present within a grid of distractors, so the subject’s task was to locate each copy (see also Daoutis et al., 2006a). Gilbert, Regier, Kay and Ivry (2006) simplified the grid to a clock-face arrangement of 12 samples, and asked the subject merely to indicate whether the left or right half
of this circle contained the single target. In Franklin, Pilling and Davies (2005),
distractors were not explicitly marked out, because the entire background (a
CRT display of uniform colour) was a distractor, with a small circle of a
different target colour appearing at one of eight set locations. In the last case,
one can speak of the target ‘popping out’ from the background, and actively
capturing attention; calling the observer’s attention to itself, even in the absence
of a directed search. An instruction to find the target was not possible here
since the subjects were infants, whose detection times were recorded via eye
movements.

2. Categorical Perception

The overall result here, that category membership can enhance the salience
of a target, can also be considered as a case of \textit{Categorical Perception} or CP.
CP is the general phenomenon in which the subjective similarities among
stimuli within some perceptual domain exhibit a pattern of distortion, in effect
partitioning the domain into a number of distinct categories. The categories
may correspond to verbal labels used to classify and describe the stimuli, but
this is not an essential part of the definition. CP was originally reported as an
auditory phenomenon, in the domain of phonetics, where it was defined as a
complete loss of discrimination within each phonemic category (Liberman et
al., 1957), so that dissimilarities within each category collapsed to zero, and
only inter-category discrimination was possible. That initial definition
explicitly excluded hue perception (Studdert-Kennedy et al., 1970). However,
the definition of CP has since been broadened to bring other perceptual
domains under its rubric, and contemporary discussions often cite the domain
of colour vision as a prime example.

Consider the sequence of hues specified in the Munsell system of
coordinates as 2.5BG, 7.5BG and 2.5B (leaving the other two variables of
lightness and saturation unspecified, as long as they are constant for all three
stimuli). The techniques used to project the Munsell coordinates into colour
space should ensure equal spacing within this sequence (Indow, 1988;
Newhall, 1940): the gap between 2.5BG and 7.5BG should be perceptually
equal to the gap between 7.5BG and 2.5B. However, the latter pair is usually
judged as more similar than the former pair. This corresponds to a category
boundary between the former pair, with English-speaking judges tending to label 2.5BG as green, and 7.5BG and 2.5B as both blue.

In general, we can select various pairs of colours, so that each pair is separated by the same distance in some model of ‘colour space’ — the three-dimensional gamut in which each point represents a different hue (in one such model, CIELUV space, distance is labelled $\Delta E_{uv}$, while in CIELAB space it is $\Delta E_{ab}$). Even so, a pair that are given the same colour label (green or blue or whatever), and fall into the same colour category, prove to be perceptually more similar than a pair that straddles a category boundary. From now on, in any reference to ‘distances’ without qualification, the phrase ‘colour space’ should be understood.

“Similarity” could be the outcome of a psychophysical measurement such as the accuracy of a ‘same / different’ decision, or the time taken in the decision. Alternatively, similarity might be elicited by direct judgment, or by comparisons in a triadic task (where stimuli are presented in groups of three, and the observer must decide which two are most similar, leaving the third stimulus as the ‘odd-one-out’).

To emphasize the evidence for some linguistic involvement, Gilbert et al. (2006) describe these phenomena as a “Whorfian effect”. The term acknowledges the argument that the linguistic categories we learn in childhood affect our similarity judgments to some extent (Whorf, 1956). More generally, the similarities we perceive among stimuli are modified by the inveterate human habit of verbalizing experience and reducing much of the diversity in our environment to a small lexicon of labels. This is a linguistic explanation for CP.

When English-speaking observers make a triadic choice among 2.5BG, 7.5BG and 2.5B, they predominantly choose the 2.5BG sample as the odd-one-out. It may be that English-speakers apply a ‘verbal strategy’ to simplify the task, pigeon-holing 7.5BG and 2.5B as ‘green’ and allowing their shared category membership to contribute to the judged similarity. These verbal strategies do not necessarily operate at a conscious level.

The Linguistic Relativity Hypothesis or LRH is a corollary. If two languages differ in their structures of colour categories — one might extend a single word across a region of colour where the other makes finer distinctions — then speakers of the languages should display corresponding differences in
their perceptions of colour similarity. Kay and Kempton (1984) presented the same three hues to speakers of Tarahumara, a Mexican language which makes no distinction between blues and greens, employing the same word *siyóname* to describe both. They perceived more balanced similarities, choosing 2.5BG and 2.5B with equal frequency.

The LRH does not mean that colour space is split into categories arbitrarily. It is noteworthy that various languages broadly agree in the way they lexicalize colour and locate category boundaries, allowing discussions of colour to be translated between languages with reasonable success. Nevertheless, differences do exist, which we can regard as natural experiments. The next section of this review will skim over a number of cross-cultural comparisons, comprising a tradition of ethnological fieldwork.

Two subsequent sections will deal with two more bodies of research where adult English-speaking observers were tested under conditions designed to disrupt or bypass verbal processing, such as competing tasks. Under some of these conditions, CP effects are weaker, as predicted by the linguistic account. But CP can still be discerned under other conditions, such as time pressure. I will argue that when CP is found, in apparent support of the innate account, in fact it highlights the counter-intuitive nature of verbal functioning and the shortcomings of our naïve beliefs about it.

One way of isolating similarity judgments from verbal strategies is to work with subjects too young to have learned any colour terms. Studies of infants and young children comprise a fourth body of research. Contrary to the predictions of the linguistic account, several of these studies report CP effects.

A fifth group of studies is an intersection between the first and fourth group. It consists of developmental studies of children in different cultures, as they acquire the colour lexicon of their respective languages.

In the course of this review, I will argue that the weight of evidence supports a form of linguistic account in which the postulated ‘verbal strategies’ for comparing colour similarity can operate below the limen of awareness. Rather than something invoked as a last resort, to tip the balance when visible similarities are too well-matched for a triadic judgment, we should regard the verbal strategies as *preceding* conscious consideration. Language has already modulated the similarities before we experience them. In this view, any attempt to disentangle verbal functioning, perception and experience cannot fully
succeed.

There is also much to be said for a rival innate position. This proposes that the category structure of colour experience is a ‘hard-wired’ feature of the human brain, a product of evolution; it predates any experiences we may receive from our culture about colour labelling (although there is room for it to be modified or fine-tuned by experience). Among other corollaries, this account predicts that children will display CP effects before they are old enough to have internalized the colour lexicon of their culture (as a special case, even infants should evince CP, if an appropriate test can be devised). Moreover, the CP effects should be the same in children from any culture and language.

Note that these rival accounts are not mutually exclusive. Nor are they exhaustive. Özgen and Davies (2002) showed that observers can be taught to subdivide an existing colour category and thereby acquire new categories, without learning distinct words for them. They suggest that in the usual case when the acquisition of new colour categories does coincide with the learning of new labels, then the causal role of the latter is indirect: the labelling focuses attention on the new inter-category boundary region, but it is not essential for the emergence of finer discriminations there or the distortion of similarities.

A few words of definition before we continue will reduce some ambiguity. Judgments of colour difference of the kind discussed here are often described in the literature as ‘discrimination’. The confusion here is that the same word is also used when thresholds of difference are measured. For instance, if the wavelength (\(\lambda\)) of a monochromatic light varies slightly, the smallest discernable change in \(\lambda\) is \(\Delta \lambda\). This threshold is a function of wavelength: it peaks in the yellow region of the spectrum (where the eye is least sensitive to wavelength change) and has troughs for red, green and blue hues. It is the wavelength discrimination function or \(\Delta \lambda(\lambda)\).

I will refer to ‘sensory discrimination’ and ‘perceptual discrimination’ respectively for judgments of threshold and of small subjective dissimilarities. Sensory discrimination involves hues separated by a discrimination threshold or a JND (Just Noticeable Difference), and the distances in colour space — generally measured by psychophysical procedures — are smaller than the supra-threshold dissimilarities of perceptual discrimination. Since JNDS provide the yardstick by which a typical colour space is constructed, and the
units for quantifying dissimilarity, it is necessary to assume that they are not subject to the phenomena of CP, and are coupled directly to the description of the visual environment captured by cone-cell responses. If they do undergo distortions at category boundaries, this leaves no independent grounds on which CP can be defined.

The CIELAB and CIELUV colour spaces do not lend themselves well to measuring distances in the units of discrimination. They are defined as simple transformations of ‘tristimulus space’ (which in turn represents the cone-cell responses to a stimulus, rather than its spectral composition). Both were designed with the goal that a constant inter-point distance, anywhere in the space, should correspond to the same sensory dissimilarity between the respective colours (prior to any distortions from CP effects). Neither, however, is fully satisfactory. CIELAB has come to be regarded as the appropriate choice for representing surface colours and the relationships between them, while CIELUV has advantages for representing self-luminous colours. In contrast, the Munsell system was constructed empirically, and is not restricted to simple transformations.

3. Cross-Cultural Comparisons

For quantifying dissimilarities larger than the threshold of discrimination, the triadic method is particularly suitable. Here three stimuli are presented concurrently, A, B and C, and the subject decides whether the pair A-B is more similar, or less so, than the pair B-C (or equivalently, whether A or C is less similar to B). This is the paradigmatic experiment used by Kay and Kempton (1984). Those authors chose their triads so that the distances A-B and B-C, measured as sums of just-noticeable differences, were balanced. English inserts a BLUE-GREEN category between B and C; Tarahumara does not. Only English speakers systematically saw that pair as less similar. They denied that their perceptions of similarity were swayed by labelling the stimuli with colour terms: that is, they were not consciously resorting to a verbal strategy to ease the cognitive burden. Subsequently the researchers covered C while calling attention to the shared blueness of A and B; then A was covered while the shared greenness of B and C was emphasized, in effect shifting the category boundary. This simple manipulation restored the equality of A-B and B-C.
Even so, the results are not conclusive. Failure to find CP effects in the Tarahumara is only negative evidence, and is not positive evidence for the absence of CP. The perceptual equivalence of A-B and B-C for Tarahumara subjects was inferred from their failure to agree on one pair as more similar that the other. This could also arise if they responded randomly (through misunderstanding the task, for instance, or by lacking our concern about colour).

Roberson, Davies and Davidoff (2000, Experiment 4) addressed this possibility by using two sets of colour triads. One set crossed the BLUE / PURPLE boundary, while the other crossed the boundary between NOL and WOR, two of the five basic colour categories in Berinmo (a language spoken in Papua New Guinea). NOL includes blues and most greens; WOR is roughly equivalent to YELLOW, but more extensive, so its boundary with NOL is different from the YELLOW / GREEN boundary in English. Speakers of both languages judged both sets, and in a kind of double dissociation, they displayed the hallmarks of CP (stretching of inter-category dissimilarities) only for the boundary marked in their own language.

Roberson, Davidoff, Davies and Shapiro (2005, Experiment 3a) extended this replication to a third language, Himba. Like Berinmo, Himba — spoken in Namibia — has five basic colour terms. DUMBU and BUROU are roughly equivalent to NOL and WOR, but DUMBU is even more extensive (including orange, brown and yellow-greens), shifting the DUMBU / BUROU boundary still further from the YELLOW / GREEN boundary. Speakers of Himba, Berinmo and English made triadic judgments, and displayed CP effects only for the boundary specific to their language. Himba and a second Namibian language, Kwanyama, also feature in a comparison with English by Daoutis et al. (2006a). Discussion of that study is deferred until Section 6 below, along with the visual-search procedure that it applied.

In a second strand of research, beginning with Laws, Davies and Andrews (1995), the Kay-Kempton paradigm was modified. Rather than Munsell papers as stimuli, this series of studies used Color-Aid® papers, which sample colour space in a less systematic way, without following a grid of cylindrical coordinates. In general, the stimuli of a given triad (call them A, B, C) are not spaced equally along a single continuum (hue or lightness). Rather than lying in a straight line, they form a triangle in colour space, so the subject has the
option of B as the odd-one-out, as well as A or C. Further, the colour-space
distances A-B and B-C are generally not balanced (i.e. the triangle is scalene
rather than isosceles). This difference in triad composition imposes a more
stringent test for CP effects. The crucial point that if one stimulus in a triad is
already isolated by sensory distances, then any cross-boundary distortions of
similarity will go undetected unless they are large enough to override this
prima facie isolation.

As well as sampling colour space less regularly than Munsell papers, Color-
Aid papers are also spaced more coarsely. Thus they probe a larger scale of
colour dissimilarity. Intra-triad distances (measured as \( \Delta E_{uv} \) in CIELUV
space), which were 20 to 40 in Kay and Kempton (1984), ranged from about
25 to 120 in Davies et al. (1998).\(^2\) One can argue that judgments become less
reliable and more prone to random fluctuations as distance increases, which
would further weaken the sensitivity of these triads to CP effects. Indeed, Laws
et al. (1995) were struck by the number of trials where a subject picked an odd-
one-out contrary to perceptual and lexical predictions.

Laws et al. (1995, Experiment 1) probed a category boundary not present in
English: that between the Russian colour terms, SINIJ and GOLUBOJ, or dark and
light blue. SINIJ and GOLUBOJ qualify as ‘Basic’ terms according to many
criteria (summarized and discussed by Paramei, 2006), so it is reasonable to
expect CP effects at the boundary between them. No systematic differences
emerged between English and Russian speakers. However, as we have seen,
the triads were not optimized for sensitivity. Paramei explores this point at
greater length. Notably, Witthoft et al. (2003) did find evidence for CP effects
in Russian speakers only (reviewed below in Section 4).

Laws et al. (1995, Experiment 2) proceeded to ask subjects to directly rate
the similarities between pairs of stimuli, on a 0-5 scale. Russian speakers
followed expectations: compared to English speakers, they provided lower
similarity ratings (relative to within-category ratings) for those pairs which
straddled the putative SINIJ / GOLUBOJ threshold. ANOVA analysis did not find

\(^2\) The limitations of CIELUV space mean that these \( \Delta E \) distances are only
approximations of what the ‘true’ sensory dissimilarities might be, prior to distortion
by CP effects. Unlike the situation with Munsell papers, the ‘true’ distances between
pairs of Color-Aid papers have not been accurately quantified.
the difference to be significant, but a statistical test that took the repeated-measures nature of the data into account might have done so.

Davies et al. (1998) compared speakers of English and Setswana (from Botswana) on triads which crossed various linguistic boundaries in both languages. As already noted, the composition of the triads reduced their responsiveness to distortions of inter-colour similarity. For the majority of triads, subjects responded in the same way, regardless of which stimulus was isolated by its colour label in their native language. The difference between the two groups was small, but significant.

Pilling and Davies (2004, Experiment 2) compared speakers of English and Ndonga (a language from northern Namibia). Stimuli were chosen from the purple-red-yellow region of colour space, to take advantage of the absence in Ndonga of basic terms distinguishing among ORANGE, PINK or PURPLE. The subject groups differed significantly in their responses to those specific triads where the two languages isolated different stimuli by their colour labels.

Finally, studies in this cross-cultural tradition also employ the concurrent and sequential forms of the Two-Alternative Forced Choice procedure (2-AFC). In its concurrent form, the 2-AFC test is simply a limiting case of an A-B-C triad, where either A or C is identical to stimulus B. The time taken for the observer to recognize the duplicated stimulus is an index of its similarity to the non-duplicate stimulus C.

When Witthoft et al. (2003, experiment 3) applied this procedure to 20 stimuli along a gradient between light and dark blue, consecutive pairs were equally discriminable for English-speaking subjects, whereas Russian speakers showed enhanced discrimination across the SINIJ / GOLUBOI boundary. Triads were displayed on a computer monitor. Winawer et al. (2007) obtained the same Russian-specific CP effect when they replicated the experiment with the same stimuli. They pointed out that this is a test of perceptual discrimination. It is possible to imagine that a conscious verbal strategy might tip the balance in a comparison between two triadic similarities, but not how it could help decide which pair of stimuli are identical. Winawer et al. also noted that colour categories contribute less when the judgments are made easier; the CP effect vanished when the non-replicated stimulus in each triad was easier to discriminate (four steps away along the 20-step sequence).

I have omitted a series of studies where speakers of various linguistic
backgrounds indicated the similarities among colour samples by sorting them into groups. Sorting is more complex than the triadic task; it is far from clear which decisions and comparisons are involved in the assignment of a colour to one group or another. In turn this raises doubts whether any cross-language differences are perceptual or cognitive in origin. Among them, Davies (1998) and Pilling and Davies (2004, Experiment 1) applied the sorting task with Color-Aid stimuli. There is certainly evidence for a linguistic influence on the responses, but this is not necessarily a marker of CP. It is that much harder to tell whether or not subjects consciously considered verbal labels for the stimuli. Even when speakers of different languages create similar groups, these may not reflect universal perceptual categories, if the distribution of stimuli within colour space was inhomogeneous (that is, if it possessed some intrinsic clustering).

This is aptly demonstrated by the case of a subject known as LEW, whose language skills have been compromised by a left-hemisphere stroke (Roberson, Davidoff & Braisby, 1999). LEW followed the normal pattern of odd-one-out choices in colour triads, and of responses in a X-AB task (see Section 4 below), including the apparent reduction of similarities across category boundaries. His normality extended to the disappearance of that distortion when the triad procedure was modified to minimize colour-name cues, in a replication of Kay and Kempton’s (1984) manipulation of conditions. This implies that LEW has an intact colour lexicon, which contributes to his judgments of colour similarity, although his access to the lexicon is poor when asked to name colours (he is quoted as declaring ‘that he did [always] know the name’ which he was unable to express).

LEW’s performance when sorting colours stands in stark contrast. His lack of overt access to his colour lexicon extended to an unawareness that a range of samples have a colour name in common, or could be connected by their proximity to a colour prototype. In lieu of such shortcuts for sorting, he is described as laboriously and repeatedly searching for the most similar pair of colour samples, out of all available (ungrouped) pairs. The resulting groups were coherent (Roberson et al., 1999, Figure 1, Appendices B and C) but cut across standard English boundaries, and linked colours by lightness level as much as by hue. LEW’s disjunction between sorting and triads nicely illustrates the multiple stages and the hidden complexity behind the seemingly
transient mapping of hues to colour terms.

4. A Working Hypothesis

The concurrent 2-AFC test was described above. In its \textit{sequential} form, this

\textsuperscript{3} It is instructive to compare LEW to the traditional Uzbeki women whom Luria (1976, pp. 24-30) asked to sort coloured wool skeins into groups. They showed little sign of recognizing any ‘family likeness’ across the range of colours covered by each Uzbeki hue name. To the contrary, they were reluctant to create groups from such colours, which were “not alike and could not be put together.” They preferred ‘analytic’ classifications of the samples into a large number of “small groups, combining them according to colour, saturation or lightness.”
is also known as the X-AB task. Here a stimulus is presented (X), then followed — after an inter-stimulus interval or ISI — by the stimuli A and B. One of these is a repeat of X, and the subject must judge which one. The similarity between A and B — the difficulty of the task — is indicated by the accuracy of these judgments, and the time taken to reach them.

Clearly there is an element of short-term memory involved in the X-AB task. Other studies have explicitly invoked short-term memory by using a sequential Same / Different task, with various delays (e.g. Uchikawa & Shinoda, 1996). These studies are not covered here since they do not draw cross-cultural comparisons. They clearly demonstrate the existence of colour categories, but provide no clues to narrow down their origins, because perception is not separated from memory.

Roberson et al. (2000, Experiment 6) applied the X-AB procedure, with a 30-second ISI, to English and Berinmo. They obtained similar results to those from the Kay-Kempton paradigm: the decision is easier when the samples are separated in one’s native language by a category boundary (BLUE / GREEN for English subjects, WOR / NOL for Berinmo). Roberson et al. (2005, Experiment 3b) extended the demonstration of CP effects to Himba.

The interpretation depends on how one understands short-term memory to function. It is possible that as well as an eidetic trace, subjects encode the stimulus in a lower-resolution way: as a verbal label. Then the judgment about stimulus identity is based partly on the equivalence or the difference between labels. Conversely, one could postulate that subjects represent the initial stimulus as some kind of eidetic trace, retaining its appearance with some accuracy, for comparison with the final stimuli. This memory trace would then become increasingly stereotyped as it fades with time, ‘regressing toward the prototype’ of the category it belongs to. When the distorted memory is compared to subsequent stimuli, the result is an increase and reduction in within-category and cross-boundary similarities respectively (Huttenlocher,

---

4 Two stimuli are presented sequentially, separated by a time delay, and the subject must decide whether they were the same or different. In those trials involving different stimuli, reaction time and accuracy are indices of the similarity between them. Less information accrues from trials where the stimuli are the same, but they are unavoidable for maintaining subject attention.
Hedges & Vevea, 2000). The emphasis here on the foci (prototypes) of categories, rather than the boundaries between them, bears comparison with the ‘perceptual magnet’ account of CP (Iverson & Kuhl, 1995).

Whether or not ‘regression to prototype’ is a feature of short-term memory, it cannot account for the CP effects observed in the studies surveyed in Section 3, where triadic-method comparisons and 2-AFC discrimination were simultaneous rather than sequential. The search for an economical explanation, for CP effects in general, leads back to some kind of verbal encoding of stimuli.

Verbal encoding implies that whenever we attend to a visual stimulus, it evokes a response from our verbal faculty, in the form of a descriptive colour term. However, informants in cross-cultural studies (Section 3) often insisted that they had eschewed the conscious use of a labelling strategy and based their decisions on perceptual grounds alone.5 Davies et al. (1998) debriefed subjects

---

5 Kempton’s 1991 remarks on the A-B-C test in Tarahumara and English are worth quoting at length. The ellipses are mine. From http://listserv.linguistlist.org/cgi-bin/wa?A2=ind9110D&L=linguist&P=R1957:

“First, as the speaker of a language subject to this perceptual effect, I would like to report that it is dramatic, even shocking. I administered the tests to informants in Chihuahua. I was so bewildered by their responses that I had trouble continuing the first few tests, and I had no idea whether or not they were answering randomly. In subsequent analysis it was clear that they were answering exactly as would be predicted by human visual discriminability, but quite unlike the English informants.

“An informal, and unreported, check of our results was more subjective: I showed some of the crucial triads to other English speakers […] All reported seeing the same effects. We tried various games with each other and ourselves like ‘We know English calls these two green and that one blue, but just looking it them, which one looks most different?’ No way, the blue one was really a lot more different. Again, the Tarahumara, lacking a lexical boundary among these colors, picked ‘correctly’ with ease and in overwhelming numbers.”

“Some of the triads which crossed hue and brightness were truly unbelievable, as it was perceptually obvious to us English speakers which one was the most different, yet all the visual discriminability data were against us (the article did not mention the hue/brightness crossovers for the sake of simplifying the argument in print).

“Our second experiment, like the original visual discrimination experiments, showed only two chips at a time. We additionally made it difficult to use the lexical
specifically on this point. Is it plausible that a stimulus could trigger a colour term without our awareness or volition, with the term never reaching consciousness, leaving its presence to be deduced from its behavioural effects only?

Conscious verbal production is clearly a linear process; even if the words and concepts are not articulated, they pass under the spotlight of awareness one at a time. The process is also slow: the latency between the presentation of a colour stimulus and the production of an appropriate word is typically 1 to 2 seconds (~1800 ms, for English speakers choosing between basic colour terms: e.g. Corbett & Davies, 1997, Table 9.1). However, verbal articulation is the final stage of a long process of semantic activity, the tip of the iceberg. Phenomena such as verbal priming and RSE (the Reverse Stroop Effect) demonstrate how much of that activity occurs outside of awareness, and prior to it. In network models of speech (e.g. Levelt, 1999), a percept can stimulate alternative candidates for ultimate verbal expression, as multiple nodes within a semantic network. These rival candidates propagate through the network in parallel, through conceptual, lemma and lexeme levels, gaining or losing activation until the most activated passes through the linear bottleneck of awareness. Interference among them can create slips of the tongue and the “tip-of-the-tongue” phenomenon.

The RSE is quite weak, but it attests to the automatic, non-conscious, parallel nature of much semantic activity. It requires a subject to read out a colour word while ignoring the colour of the letters it is printed in. The print colour appears to triggers an automatic verbal response from the subject. This may reinforce the conscious response (the word to be read) if the word and the print are congruent, resulting in a shorter response time. Conversely, if word and print conflict, the automatic response interferes with the conscious response, impeding its path to articulation and increasing response time. When the situation is reversed by asking the subject to ignore the lexical content of a word and announce the colour in which it is printed, the automatic response (i.e. reading the word) seems to be stronger, since it facilitates or hinders the categories. And we got visual discrimination-based results, even from English speakers. So there are ways to overcome our linguistic blinders.”
conscious response considerably more. This is the standard Stroop Effect. Davies et al. (1991) examined Russian speakers for Stroop and reverse-Stroop interference between GOLUBOJ and SINIJ.

In Section 6 we will encounter ways of measuring colour similarity that entail the division of attention across multiple stimuli, or response times of less than 1800 ms. They were intended to rule out verbal labelling, and thereby determine whether other explanations are needed for CP, such as hard-wired mechanisms. However, if those measurements reveal the hallmark of CP (similarity distortions across boundaries), one could retain a role for ‘verbal strategies’ by arguing that they function more rapidly and with a higher degree of parallelism than introspection would lead us to expect. This is possible if the assignment of colour labels occurs within the early stages of semantic classification, where it is concealed from awareness and volition. This line of argument allows one to interpret the presence of CP as information about the method of measurement. Namely, it is evidence that the measurement accesses or is influenced by some stage of semantic processing, even if this seems counter-intuitive.

5. Bypassing or Discouraging Verbal Processing

The short version of Section 4 is this: On seeing a coloured stimulus, a red book for instance, it is hard to prevent the word ‘red’ from passing through one’s mind at some level (which is not to say that the word modulates one’s perception of the stimulus; I suggest it is part of the perception). The common theme of the following studies is that they apply conditions which might bypass or penalize that verbal response. Thus they test whether verbal processing contributes to the categorical perception of colours. In most cases the subjects are English-speaking adults.

Roberson and Davidoff (2000, Experiment 1) introduced the idea of verbal distraction. They used the X-AB procedure from Section 4 to examine a series of four stimuli at hue intervals of 2.5 Munsell steps. Normally the series is discontinuous: the dissimilarity between the pair of stimuli straddling the BLUE / GREEN boundary (measured by proportion of correct trials) is greater than within-category dissimilarities. This difference disappeared when the subjects’ verbal mechanisms were occupied by a verbal task during the 5-second or 10-
The disappearance was not caused by the mental activity per se, since a non-verbal task had no effect. The authors obtained the same result when they switched from physical Munsell samples to computer-rendered equivalents on a CRT screen (their Experiment 2) and when they varied the verbal interference (Experiment 3). They speculated that verbal interference affects the outcome by competing for processing resources such as the phonological loop. They also attempted to distinguish between interference at the time of viewing the stimulus (encoding) and during the period of storage, though in this case (Experiment 4) the interference task — counting down by twos from a random number — appears more demanding and more numerical than verbal.

Pilling et al. (2003, Experiment 1) extended this study and generalized it to a sequential Same / Different procedure. In both procedures, the distraction of a verbal task removed the between-category advantage. To complicate the picture, however, the authors found in their Experiment 2 that when the verbal and spatial interference tasks were imposed in random order, rather than in blocks (so subjects had no advance warning which to expect), neither task removed the between-category advantage.

Witthoft et al. (2003, Experiments 1 and 2) ruled out short-term memory as an explanation for the CP effects by presenting the three colour patches of each X-AB choice concurrently rather than in sequence. In a second departure, they replicated Roberson and Davidoff (2000, Experiment 2), using the same CRT-displayed simulations of Munsell stimuli. The outcome was also the same: subjects showed a 'between-category advantage', which was eliminated by verbal interference but not by a spatial task.

As noted in Section 4, Witholt et al. (2003, Experiment 3) and Winawer et al. (2007) compared the performance of English and Russian speakers across the X-AB boundary, so at least some of the categorical effects eliminated by distraction (but observed in its absence) can be attributed to CP. The categorical characteristics remained after reducing the delay to 300 ms.
the putative SINIJ / GOLUBOJ boundary. Consistent with the idea that the colour terms SINIJ and GOLUBOJ were basic for the Russian speakers, their decisions were affected by verbal (but not spatial) interference: a between-category advantage disappeared. In contrast, English subjects showed no difference between baseline and either interference condition.

Wiggett and Davies (2008) created the verbal interference by presenting the first colour stimulus (the X of the X-AB task) in the form of letters, spelling out the name of a colour (BLUE or GREEN), not always concordant with the actual physical colour of X. Discordance between the lexical and physical names proved to remove the between-category advantage for identifying X at the AB stage of the task if the distractor comes from another category. In the classic Stroop effect (described in the previous section), involuntarily reading the letters of X evokes the lexical colour name within a subject’s verbal faculties, where it interferes with naming the physical colour. In this case, the evoked name appears to interfere with either the initial encoding of the physical colour into a short-term memory trace that contributes to later recognition; or to the retention of that trace. But to confuse the picture, there was no sign of interference if the colour-word format was used to present the A and B stimuli instead of the X, suggesting that the memory trace was not verbal in nature.

Pilling et al. (2003) postulated that the categorical effects arose not as products of perception, but through distortions during the ISI, as the short-term memory of the initial stimulus underwent ‘regression to prototype’ (Section 4). In this model, categorical distortion would have no effect on whichever stimulus in a pair was presented last, and it follows that the similarity between

<table>
<thead>
<tr>
<th>Study</th>
<th>Method</th>
<th>Stimuli</th>
<th>Boundaries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roberson and Davidoff (2000)</td>
<td>X-AB</td>
<td>Munsell CRT</td>
<td>BLUE / GREEN</td>
</tr>
<tr>
<td>(Exp. 1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Exp. 2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pilling et al. (2003, Experiment 1)</td>
<td>X-AB; S / D</td>
<td>CRT</td>
<td>BLUE / GREEN</td>
</tr>
<tr>
<td>Witthoft et al. (2003)</td>
<td>2-AFC</td>
<td>CRT</td>
<td>BLUE / GREEN</td>
</tr>
<tr>
<td>(Exp. 3)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winawer et al. (2007)</td>
<td>2-AFC</td>
<td>CRT</td>
<td>SINIJ / GOLUBOJ</td>
</tr>
<tr>
<td>Wiggett and Davies (2008)</td>
<td>X-AB</td>
<td>CRT</td>
<td>BLUE / GREEN</td>
</tr>
</tbody>
</table>
some pairs should depend on the order of presentation. The authors examined their data for order asymmetries, but the results did not conform to predictions.

6. Speeded Visual Search and “Pop-Out”

We turn now to a different procedure: visual search. The subject is presented with an array of coloured samples, of which most are distractors. The task is to find the sample (or samples) which match a target colour, as accurately and rapidly as possible. For instance, it might take 30 to 50 seconds to locate 19 targets amid 149 distractors in a 12-by-14 array (Pilling & Davies, 2004, Experiment 3). The more the targets stand out, the faster they are to find, so search time is an index of target-distractor similarity. This index decreases with distance in colour space (Carter, 1982), but there is a categorical component: targets from a different category from the distractors are found more rapidly than distance alone can explain (Daoutis et al., 2006b). There is also an asymmetry between good and less-good members of a category: magenta targets (red mixed with blue) stand out more against a background of pure-red distractors, compared to red targets against magenta distractors (Treisman & Gormican, 1988).

It is natural to bring these visual-search results under the aegis of the CP paradigm (e.g. Roberson, Pak & Hanley, 2008). If we interpret search time as an index of target / distractor similarity — more similar combinations taking longer to reliably identify the presence of the target — then these discrepancies in search time reflect the distorted similarities of CP. However, this account does not explain the category advantage — mentioned in Section 1 — when there two distractor hues which flank the target in colour space (linear non-separability). To avoid begging the question or prematurely making assumptions about mechanisms, in the context of visual search, ‘CP’ could stand for ‘categorical phenomena’.

Pilling and Davies (2004, Experiment 3) tested for CP effects by including distractors of the same (English) category as the target. Indeed, the task became harder for English speakers, and search time increased, compared to other-category distractors only. The authors also included a cross-cultural comparison with the Ndonga language. Ndonga does not distinguish the categories in question (RED, ORANGE, YELLOW), or mark any boundaries
between them, and the additional distractors did not produce so great an increase in search time for Ndonga speakers. As with the cross-cultural studies considered earlier (Section 3), this outcome militates against preclude innate or hard-wired categories as an explanation for the effects shown by English speakers.

English speakers showed the same CP effect for sequences of distractors and targets crossing the boundaries BLUE / GREEN, BLUE / PURPLE, RED / PINK and GREEN / YELLOW (Daoutis et al., 2006a). They also showed the expected decrease in search time for more widely-spaced sequences. The authors’ Experiment 1 was a comparison with the Namibian language Kwanyama, which subsumes red and pink hues within a single colour term, and blues, greens and purples within another term. No CP effects were found when Kwanyama subjects were tested for the first three of these boundaries.

Experiment 2 was a comparison with Himba across BLUE / GREEN and GREEN / YELLOW. It should come as no surprise that Himba subjects failed to display any CP effect across either boundary. In Himba, the sequence of distractors and targets traversing the BLUE / GREEN boundary fall within a combined ‘grue’ term BLOU,7 as we saw in Section 3. However, BLOU does not extend to yellow-greens, which are covered by DUMBU — a broader version of YELLOW. In particular, the targets and distractors used by Daoutis et al. (2006a) to test the YELLOW / GREEN boundary would all be described as DUMBU (Roberson et al., 2005).

Using stimuli described by English-speaking subjects as BLUE, Davies et al. (n.d.) did not replicate the reports of a SINIJ / GOLUBOJ boundary for Russian speakers. However, they found that Turkish speakers provided a significantly different pattern of RTs, consistent with a boundary between the Turkish words for ‘navy blue’ and ‘sky blue’ (LACIVERT and MAVI).

Does the speeded nature of the task exclude an explanation in terms of the verbal strategies which we considered previously? “The English search times are equivalent to about 600 ms per stimulus, which is probably too fast for all the stimuli to be named” (Daoutis et al., 2006a). Davies et al. (n.d.) pointed out that “[…] efficient search is a signature of pre-attentive processes”, combining

---

7 Transliterated as BUROU by Roberson et al. (2005), and as OSHIBUROU by Franklin et al. (2005).
this with an assumption that any access to verbal labels for stimuli is necessarily attentive. Pilling and Davies (2004) followed a similar argument to reach the conclusion that “Search is too rapid for name generation to be useful.” They therefore prefer an indirect role for colour terms, whereby a term is not directly involved during ‘on-line’ manifestations of the category (though knowing the term may have aided the acquisition of the category). I am promoting a contrary view: that although the processes of visual search may be rapid, pre-attentive and parallel, these are all compatible with the early stages of verbal processing.

This is supported by a series of ingenious studies which bypassed naming strategies by using the fact that language is largely a specialty of the left cerebral hemisphere, whereas for the vast majority of right-handed subjects, the right hemisphere’s language abilities are limited. A second relevant fact is that the optic nerves provide each hemisphere with direct access to stimuli that are presented in the contralateral visual field, but not the ipsilateral field. Drivonikou et al. (2007) simplified speeded-search to the case of a single target appearing in one of 12 possible locations within a uniform area of the distractor hue, while subjects were merely required to state whether it fell to the left or the right of the central fixation point. When the target fell in the right half-field, to be found by the left hemisphere, reaction times were shorter for target / background pairs that crossed category boundaries. The no-language condition was when the target fell in the left visual half-field where it registered first with the right hemisphere, and here the CP effect was weaker.

In Gilbert et al. (2006), the single target appeared somewhere in a ring of 11 samples of the distractor hue. A cross-category target was located more rapidly when it appeared in the right visual field (RVF). Moreover, this other-category advantage vanished when a verbal interference task was added to the visual search. In their Experiment 3, Gilbert et al. studied a post-callosotomy patient for whom there is less sharing of visual half-field information between the left or right hemisphere, and confirmed that the other-category advantage is restricted to the linguistically-able hemisphere.

This procedure has been extended to boundaries not present in English. The Korean colour lexicon includes the terms YEONDU (yellow-green) and CHOROK (green). As long as the cross-boundary target appeared in the RVF, it was relatively salient for Korean subjects — but not for English speakers, for
whom all stimuli were green (Roberson, Pak & Hanley, 2008).

Paluy et al. (2007) applied the procedure to subjects whose left-hemisphere language abilities were impaired (aphasics). Unlike normal controls, they showed no category advantage for the RVF. If anything, their responses were shorter for cross-category targets in the left visual field, suggesting that their intact right hemispheres had acquired some of the language functions that would originally have been lateralized in their left hemispheres.

Finally, as noted in the Introduction, Daoutis et al. (2006b) recorded search times in the situation of two distractor hues which bracketed the target hue in colour space. They found that even without the aid of linear separability, the presence of the target could be recognized rapidly if it was distinguished by category membership. The role of language in this effect becomes evident when the location of the target, relative to the initial fixation point, is taken into account. Drivonikou (2007) included this variable in a re-analysis of the data.

Table 3. Studies using speeded search to test for category phenomena (some with lateralization).

<table>
<thead>
<tr>
<th>Study</th>
<th>Languages (if not English)</th>
<th>Lateralization?</th>
<th>Boundaries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pilling &amp; Davies (2004, Exp. 3)</td>
<td>Ndonga</td>
<td></td>
<td>RED, ORANGE, YELLOW</td>
</tr>
<tr>
<td>Daoutis et al. (2006a, Exp. 1)</td>
<td>Kwanyama (Exp. 2)</td>
<td></td>
<td>RED / PINK, BLUE / GREEN, BLUE / PURPLE</td>
</tr>
<tr>
<td>Daoutis et al. (2006b)</td>
<td>Himba</td>
<td></td>
<td>BLUE / GREEN, GREEN / YELLOW</td>
</tr>
<tr>
<td>Gilbert et al. (2006)</td>
<td></td>
<td>Yes * †</td>
<td>BLUE / GREEN</td>
</tr>
<tr>
<td>Drivonikou et al. (2007)</td>
<td></td>
<td>Yes</td>
<td>BLUE / GREEN / PINK; BLUE / PURPLE</td>
</tr>
<tr>
<td>Reanalysis of Daoutis et al. (2006b)</td>
<td></td>
<td></td>
<td>GREEN / BLUE / PURPLE</td>
</tr>
<tr>
<td>Drivonikou et al. (2007)</td>
<td></td>
<td>Yes</td>
<td>BLUE / GREEN; BLUE / PURPLE</td>
</tr>
<tr>
<td>Roberson, Pak &amp; Hanley (2008)</td>
<td>Korean</td>
<td>Yes</td>
<td>YEONDU / CHOROK</td>
</tr>
<tr>
<td>Paluy et al. (2007)</td>
<td></td>
<td>Yes‡</td>
<td></td>
</tr>
</tbody>
</table>

* Verbal distraction task † Split-hemisphere patient ‡ Comparison with aphasic group
and found that the category advantage was limited to trials where the target appeared in the RVF.

According to this accumulating evidence, the visual search task can call upon the linguistic abilities of an undistracted left hemisphere, resulting in CP effects. If the pressure of time precluded linguistic processing, then neither (a) confining colour processing to the non-linguistic hemisphere, nor (b) verbal interference would make any difference to speeded search. As it is, (a) and (b) do affect the pattern of search times, in the same way. It would seem that some components of verbal processing are pre-attentive, parallel and rapid.

7. Per-Lingual Subjects: Procedures

One way to eliminate any linguistic contribution to judgments of colour similarity is to work with subjects who have not learned any colour terms. Thus studies which look for CP effects in infants or young children are a special case of the previous topic, but the literature is large enough to review them in a separate section. Obviously infants cannot listen to instructions or respond verbally. But research is still possible, because they soon become bored with a colour (or any other stimulus) after prolonged presentation; in search of visual stimulation, their gaze is directed elsewhere in their environment. The amount of time infants spend gazing at a new colour, when it takes the place of the first, indicates the dissimilarity between the novel colour and the familiarized stimulus. Infants possess full trichromatic vision when they are three or four months old, so this age often features in these studies.

The sequential X-AB procedure also translates into terms of preference for novelty. After colour X has been presented for long enough to familiarize the infant, colours A and B are presented side-by-side, while an observer records how long the infant’s gaze is directed to the left or the right. If A is a repeat of X, the ratio of the times spent gazing to each side is an index of the novelty of B, i.e. its dissimilarity from X. In the standard triadic paradigm, where A and B both differ from X, the ratio of gaze times indicates whether A or B is more dissimilar. If there is no prior exposure to X, then the test becomes a preference comparison — one can speak of the gaze times for A and B as indices of how much the infant prefers one to the other.

Speeded-search techniques are suitable for infants, if only a few distractors
are used. Indeed, use of this sparse form of the technique with infants (Franklin et al., 2005) predates its application to adults. Clearly it is not possible to give infants explicit instructions to find the target among distractors. Instead, the discrepant colour of the target is expected to increase its salience, causing it to ‘pop out’ and capture the infant’s attention, which in turn is measured by eye-tracking.

Another way to exploit the infant’s preference for novel visual stimulation is to present a composite display, containing the two stimuli simultaneously, or alternating between them. The rationale is that as the components of the display become increasingly different, the display becomes more complex and will retain the interest of the infant for longer. In this case there is no need to familiarize or habituate the infants.

Clearly there is no special role for short-term memory when a composite display is presented, whereas the sequential X-AB procedure does rely on a memory of the initial familiarized stimulus. The role of memory is increased in some of the studies reviewed below, which impose a delay after the initial stimulus and before the novelty preference. Unlike the situation with adults, it is not so important to disentangle memory from perceptual effects. If infants were shown to be dividing colour space into categories, but the discontinuities were shown to be a product of their memory, this would be just as important as “true” categorical perception. We would be left with innate categories to account for, embodied somewhere in the neural substrate of colour cognition, whichever route they took to influence the infants’ behaviour.

8. Pre-Lingual Subjects: Studies

Bornstein (1975) collected gaze times for pairs of eight monochromatic lights. The results correlate well with saturation, which varies considerably across the spectrum. For instance, monochromatic yellow light is less saturated than blue or red lights, and is also less attractive to infants. To account for departures from this general trend, Bornstein posited a secondary trend to prefer better category examples. Four category prototypes (red, yellow, green blue) were more preferable than saturation would explain, while four poor examples (red / yellow, etc.) were less preferable.

Bornstein, Kessen and Weiskopf (1976, Study 1) familiarized infants to a
light B. Then their gaze times were recorded while alternating B with two novel colours, A and C, with wavelengths higher and lower than B by the same amount. For instance, B = 480 nm, A = 450 nm, C = 510 nm. Bornstein et al. asked whether A or C was gazed at longer, i.e. which was more dissimilar from B. For seven triads which traversed the category boundaries BLUE / GREEN, GREEN / YELLOW and YELLOW / RED, the categorical isolate was also more dissimilar. The authors concluded that ‘hard-wired’ colour categories are part of the human genetic endowment. Three other triads showed no significant asymmetry in the region of the YELLOW / RED boundary, suggesting that these categories are less distinct. No evidence was found for a category boundary at the short-wavelength end of the spectrum, between blue and violet. Franklin and Davies (2004) noted that the infants did not distinguish the habituating hue from the more-similar test hue in many of the triads, and gave the same length of gaze time to both.

The use of monochromatic stimuli creates technical difficulties (Werner & Wooten, 1987). The first is that any asymmetries in the data are complicated by an overall preference gradient. Second, the human retina does not respond equally across the spectrum to a given change in wavelength: as noted already, different zones of the spectrum require a different change in wavelength before the change in the stimulus is discernable. That is, the “wavelength discrimination function” \( \Delta \lambda(\lambda) \) is not flat. This undercuts the authors’ attempt to create balanced intervals within each ABC triad, and means that the infants’ indications of novelty are not immediately useful. It would be an uninteresting kind of ‘quasi-CP’ if the asymmetry in each triad was merely a reflection of non-uniformity in the retinal mechanics of trichromatic vision.

Franklin and Davies (2004) addressed these technical difficulties by replotting the stimulus combinations from Bornstein et al. (1976) into CIELUV space. For several of the triadic combinations (though not all of them), the responses from the infants were now in line with the resulting sensory distances. This highlights the need for more rigorous research along these lines.

Catherwood, Crassini and Freiberg (1989) introduced a computer CRT to present the stimuli, and provided the infants with a preference choice between two simultaneous hues, rather than three options sequentially. They focused on a single triad of hues — two greens (GA and GF) and a blue (BF). After being familiarized to the anchor GA, infants spent more time looking at BF than at
There is no need to invoke the blue / green boundary as an explanation for this, since it turns out that in terms of distance in CIELAB space, GA was three times as far from BF as from GF (reconstructing CIELAB locations from chromaticity coordinates provided in the paper). This was although the stimuli were equally spaced along the spectrum — dominant wavelengths were 548 nm for GF, 515 nm for GA and 482 nm for BF — and reinforces the problem with using wavelength as a metric of colour distance.

The first researchers to use Munsell papers were Gerhardstein, Renner and Rovee-Collier (1999). Infants saw two stimuli at a time, each pair comprising a composite display; they had learned to tug on a string (by kicking their legs) once the display exhausted its appeal. Their response times could be interpreted as estimates of inter-stimulus dissimilarity. These times increased with distance in Munsell space, but did not differ between cross-boundary and within-category pairs. In contrast, when adults rated the same similarities, they judged the cross-boundary pair as less similar than their within-category ratings would predict. The categories were red and orange. Note that orange is a ‘secondary’ or ‘derived’ colour term, and one can argue on various grounds that it has a different status and a lesser claim to be innate (for instance, subjects cope without it in colour-naming studies, where they can describe ‘orange’ stimuli as combinations of red and yellow). Thus it is possible that infants only notice boundaries such as red / yellow, between ‘primary’ terms.

Davies and Franklin (2002) pointed out that the stimuli in Gerhardstein et al. (1999) varied in lightness as well as hue, introducing the former as a potential confounding factor. Second, they were seen under incandescent light, rather than the day-like balanced-spectrum Illuminant C for which Munsell papers are designed. Indeed, when Davies and Franklin repeated the control experiment of eliciting pair-wise similarity ratings from adults, but using incandescent lighting, the evidence for adult CP was less clear-cut — that is, the contrast with the infants was weaker. Finally, the same-category pair may have crossed a category boundary itself, since one of the two putatively red stimuli was in practice identified as ‘pink’.

Franklin and Davies (2004) went on to conduct their own research. They continued the use of Munsell papers, varying the stimuli in hue only while keeping lightness and saturation constant (Value = 6, Chroma = 10). Their Study 1 looked for evidence of blue / green and blue / purple boundaries. In
both cases, they selected within-category and cross-boundary stimulus pairs, separated by 5 and 7.5 Munsell hue steps. They returned to the X-AB paradigm, with a 5-second ISI after familiarizing the subjects to stimulus X and before presenting the choice of A and B.

The infants’ dissimilarity data contained purely categorical effects, showing no dependence on colour-space distance at all. That is, subjects preferred the novelty of B if a boundary separated it from X, but their preference grew no stronger as the distance increased. Conversely, they showed no preference for B if it occupied the same category as X, even when 7.5 hue steps separated the two. The same pattern emerged across the RED / PINK boundary (Franklin & Davies, 2004, Study 3).

We turn now to the visual-search approach to quantifying dissimilarity, as in Section 6 (Franklin et al., 2005). Within the background of a second hue, a circle of the target hue appears at one of eight possible locations. The time taken to fixate on the target is recorded by an eye-tracker and indicates how strongly it stands out against the background, attracting the subject’s gaze by virtue of its contrast. Two hue pairs were assessed, involving a fixed target (GT) and two background screens (GS and BS). The dissimilarity between GT and GS was not zero — that is, infants did fixate a green target against a green background in finite time, suggesting that they were making within-category discriminations. More importantly, the cross-boundary dissimilarity between GT and BS significantly exceeded this within-category value, although the two target / background separations were the same in CIELUV space (40ΔE) and nearly the same in Munsell hue steps. Adults showed the same category advantage.

This measure of hue-pair distance (Franklin et al., 2005) is not necessarily symmetric. That is, we do not know whether a blue target on a green screen would attract an infant’s gaze as rapidly as the green target on the blue screen. Certainly adults can display asymmetries in colour search (Treisman & Gormican, 1988). There is the complication of preference bias. This is not intended to minimize the technical difficulty or the laborious nature of this kind of research with infants. Nevertheless it will be reassuring when these measurements are repeated with other hues as the target, and other hue pairs crossing other boundaries.

Following Gilbert et al. (2006), Franklin et al. (2008a) went on to test
whether this infant category effect is lateralized, this time specifying stimuli so that target / boundary pairs were equidistant in Munsell units rather than in CIELUV space. Unexpectedly, they found that the same-category target took significantly longer to be noticed when it appeared in the left visual field, whereas same-category and cross-boundary times were the same when the target appeared in the RVF. It appears that the mechanisms producing this prelingual category boundary in infants are based in the right cerebral hemisphere. The implication is that they are not directly related to colour categorization by adults, which has a left-hemisphere locus.

9. Implications of Categorical Colour Vision in Infants: Features vs. Categories

As noted in the previous section, three studies have reported persuasive evidence that infants display categorical phenomena (Franklin & Davies, 2005; Franklin et al., 2005; Franklin et al., 2008a). Another study failed to observe CP in infant colour vision, but it is open to question whether it examined suitable conditions where CP might actually apply.

It appears that infant vision can recognize ‘greenness’ as a discrete feature, one that attracts the infant’s attention when a target ‘pops out’ from a scene that
otherwise fails to meet the threshold of greenness. However, Section 2 noted that the phenomenon active in visual search is not necessarily identical to classic ‘categorical perception’, with its hallmark of distorted cross-boundary similarities in an A-B-C triad or a 2-AFC task. Presumably some other categories of colour are also recognized by infant vision as features that can guide visual search. There are reports that infant vision can also process colour as a continuously varying domain, in parallel with this extraction of discrete features (e.g. Gerhardstein et al., 1999).

If these features are experienced in the same way that adults experience vivid, distinct percepts of colour — for instance, the special quality of ‘redness’ we obtain from a good example of RED — then colour language would be easy to learn. The infant would merely need to translate ‘mentalese’ into common language, as it were, by learning the accepted label for each percept. In practice, however, the acquisition of colour labels seems to be lengthy and laborious (e.g. Bornstein, 1985; Roberson et al, 2004), which hardly fits the idea that colour categories are manifestations of innate experiential qualities. Pitchford and Mullen (2002) have argued that a general developmental delay in acquiring abstract object attributes includes and explains this tardiness with colour names. *The Journal of Experimental Child Psychology* devoted a special issue to this question (Volume 94 (4), 2006).

The apparent paradox is resolved when we learn that the mechanisms responsible for pre-lingual categorical pop-out operate in the right hemisphere (Franklin et al., 2008a). In contrast, the cortical systems involved in the infant’s acquisition of language are concentrated in the *left* hemisphere, restricting their access to these categorical features being detected. This suggests a model in which the innate right-hemisphere mechanism that detects a given colour feature is not directly linked to the nearest quality or percept: ‘redness’, for instance, or ‘dumbu-ness’. The percept comes *after* acquiring that colour term, and learning the range of hues to which it can be applied (perhaps assisted by the feature-detection).

This dual model poses a number of questions. Primarily, what purpose do these colour features serve? Why do they disappear later? (there is little evidence for colour categories when an adult’s right hemisphere is presented with the visual-search target: Gilbert et al., 2006). At this point one can only speculate. It may be that these features provide an infant with a basic...
framework for structuring colour experiences, and for coding them in a suitable form for memory, until the accumulation of experiences provides a more nuanced coding scheme, tailored to the child’s visual environment.

The number of innate colour features is also unknown. We have already encountered the intuitively-appealing hierarchy of ‘basicness’, which ranks the Primary (elemental) categories of RED, YELLOW, GREEN and BLUE as more basic than the ‘secondary’ or ‘derived’ categories such as ORANGE, PINK and PURPLE (e.g. Kay & McDaniel, 1978). A parsimonious model would include the primary categories as innate features, serving as landmarks or cardinal directions in colour space. It is not necessary to assume the secondary categories to be hardwired as well, since their possible locations in colour space are already restricted to the gaps between the primaries. Thus one suggestion, when infants showed no sign of CP between RED and ORANGE (Gerhardstein et al., 1999), was that the boundary of a secondary category may be the wrong place to look for a CP effect. However, of the three boundaries where Franklin and Davies (2005) found CP effects, using the same X-AB task as Gerhardstein et al., two involved PINK and PURPLE. This would suggest that secondary categories are innate after all.

Cross-cultural data reveal languages that impose an ‘unnatural’ systems of categories: systems that lack some of the thresholds between innate features, thereby merging categories (e.g. BLUE with GREEN), or insert additional boundaries (e.g. between SINU and GOLUBOJ), or displace boundaries from their natural locations in colour space (e.g. BOROU and DUMBU). Clearly this would be quite difficult if infants did in fact experience categories as vivid percepts, comparable to an adult’s ‘redness’. Himba children, for instance, would have to learn to suppress the innate percept of ‘greenness’ evoked by yellowish-green hues (as well as by focal green and bluish-green), and instead to experience the percept of ‘dumbu’ that yellow hues also evoked for them.

Nevertheless, English is often regarded as a ‘natural’ language, at least by English-speaking researchers. It is presumably closer than others to the grounding of neural reality; its lexicon retains more of the innate features and thresholds. When Franklin and Davies concluded (2004, p. 375) that “… 4-month-old infants seem to have adult perceptual categories, at least to some degree”, they implicitly specified ‘English-speaking adults’. It may that ‘natural’ colour lexicons are easier and faster to learn, especially if any
unlearning and unlearning is involved (Bornstein, 1985). Thus some researchers have looked at older children who can receive instructions, but are still acquiring competence in colour language. This also finesses the considerable technical difficulties of studying colour vision in infants.

10. Developmental and Cross-Cultural Studies

The colour lexicon is not learned all at once, in a sudden transition. Even among the 11 basic terms of English, some tend to lag behind others by up to a year, though there is no single standard sequence for acquisition. Production and comprehension are different. This leaves some uncertainty about the age when children begin to internalize the colour lexicon. After reviewing a substantial literature, Roberson, Davidoff, Davies and Shapiro (2004) concluded that competence is seldom found in children younger than two.

Daoutis et al. (2005a, Experiment 1) set an upper limit when they compared visual search between English and Kwanyama, as summarized in Section 6. There the subjects were children ranging from 4- to 7-year-olds. The relevant point here is that in both languages, the youngest children had learned the colour lexicon as well as the oldest (i.e. they were equally able to produce the appropriate consensus label for each target and distractor hue). All ages of English-speakers, and no ages of Kwanyama, displayed the CP effects expected from English categories.

Roberson et al. (2004) conducted a longitudinal study of colour-lexicon acquisition in English and Himba children aged 3 to 6 (see also Roberson et al., 2006). Children from the two cultures learned colour words in the same gradual fashion, in parallel trajectories, with no sign that the process for English children was facilitated by an innate structure. The Himba children

8 In the RED / PINK triad, many Kwanyama children who labelled the two reds as OSHITILYANA were unwilling to extend the term as far as the pink stimulus, which they preferred not to label at all (“I don’t know”). This could be a sign of graded category membership, ranging from good examples of OSHITILYANA (the reds) to a poor example (the pink). The same pattern appeared in the BLUE / PURPLE triad where the two blue stimuli were good examples of the ‘grue’ category OSHITWIMA, while the purple stimulus was not a good example for many children, who gave no label for it.
lagged consistently behind English speakers, but this can be explained by the higher salience of colour in modern literate cultures, and the educational investment it receives.

As part of the six-monthly examination of the children, an array of 22 samples was used to test their colour memory. For English children, as they grew more competent in colour terms, samples which were focal colours in English showed an increasing memory advantage (i.e. fewer mistakes when matching a target after a 5-second delay) over samples which were only focal in Himba; and vice versa for Himba speakers. Another way in which the two groups diverged was that their patterns of mistakes became increasingly dominated by confusions between samples with the same colour term. In comparison, for the children knowing fewest terms (in both language groups), the inter-sample confusions they made were associated with the inter-sample distances in colour space.

Note, however, that these trends can be explained by the use of colour terms as memory aids, and do not establish that the terms are involved in perception. Bonnardel and Pitchford (2006) used a task with no memory component: children were given eight targets and asked to judge which target was the closest for each of 100 samples. In effect they partitioned colour space into eight ‘catchment areas’, each centred on one of the targets, separated by ‘watersheds’. Targets were foci of chromatic colour categories. When the subjects were split into three bands according to their knowledge of the colour lexicon (plus a fourth band composed of adults), differences between bands were very small, and the ‘watersheds’ consistently matched consensus category boundaries. At the relatively broad scale probed in this experiment, colour similarities were not greatly altered by knowledge of colour terms.

Franklin, Clifford, Williamson and Davies (2005, Experiment 1) tested 60 English toddlers with a mean age of 36 months. In the age band 2 to 2½ years old, about half the children could understand a given basic colour term; after that the proportion rose rapidly, to level off in the age band 3 to 3½ years.

The same toddlers were asked to label the stimuli they saw in a X-AB task with a 5-second ISI (and age-appropriate wording). 37 recognised a category boundary built into the task (that is, they gave the same label to two same-category stimuli, and a different label for an other-category stimulus); 14 followed a ‘reverse boundary’ pattern of labelling; nine followed a ‘no
boundary’ pattern. All three groups evinced an other-category advantage across the boundary in the task. However, this grouping was based on a single test, of production only. Children might have understood the correct colour terms, but failed to provide them out of (e.g.) shyness; or they might not have known them, but guessed correctly. “I don’t know” was interpreted as a category in its own right, so the ‘reverse-boundary’ pattern would include, for instance, a child who correctly labelled the good example of green in a sequence but declined to label the poorer exemplars — blue-green and green-blue — straddling the boundary. This work was recently replicated by Goldstein, Davidoff & Roberson (2009, Experiment 1), but with a more elaborate test for colour-language competence. They found a CP effect across two boundaries, but restricted to the group of toddlers who did know the categories involved.

Franklin et al. (2005, Experiment 2) applied the same task to a group of Himba children, aged about 3, and found a CP effect across the blue / purple boundary. Innate-category and linguistic accounts both predict a boundary here. Himba lacks a specific equivalent for purple, but according to Roberson et al. (2004, Figure 1), the blue / purple stimulus triad includes at least one linguistic boundary because the ‘black’ term ZOOUZU generalizes to encompass purple hues, and even mauves. In a replication, Goldstein et al. (2009, Experiment 2) found the same blue / purple CP effect, but no evidence for an effect across the blue / green boundary (which Himba does not recognise). The Himba children in these studies were not rated for category knowledge, but Roberson et al. (2004) reported that Himba children of this age already possess a reasonable grasp of their colour lexicon.

Franklin et al. (2008b) applied the speeded-search procedure (where a discrepant stimulus appears on a uniform background in one of 12 ‘clock-face’ locations) to test the cerebral lateralisation of any category phenomenon in a group of toddlers aged 2 to 5. This age range varied widely in colour-language competence, according to a robust score, allowing Franklin et al. to compare

---

9 This is a more abstract version of the process performed in practice by the children asked by Bonnardel and Pitchford (2006) to group colour samples (from a Munsell array) around targets. In the field of computer vision, Lammens (1994) proposed a similar model for how colour categorization could work in practice.
the least and most competent toddlers. Both displayed CP, but confined to the left visual field in the former, and to the RVF (left hemisphere) in the latter — in effect, bridging the gap between pre-lingual infants and adults.

### 11. Summing Up

By one definition, categorical perception (CP) consists of the perceptual ‘sharpening’ of category membership: an exaggeration of the subjective dissimilarity between two stimuli (in comparison to the objective dissimilarity between them) when they straddle a boundary and fall in different categories. Here the emphasis is on the boundaries that demarcate categories. In an alternative account, similarities among the members of the category are exaggerated because less-good members of a category are perceived as being closer to the category prototype than is actually the case. That is, the prototypes (foci) produce CP by acting as ‘perceptual magnets’ (Iverson & Kuhl, 1995).

Subjects easily learn to recognize a novel category and to assign stimuli to it. The category may or may not be accompanied by CP effects at its boundaries with adjacent categories. The research surveyed here centres on the possibility that some categories of colour vision are innate — in the specific sense that their boundaries are marked out by CP effects prior to any training, so that they are found, for instance, among young children who have not yet learned the lexicon of colour terms. If this proved to be true, it would lead to a host of research questions. The alternative possibility, that colour categories are all the product of experience, would be equally productive.
Many of the studies under review involve the visual search methodology (Section 6). When searching for targets amongst a background of distractors, detection time increases with the target / distractor similarity. In particular, ‘dissimilarity’ in this context includes a contribution from cross-boundary CP effects (that is, using the terms introduced in Section 2, detection time is a function of the perceptual target / distractor distance in colour space, rather than the sensory distance). However, altered dissimilarities are not enough to explain why search is ‘efficient’ and ‘inefficient’ for other-category and same-category targets respectively (Daoutis et al., 2006a). That is, search time for an other-category target is relatively independent of the number of distractors, as if the search can inspect multiple distractors at once. Nor do altered dissimilarities account for the ‘category advantage’ observed when there are several distractor hues, from which the target is not linearly separable. Instead, category membership seems to function as a feature that can guide visual search. In this specific context, it may be that the exaggerated cross-category dissimilarities of classical CP are a secondary effect.

Some authors have inferred from the rapidity of visual search that it bypasses verbal processing, and invokes only pre-attentive mechanisms (e.g. Davies et al., n.d.). They go on to argue that the observed CP effects are a case of ‘true’ CP, a purely perceptual phenomenon, in contrast to the possibility that the differences in an observer’s cross-boundary and within-category responses stem from verbal functions (‘labelling strategies’). Daoutis et al. (2006a) suggest that the term “CP may be a misnomer [, if] it results from labelling rather than from perceptual processes.”

The same contrast is applied to results from other research techniques. In the case of the 2-AFC delayed-discrimination task, Pilling et al. (2003, p. 549) conclude “… that categorical perception in delayed discrimination tasks may be due to comparison of verbal labels, rather than warping of perceptual space, [although] this does not mean that true CP does not occur under other circumstances”. Essentially, language is treated as the final stage of our cognition, a ‘user interface’ that translates our perceptions, memories and thoughts from ‘mentalese’ into a medium understood by our listeners.

This review has addressed both steps in this argument: (1) that the rapidity of visual search makes the participation of verbal processing impossible; and (2) that behavioural responses are not genuinely ‘perceptual’ if they have been
influenced by verbal responses.

(1) appears to rest on a confusion of the final stage of verbal processing (that is, articulation of a colour term) with the entire complex process. In Section 4 I argued that much of the verbal response to a stimulus is pre-attentive, involuntary and parallel. To use the connectionist idiom, catching sight of a colour stimulates the appropriate node in the observer’s semantic network. A cascade of stimulation propagates upwards and can trigger an eye movement towards a target (or elicit the corresponding key-stroke response) well before it enters the domain of verbal articulation. Judging from the evidence from adult speeded search, several of these cascades of stimulation can proceed concurrently, one for each of multiple stimuli in the visual field, until the linear limitation of consciousness acts as a bottleneck and singles out the most salient one. In this view, labelling is a bottom-up, pre-attentive process rather than a top-down factor.

(2) posits a concept of pure Perception, enshrined above any contamination from cultural conditioning — an arcadian ideal belonging to the romantic world-view of Rousseau. This underestimates the degree of integration of perception and symbolic (verbal) cognition. Semantic processing and subjective experience overlap. I am arguing that verbal comparisons and labelling are an integral component of perception, and when they seem to distort dissimilarities across a language’s category boundaries, that this is an authentic change of perceptual space rather than an illusion.

A weak version of the Whorfian hypothesis generates a number of testable hypotheses. For instance, using semantic primes to predispose an observer to think about words like ‘red’ or ‘blue’ should also affect the strength and locations of the corresponding boundaries.

One reason to think that the role of categories in visual search is mediated by language (i.e. that it deserves to be called a Whorfian effect) is its greater strength in the right visual field (Section 6). A second reason is that CP effects disappear when language functions are otherwise engaged. Several studies have distracted the subjects with secondary tasks while they made colour-similarity judgments (Section 5) or pursued a visual search (Gilbert et al., 2006). A parallel tradition of cross-cultural research, concentrating on similarity judgments (Section 3), began with Kay and Kempton (1984). Apart from Laws et al. (1995), the literature consistently shows that when subjects
display CP effects, these correspond to linguistic boundaries specific to the subjects’ own language.

This does not settle the question of the origin of these categories. The lexical differences between languages — differences on how to carve colour space along category boundaries — are superimposed on a general trend towards cross-cultural agreement (e.g. Regier et al., 2005). The fact that lexicons have diverged relatively little implies that they are presumably optimal in some way: best-suited to the environment, perhaps, or to human physiology (Bimler, 2005).

Some researchers have examined similarity judgments or visual-search performance from children, either comparing groups of children according to their grasp of colour language or following the same children longitudinally as they mature, in case there were associated differences in their pattern of responses (Section 10). Such comparisons can be combined with the cross-cultural approach. The results have been indecisive.

A final research tradition has focused on infants (Section 8). Infant novelty preferences may depend purely on the categories of the colour stimuli, with no contribution from their separation in colour space (Franklin & Davies, 2004), or purely on the separation with no cross-boundary contribution (Gerhardstein et al., 1999). Using an analogue of the visual-search paradigm, Franklin et al. (2008a) found evidence for a category boundary between green and blue. Unexpectedly, this is lateralized in the left visual field, suggesting that the categories in question are not directly coupled to the way that language partitions the colour domain in later life.

This heterogeneous catalogue of findings can be reconciled within a single framework as follows: Adult perception of colour includes cognitive systems that develop during childhood, have a left-hemisphere neural substrate, and are strongly responsive to cultural transmission — in particular, to the acquisition of a colour vocabulary. These systems are built upon a foundation of hard-wired right-hemisphere mechanisms that are active in infancy, detecting a basic set of features within colour stimuli (‘greenness’, ‘blueness’, etc.). One could draw an analogy with the way that the adult mode of memory is preceded in many individuals by eidetic or ‘photographic’ memory (e.g. Haber, 1979) — a more concrete but less efficient way of encoding experience, less specific to one’s cultural environment. The thresholds between these putative features
may be passed on to inter-category boundaries (as in the case of the BLUE / GREEN boundary in English, but this inheritance is not obligatory.

Regier, Kay and Cook (2005) have pointed out that across languages, the foci of colour categories vary less than the boundaries between them. They present a model of colour labelling in which a given hue sample is subjected to a ‘gravitational pull’ from each focus, which falls off according to the colour-space distance between hue and focus, and depends on the “strength” of that category, where these ‘strength’ parameters are variable. The sample is assigned to the category that exerts the greatest attraction, so increasing a given strength parameter extends that category’s share of colour space by annexing marginal hues, pushing back the boundaries (‘watersheds’) with neighbouring categories.

References


Benjamins.


