Available online at www.sciencedirect.com
ScienceDirect
Brain and Language 103 (2007) 251–263
Brain and Language
www.elsevier.com/locate/b&l

Does maintenance of colour categories rely on language? Evidence to the contrary from a case of semantic dementia

Catherine Haslam *, A.J. Wills, S. Alexander Haslam, Janice Kay, Rachel Baron, Fiona McNab

School of Psychology, University of Exeter, Exeter EX4 4QG, UK

Accepted 13 August 2007

Abstract

Recent neuropsychological evidence, supporting a strong version of Whorfian principles of linguistic relativity, has reinvigorated debate about the role of language in colour categorisation. This paper questions the methodology used in this research and uses a novel approach to examine the unique contribution of language to categorisation behaviour. Results of three investigations are reported. The first required development of objective measures of category coherence and consistency to clarify questions about healthy control performance on the freesorting colour categorisation task used in previous studies. Between-participant consistency was found to be only moderate and the number of colour categories generated was found to vary markedly between individuals. The second study involved longitudinal neuropsychological examination of a patient whose colour categorisation strategy was monitored in the context of a progressive decline in language due to semantic dementia. Performance on measures of category coherence and consistency was found to be relatively stable over time despite a profound decline in the patient's colour language. In a final investigation we demonstrated that, for both the patient and controls, between- and within-participant consistency were higher than expected by (a) random sorting and (b) sorting perceptually similar chips together. These findings indicate that the maintenance of colour categorisation need not depend on language.

Keywords: Colour categorisation; Language; Semantic dementia

1. Introduction

There has been considerable debate over the role that language plays in guiding thought and behaviour, particularly in the context of colour categorisation. Indeed, this has been identified as one of the defining debates in psychology (e.g., Hock, 1992). In particular, the classic cross-cultural work of Rosch and colleagues challenged the hitherto influential view that language systems constrain and shape thinking (Whorf, 1956)—arguing instead that categories exist naturally and need not arise from language (e.g., Rosch, 1973; Rosch Heider & Olivier, 1972). Recent contradictory evidence from neuropsychological (Davidoff & Roberson, 2004; Roberson, Davidoff, & Braisby, 1999) as well as adult (Roberson, Davidoff, Davies, & Shapiro, 2005; Roberson, Davies, Corbett, & Vandervyver, 2005; Roberson, Davies, & Davidoff, 2000) and developmental (Roberson, Davidoff, Davies, & Shapiro, 2004) cross-cultural studies has led researchers to reconsider and indeed, advocate anew, original principles of linguistic dependence. In this paper we question the neuropsychological evidence and offer an alternative methodology, involving longitudinal examination of a patient who, as a result of neurological disease, suffered progressive loss of language. Findings drawn from study of this condition and that of healthy controls challenges previous neuropsychological evidence suggesting that language is essential in maintaining colour categorisation. This neuropsychological critique adds to, and complements, recent critiques of

* Corresponding author. Fax: +44 (0) 1392 264623.
E-mail address: c.haslam@exeter.ac.uk (C. Haslam).

0093-934X/$ - see front matter © 2007 Published by Elsevier Inc.
doi:10.1016/j.bandl.2007.08.007

the cross-cultural (e.g., Kay & Regier, 2007) and developmental (e.g., Bornstein, Kessen, & Weiskopf, 1976; Franklin, Clifford, Williamson, & Davies, 2005) work.

1.1. Natural categories

The original challenge to the view that language constrains colour categorisation comes from the work of Rosch and colleagues. Rosch Heider and Olivier (1972) investigated colour knowledge in two cultures that differed markedly in their use of colour terms: the Dani who relied largely on two colour terms and American English speakers who used a range of basic and non-basic colour names. These researchers argued that if the principle of linguistic relativity were correct, the structure (or organisation) of colour in memory should resemble the structure of colour names in each language. Yet they found a universal pattern of colour confusions in a memory task, regardless of cultural differences in the naming of colours. A subsequent investigation (Rosch, 1973) found that the Dani learned focal colours faster than non-focal colours, where focal was defined as the most representative examples of Berlin and Kay's (1969) basic chromatic colour terms (e.g., green, yellow), and non-focal colours were internominal (e.g., greenish-yellow). These results suggest that focal colours were psychologically real for the Dani despite their lack of terms to describe them. Together, these data were used to support the argument that perception of colour is universal and that categorisation is dependent on the degree to which items matched a natural prototype, not a linguistic definition.

1.2. A Whorfian revival

Methodological criticism of the above work (see Davidoff, 2001; Lucy & Shweder, 1979) encouraged Roberson, Davidoff and their colleagues to replicate it, this time using Berinmo and English speakers. However, they found that performance on tests of categorical perception and category learning was driven primarily by knowledge of colour terms specific to each culture (Roberson et al., 2000). These findings contradicted Rosch's universality principle and led the researchers to conclude that "...colour categories are formed from boundary demarcation based predominantly on language" (Roberson et al., 2000, p. 394). Subsequent work continued to support this conclusion, with Roberson et al. (2004) showing that as children from different cultures acquire knowledge of colour terms, their recognition memory errors become more consistent with the labels of their culture.

The neuropsychological evidence is of particular relevance to this paper and is based on several reports of the same patient, LEW, who became severely anomic after suffering a left hemisphere stroke (Davidoff, 2001; Davidoff & Roberson, 2004; Roberson et al., 1999). While limited pathological data is provided in reports involving this patient (Davidoff & Roberson, 2004; Druks & Shallice, 2000; Roberson et al., 1999), the neuropsychological results suggest the stroke resulted in significant aphasia (affecting spoken and written language), left-sided inattention and at best a mild impairment in comprehension (but reports of these latter symptoms are inconsistent, see Druks & Shallice, 2000; Roberson et al., 1999). In an initial study, Roberson et al. (1999) found that LEW could neither name colours nor identify them through pointing or recognition. Hence, the deficit was not confined to name retrieval, but appeared to affect colour comprehension; the latter implied by LEW's poor performance in the pointing task which is of the type typically used to examine semantic knowledge (i.e., spoken-word item matching).

The status of LEW's internal representation of colour space was then tested using freesorting colour categorisation tasks based on an adapted version of the Berlin and Kay (1969) colour space. Two sets of colour chips were used in freesorting: a small set and a large set. In both cases LEW was simply asked to place items into as many groups as he thought appropriate. The small set comprised "four examples each from focal areas of red, yellow, pink, blue and green" (Roberson et al., 1999, p. 7) and, as a result, the within-group similarity was much greater than the between-group similarity. LEW sorted items in this set slowly, using pairwise similarity comparisons, but produced similar groupings to healthy controls. The same approach was used in attempting to freesort items in the large set, containing 58 items which varied widely in their hue and value. LEW failed on this task despite several attempts which included one occasion of attempting to copy the examiner's sort. Davidoff claims that LEW's failure in the freesort task "lead(s) directly to the conclusion that it depends on language" (Davidoff, 2004, p. 572), implying that colour vocabulary is necessary to maintain colour categorisation.

1.3. Critique of the Whorfian revival

The above findings from LEW appear to support the view that language determines categorisation, but closer inspection of the data presented in the 1999 paper raises some questions.

One of the more intriguing conclusions drawn in that paper is that whilst LEW has a serious deficit in his *explicit* access to colour categories, his implicit access to this information is largely preserved. Specifically, LEW appears to exhibit normal categorical perception for colour—in other words, he finds it easier to discriminate between two colours that cross a category boundary (e.g., blue and purple) than two colours from the same category (e.g., two blues). If one assumes that, in terms of perceptual similarity, cross-category stimulus pairs are comparable to within-category pairs then this difference in discriminability can be taken as evidence that LEW has preserved knowledge of the category boundary.

Such a conclusion would imply that the role of language in the maintenance of colour categories depends on the way

colour categories are accessed. However, the validity of this conclusion depends on the assumption that perceptual similarity has been appropriately controlled for in the tests of categorical perception. Although Roberson et al. (1999) do not formally assess the perceptual similarity of their stimuli, they have conducted such analyses in subsequent research, using Euclidean distance in the CIE L*u*v* colour space as a metric of perceptual similarity (e.g., Roberson, Davidoff, et al., 2005). One can apply the same analysis to Roberson et al.'s (1999) stimuli by converting the Munsell notations they provide for their stimuli into L*u*v* co-ordinates.[1] As Fig. 1 illustrates, the colours used by Roberson et al. (1999) were not equally spaced on this metric (only the blue–purple range is shown, but the blue–green range used is also unequally spaced).

The problem this causes is best illustrated by an example. In Experiment 1 of Roberson et al. (1999), one of the decisions facing the participant is to pick the odd-one-out from 7.5B, 2.5PB and 7.5PB. In terms of the colour categories, 7.5PB is the odd-one-out because it is typically called "purple", whilst the remaining chips are typically "blue". However, as Fig. 1 shows, 7.5PB is also the odd-one-out on the basis of perceptual similarity: 7.5PB is about 1.4 times as far from 2.5PB as 7.5B. In 13 of the 16 stimulus triads of Experiment 1, one cannot distinguish between categorical and perceptual decisions for this reason. Thus, a participant responding on the basis of perceptual similarity alone can score in excess of 80% "categorical" responses on this task.

Experiment 3 of Roberson et al. (1999) contains the same ambiguity. Experiment 2, however, was closely based on the work of Kay and Kempton (1984), in which perceptual similarity was formally assessed. Kay and Kempton demonstrated that people can be led to produce either category-based or perceptually-based similarity judgements by the manipulation of task demands. LEW was reported to be sensitive to the same manipulation, on the basis that he was within 1.5 standard deviations of control performance. However, the critical question is whether LEW's performance was reliably affected by the experimental manipulation. Based on data presented in Table 3 of Roberson et al. (1999, p. 21) it is not, $\chi^2 = 2.29$, $p > 0.05$. In summary, the evidence that LEW has preserved implicit colour category knowledge is inconclusive, either because perceptual similarity was not fully controlled (Experiment 1 and 3) or because LEW was not reliably affected by the experimental manipulation (Experiment 2).

Another problem with interpretation of LEW's data involves comparisons that were made with control participants. Roberson et al. (1999) argue that LEW's attempts to freesort the large set resulted in groupings which were markedly abnormal compared to controls. In support of this, three control participants are reported to have sorted the large set rapidly into eight groups whilst LEW, in contrast, produced 4 or 5 groups that were clearly far from random, but differed from those produced by controls. In the absence of any objective measure of the level of discrepancy between LEW and controls, and with such a small control sample, these apparent differences must be interpreted with some caution. Whilst Roberson and colleagues have subsequently employed larger control samples, these data are not directly relevant to the question of the abnormality of LEW's freesort. In one study the controls did not perform a freesort task (Roberson et al., 2000) and in a second study, where freesort was used, the colour chips differed in several ways from those used in the 1999 paper (Roberson, Davies, et al., 2005). In the latter study, the stimulus set encompassed a wider range of saturation and brightness than that used to test LEW, and it also included seven achromatic stimuli—black, white, and five shades of grey—that were not presented to LEW. Clearly, there is a need for more control data to assist in interpretation of patient performance on this large freesort task.

Davidoff has used the neuropsychological, adult cross-cultural and developmental evidence discussed above to support the Whorfian view that "perceptual categories are organised by the linguistic systems of our mind." (Davidoff, 2001, pp. 382). However, all three sources of evidence for this statement are less than definitive. The issues surrounding the current neuropsychological evidence have been discussed above. Interpretation of the developmental evidence is complicated by the finding that Western pre-linguistic infants show categorical perception for both primary (blue–green) and secondary (blue–purple) colour boundaries (Bornstein et al., 1976; Franklin et al., 2005), which does not seem to support a Whorfian view. Interpretation of the adult cross-cultural evidence is complicated by the fact that it is largely based on memory performance, and hence may result from participants using different verbal codes to represent the same underlying percepts. A number of authors (e.g., Davidoff, 2004; Goldstone, 1995) consider the Whorfian view to hold that language affects not merely the memory of percepts, but perception itself; something that is not demonstrated by these experiments.

1.4. The present studies

Patients like LEW, whose colour vocabulary is severely impaired, provide an interesting test of the relationship between linguistic systems and the organisation of perceptual categories. More specifically, they provide a way to address the question of whether the ability to use verbal codes is critical to the maintenance of perceptual categories such as colour. LEW's colour freesorting behaviour has been interpreted as demonstrating that maintenance depends on language (Davidoff, 2004). However, LEW is

[1] We did this by first converting the Munsell notation into CIE x,y,Y co-ordinates using the tables provided by Newhall, Nickerson, and Judd (1943); as recommended by the Munsell Colour Science Laboratory (2005) and then converting to CIE L*u*v* co-ordinates using the standard formulae, defining the white-point as CIE illiminant C (used by Newhall et al., 1943).

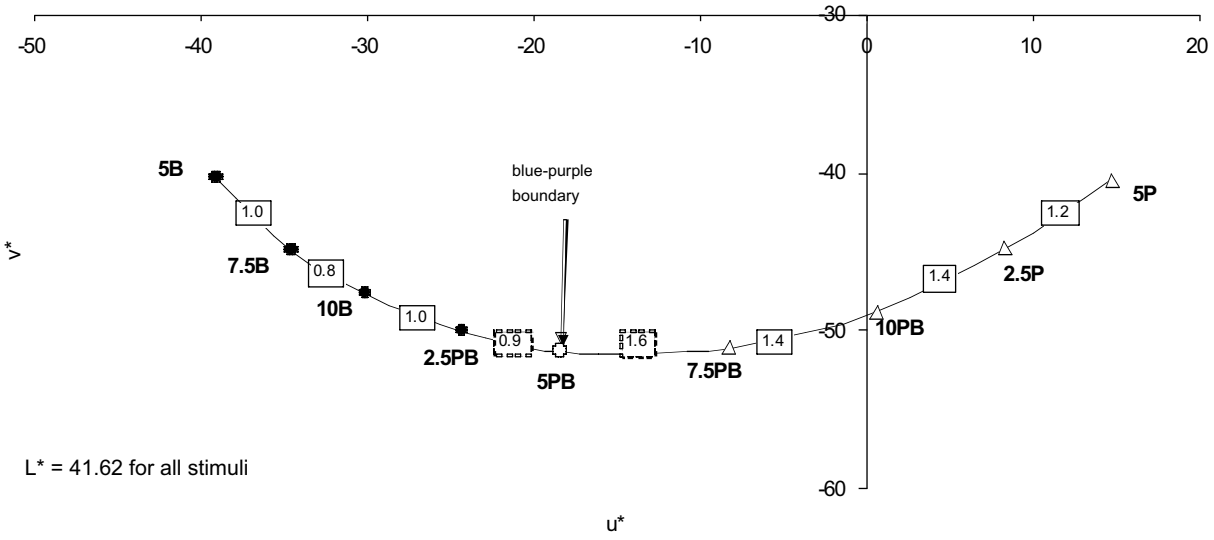

Fig. 1. The blue–purple range of stimuli used in Roberson et al. (1999) shown in CIE L*u*v* colour space. Filled circles denote stimuli typically named "blue". Open triangles denote stimuli typically named "purple". 5PB is on the boundary between blue and purple. Bold type denotes the Munsell hue. The numbers between stimuli indicate the Euclidean distance between those stimuli.

only one case and, in addition to the concerns we have already expressed about Roberson et al.'s (1999) study, we note that LEW lost knowledge of colour language *prior* to being tested. Arguably, longitudinal evidence from individuals who lose colour vocabulary *during* the period of observation would be more instructive; in a longitudinal study, it would be possible to observe the direct impact of a progressive deterioration in language.

Our critique of Roberson et al. (1999) focussed on two issues. The first concerned questions over what constitutes normal performance in freesort categorisation. We address this issue in our first experiment with the aid of two objective measures of categorisation performance taken from the cognitive and social psychological literatures. The second concerned the neuropsychological data from LEW and, more specifically, whether these data provide convincing support for the direct relationship between language and colour categorisation that Roberson, Davidoff and their colleagues assert. To explore this, we conducted a longitudinal study of a patient whose neurological condition—semantic dementia—results in a progressive deterioration in language. There are two reasons why semantic dementia provides a good model to test the relationship between language and colour cognition. First, it has been shown that colour naming is a relative strength for these patients and we capitalised on this—expecting that access to colour terms would deteriorate as the disease progressed (Robinson & Cipolotti, 2001).[2] Second, both colour naming and recognition was impaired in LEW and both deficits would be expected in a case of semantic dementia. Our goal was to determine whether the patient's categorisation strategy changed as a function of a decline in his use of language. There are three possible outcomes. If colour vocabulary is essential in maintaining categorisation, then a marked reduction in performance should be observed with increasingly more random sorts over time. Alternatively, if language is irrelevant to categorisation behaviour, performance should not change. A third possibility is that language is involved, but plays a non-deterministic role in categorisation, and in this event we would expect some reduction in performance but not one that results in random groupings. In a final section we evaluate the freesorts generated by our patient and controls against two models of freesorting to determine whether their groupings are better than would be expected by (a) chance or (b) grouping similar chips together.

2. Study 1

What constitutes normal performance in a freesort categorisation task? Previous findings (i.e., Roberson et al., 1999) suggest that healthy adults show little between-subject variability in their categorisation both in terms of (a) the number of groups produced and (b) the content of groups. Yet the relevant control data from Roberson et al. (1999) is very limited, and considerable between-subject variability has been shown in freesorts of other, albeit more abstract, stimulus types (Milton & Wills, 2004; Pothos & Chater, 2005; Wills & McLaren, 1998).

Clearly, more objective measurement of freesort performance is required to establish normal performance in the freesorting of colours. In identifying and developing objective measures we were guided by two principles commonly

[2] This is not an unreasonable assumption given that another reported area of relative preservation in semantic dementia, number knowledge, has been shown to deteriorate with disease progression (Cappelletti, Kopelman, Morton, & Butterworth, 2005; Jefferies, Bateman, & Lambon Ralph, 2005).

employed in the categorisation literature—category coherence and classification consistency.

Our first measure, the *meta-contrast co-efficient (MCC)*, quantifies category coherence, also variously referred to as *entitativity* (Campbell, 1958), *fit* (Bruner, 1957), or *comparative fit* (Turner, 1985). The MCC is an adaptation of the *meta-contrast ratio* (MCR) widely used in the social categorisation literature (e.g., Haslam & Turner, 1992; see McGarty, 1999, pp. 112–113, for general principles of calculation). It basically quantifies the extent to which within-category differences are smaller than between-category differences.

Establishing the MCR for any one set of colour chips sorted into the same group involves computing the average lateral[3] distance between each chip and all other chips *not* in the group and dividing this by the average lateral distance between each chip and all others that *are* in the group. The resulting MCR for a participant is the mean of the MCRs for the participant's groups. To standardise scores across contexts (which is relevant in freesorting where the number of groups generated can differ), the obtained meta-contrast ratio (MCR) is divided by the maximum MCR that would be obtained if chips were *optimally* sorted into the same number of groups (i.e., in such a way as to maximise inter-set difference and to minimise intra-set difference). This creates a meta-contrast co-efficient (MCC) with the desirable property that maximum category coherence is always represented by an MCC of 1, regardless of the number of groups used.

Our second measure, *Cramér's phi* (Φ_C, Cramér, 1946), indexes classification consistency; in other words, the extent to which two classifications are consistent with each other. In tasks where participants allocate one of an experimenter-defined set of labels to each stimulus, the consistency of two participants' classifications can be indexed by simply calculating the percentage of agreement. In a freesort task, however, neither the participants nor the experimenter explicitly provide category labels, and participants often differ substantially in the number of categories they produce. As Wills and McLaren (1998) have previously argued, a statistic that indexes the level of association (i.e., co-prediction) between two categorical variables is a more appropriate measure of consistency in a freesorting task. Cramer's phi (see Wills & McLaren, 1998, pp. 238–241, for general principles of calculation) provides such a measure.[4] Specifically, it indexes the *consistency* between two different categorisations of the same set of stimuli—in our case, two attempts at freesorting the large set of Munsell colour chips. Φ_C has the desirable characteristic of varying from 0 (i.e., no association) to 1 (maximum association), however many groups each participant produces

2.1. Methods

2.1.1. Participant

Twenty-seven people took part in this study, nine in each of three age groups: younger adults (mean age = 27.7 years, $SD = 4.1$), adults (mean age = 48.2 years, $SD = 7.4$) and older adults (mean age = 66.2 years, $SD = 4.9$). The ratio of males to females was 4:5, 4:5 and 5:4 in these groups respectively. All had English as their first language and were screened for colour blindness using the Ishihara Tests for Colour Blindness (Ishihara, 1992).

2.1.2. Materials and procedure

The colour categorisation task and general procedure were based on those used by Roberson et al. (1999) in testing LEW. Materials comprised a set of ten colour tiles (the best examples of the eight basic chromatic colour categories (red, blue, violet, green, yellow, pink, brown and orange), plus black and white) and two sets of Munsell colour chips: one small and the other large. The small set comprised four good examples of each of red, green, blue and yellow; grouped closely in hue, value and chroma. The boundary between these colours was well defined. The large set comprised 60 chips[5] encompassing a wide variation in hue and lightness.

Participants were asked to perform two tasks. The first involved naming and recognition of the 10 colour tiles. For naming, the tiles were presented individually in a random order and participants were simply asked to name them. For recognition, the 10 tiles were presented simultaneously on a table in front of participants and they were asked to point to a named colour (e.g., "point to the black tile"). The second task involved freesorting the small and large sets of colour chips, respectively. In each case the items were spread randomly on a table in front of participants who were asked to place them into as many groups as they felt appropriate.

2.2. Results and discussion

All participants named and recognised the 10 colour tiles without error. Similarly, all sorted rapidly the small set of colour chips into the same four groups. The data of particular interest is that from the large freesorting task and this will be reported in three sections.

[3] The lateral difference is based on the dimension of hue in the stimulus set (i.e., 20) and is the one with the greatest potential variability in the sorting task. The shortest lateral distance is calculated within a stimulus domain that is treated as cylindrical.

[4] φ_C was applied, rather than the inter-rater reliability measure kappa (Cohen, 1960), because the latter is a chance-corrected percentage of agreement measure, and is limited by an assumption that the two observers use the same number of groups.

[5] Roberson et al. (1999) report using 58 colour chips to which we added an additional two to complete the 20×3 (i.e., hue by brightness) graphical representation of colour space.

2.2.1. Number of categories

There was a wide range in the number of categories generated: from 3 to 16 for younger adults (median = 7), 4 to 21 for adults (median = 5) and 3 to 15 for older adults (median = 5). Given this variability, it was clearly unwise to represent the group data in a single figure. Accordingly, Fig. 2 shows the freesorts of three individuals who generated the median number of categories in freesorting: one from each age group. The figure is represented as a colour wheel, to illustrate both the colours used together with the variation in hue (circumference) and value (radius) of the stimulus set. The black lines show the items that were grouped together. The mean time taken to sort the chips was 307.3 ($SD = 145.2$), 296.5 ($SD = 184.5$) and 272.2 ($SD = 72.9$) seconds for younger adults, adults and older adults, respectively, and this did not differ between groups, $F(2,17) = .131$, $p = .88$.

2.2.2. Category coherence

The meta-contrast co-efficient (MCC) values ranged from 0.35 to 0.88 (mean = 0.67) for younger adults, 0.5 to 0.9 (mean = 0.76) for adults, and 0.34 to 0.92 (mean = 0.67) for older adults. There was no difference in the MCC values between age groups, $F(2,24) = .83$, $p = .45$, but the range in scores within age groups suggests some between-subject variability in the coherence of categories generated. It is helpful to consider the MCC values for the three controls whose groupings are produced in Fig. 2. Of these participants, the median older adult and the median adult produced the most coherent groupings along the dimension of hue. This is reflected in their MCC values of 0.88 and 0.86, respectively—appreciably better than the median younger adult's MCC value of 0.66.

2.2.3. Classification consistency

Cramer's phi (Φ_C) scores were calculated for each pair of participants within each age group, and also for each pair of participants irrespective of age. The mean and standard deviation of these scores are presented in Table 1 which shows no difference between age groups in Φ_C scores. Clearly the level of agreement in the content of categories generated was moderate (although comparable to freesorts of other stimulus types; see Wills & McLaren, 1998), and appears to be below that indicated in previous research. Importantly, this raises questions about the interpretation of LEW's performance. If controls do not perform the large freesort task consistently, then perhaps LEW's performance is not as impaired as previously thought. We return to this issue in Section 5. Furthermore, as our controls did not produce the standard eight categories reported in previous research, then one also might question the degree to which knowledge of colour terms influences freesort behaviour.

Table 1
Mean pairwise comparison (i.e., Φ_C) scores for controls

Participants	Mean	*SD*
Younger adults	0.78	0.09
Adults	0.79	0.05
Older adults	0.77	0.12
All controls	0.77	0.09

3. Study 2

Having established the performance of healthy adults on the freesort task using our two measures, we proceeded to investigate further the role of language in colour categorisation through longitudinal examination of freesort behaviour in a patient whose knowledge of colour language was predicted to decline through disease.

3.1. Methods

3.1.1. Participants

Six people took part in this study: JB, a patient diagnosed with semantic dementia and a group of five healthy adult controls (mean age = 47.2 years; $SD = 6.8$ years). Data from the latter group of controls were used to provide

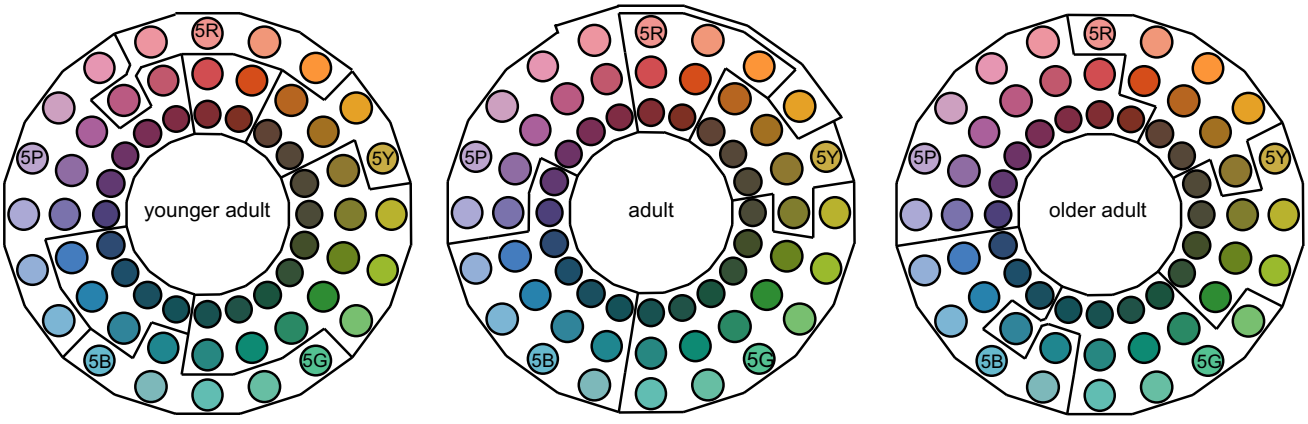

Fig. 2. Freesorting performance of the median younger adult, adult and older adult represented as a function of hue (5R to 10RP represented on the circumference) and value (3, 5 & 7 represented on the radius).

insight into the consistency between the same respondent's sort on two occasions.

JB was initially referred for medical assessment in October 2000, at the age of 69, following a 4-year history of a progressive decline in receptive and expressive language. Structural MRI showed atrophic changes, predominantly in the temporal lobes, with no other focal abnormality nor evidence of cerebrovascular disease. On examination JB had difficulty following conversation requiring him on occasion to question the meaning of words (e.g., "What are fireworks?", "What are crisps?"). Limited testing was conducted given the severity of his language problems and showed evidence of impairment in fluency, semantic association (e.g., Pyramids and Palm Trees Test, 3 pictures = 45/52) and naming (e.g., Graded Naming Test = 0/30). In the case of naming, neither semantic nor phonic cueing aided retrieval. This presentation together with the medical evidence supported the diagnosis of semantic dementia.

JB was referred to our research program in October 2002 at which time results of testing showed his condition had deteriorated further (see Table 2). Spontaneous speech was empty in content and characterized by profound word-finding difficulties. Results of standard tests, including word-picture matching and confrontation naming, showed profound impairment in naming and comprehension. As often reported in semantic dementia (e.g., Hodges, Patterson, Oxbury, & Funnell, 1992; Snowden, Neary, & Mann, 1996), performance was within normal limits on tests of more general cognitive function including spatial memory span, planning, problem solving and perception.

Table 2
Results of standard neuropsychological testing conducted in October 2002

General function	
WASI: Block Design	50
Matrix reasoning	48
Spatial span	6 Forwards, 0 backwards
Rey complex figure: copy	34/36
RPM: short form[a]	9/12
BORB: Object decision (easy)	0/10 (discontinued)
VOSP Object Perception	
Screening Test	20/20
Incomplete letters	20/20
VOSP Space Perception	
Dot counting	10/10
Position discrimination	20/20
Number location	9/10
Cube analysis	8/10
Language Function	
PALPA 48 (written word-picture match)	12/40
PALPA 53 (naming)	6/40

Notes: WASI, Wechsler Abbreviated Scale of Intelligence; RPM, Raven's Progressive Matrices; BORB, Birmingham Object Recognition Battery; VOSP, Visual Object and Space Perception Battery; PALPA, Psycholinguistic Assessment of Language Processing in Aphasia.
Raw scores are presented for all tests except the WASI where T-scores are provided.

[a] Normative information on the RPM: short form is available in Lyons, Hanley, and Kay (2002).

Table 3
Accuracy in colour naming and recognition with category coherence (MCC) for JB

Assessment date	Naming (%)	Recognition (%)	MCC
March 2003	70	70	0.85
June 2003	80	70	0.78
June 2004	0	20	0.88

3.1.2. Materials and procedure

The procedure used in Study 1 was repeated here. However, the naming, recognition and freesorting categorisation tasks were presented to JB on three occasions over a period of 15 months (the time between Sessions 1 and 2 was 3 months and that between Sessions 2 and 3 was 12 months) and to the healthy adults on two occasions where the time between sessions was 1 month. Further evidence of JB's colour knowledge was obtained from two additional tests administered in the first and third testing sessions. The first was a version of the Weigl colour form sorting task (Weigl, 1941) involving presentation of 15 items of different colour (red, green, yellow, blue and white) and shape (triangle, circle, square). JB was required to sort these items into groups that belonged together, asked to identify his sorting principle, and then to sort the items in a different way. The second task involved object-colour matching and comprised 15 unshaded line drawings of objects (e.g., strawberry, banana, tomato). JB was presented with the line drawings, individually, along with six felt tipped pens of different colours and asked to fill the item with its appropriate colour.

3.2. Results and discussion

JB's performance in colour naming and recognition deteriorated over time as predicted given his neurological condition. JB's naming and recognition, although lower than controls in the first two sessions, were reasonably high indicating that he had knowledge of many colour terms on initial testing (see Table 3). In Session 1 errors in naming and recognition were consistent (involving pink, violet and yellow). The same errors in recognition were made in Session 2, although on this occasion he was able to name the colour pink.

Further evidence of intact colour knowledge in Session 1 was demonstrated in JB's performance on colour-form sorting and object-colour matching. In the former task JB immediately proceeded to sort items correctly according to colour and identified this as his sorting principle. He failed to sort items according to shape. In object-colour matching, JB was only asked to colour those 12 of the 15 items that he recognised. Inclusion of the three non-recognised items would create a confound as errors could be attributed to impairment in either object or colour knowledge. He performed this task perfectly. By Session 3, naming was at floor and recognition was only slightly better indicating limited, if any, knowledge of colour concepts.

At this time, JB's comprehension of objects was so poor that the object-matching task could not be attempted. Nevertheless, he categorised items according to colour in the colour form sorting task, but could no longer state his organising principle.

In light of this marked deterioration in language, JB's categorisation of the Munsell colour patches over time provides a critical test of the role of language in the maintenance of knowledge of colour categories. In each session, JB sorted the small set of colour patches into four groups (i.e., green, yellow, red and blue) that were identical to those produced by controls in both this and our first study. JB's freesorting of the large set of colour chips on the three occasions of testing is presented in Fig. 3. In all sessions he sorted the large set without hesitation, with no attempt to perform laborious pairwise comparisons nor to place chips alongside each other, as reported in the case of LEW (Roberson et al., 1999). His completion times of 353, 410 and 386 s for Sessions 1 to 3, respectively, were slower than our older adult controls, falling between 1 to 2 standard deviations of their mean time (mean = 272.2 s, $SD = 72.9$). However, this is perhaps unsurprising given the severity of JB's condition from the first session. More important is the fact that JB's completion time did not decline dramatically in Session 3 when his knowledge of colour terms was at floor. Fig. 3 shows some variation in JB's sorting strategy, with six groups produced in Session 1 and 5 in the remaining sessions. However, this number is well within the range produced by our healthy younger adults, adults and older adults in Study 1. Hence the number of groupings is not abnormal relative to our controls.

JB's MCC scores are presented in Table 3 and indicate relative stability in his categorisation strategy (with scores ranging from 0.78 to 0.88) despite the marked deterioration in his colour language. Furthermore, his scores were within the normal range of our older adults in Study 1, falling within one standard deviation of the control mean value of 0.67. The top half of Table 4 shows the mean consistency (i.e., Φ_C) between JB and controls. The mean values reported were calculated from all possible pairwise comparisons between JB and older adult controls (which is appropriate given JB's age) and JB and all controls combined. As indicated in this table, Φ_C values ranged from 0.72 to 0.75 and showed little variation across sessions. Thus, despite a profound deterioration in JB's knowledge of colour terms there was no evidence of a decline in the content of his groupings when compared with that of controls. Furthermore, these consistency values compare favourably with the mean pairwise comparison values calculated for controls in Study 1. For example, JB's mean agreement with older adult controls in Session 3 is within one standard deviation of the mean agreement between older adult controls. Clearly, the content of JB's categories was not abnormal relative to controls.

Table 4
Mean consistency (Φ_C) between controls and JB as a function of session and between controls and the coherent category model

	Older adult controls	All controls
JB		
Session 1	0.74 (0.07)	0.75 (0.07)
Session 2	0.72 (0.08)	0.73 (0.08)
Session 3	0.75 (0.09)	0.73 (0.07)
Coherent category model		
Session1 (6 categories)	0.64 (0.04)	0.63 (0.06)
Sessions 2 and 3 (5 categories)	0.62 (0.05)	0.62 (0.06)

We also looked at the association between the three separate occasions on which JB was tested using Φ_C to determine whether there was any evidence of change in his particular categorisation strategy during the course of his illness (see the top half of Table 5). Consistency values were reasonably stable despite the very marked decline in his knowledge of colour terms. One might ask how these levels of self-consistency compare with healthy controls. Whilst between-subject consistency is now known (see Table 1), it is possible that controls agree with themselves more than they agree with each other.

To address this question, a second group of healthy adults were asked to complete the task twice and self-consistency values were calculated. All participants named and

Fig. 3. JB's categorisation of Munsell colour chips in sessions 1–3 represented as a function of hue (5R to 10RP represented on the circumference) and value (3, 5 & 7 represented on the radius).

Table 5
Self-consistency (Φ_C) scores for JB, controls and expected models

Participants/models	Cramer's phi
JB	
Session 1 vs. Session 2	0.84
Session 1 vs. Session 3	0.69
Session 2 vs. Session 3	0.67
Controls	
Session 1 vs. Session 2	0.77 (0.07)
Expected models	
Coherent (controls)	
Coherent (JB: S1 vs. S2)	0.49 (0.06)
Coherent (JB: S1 vs. S3)	0.55 (0.09)
Coherent (JB: S2 vs. S3)	0.55 (0.09)
	0.51 (0.10)

recognised colours perfectly. Self-consistency values (i.e., Φ_C), ranged from 0.66 to 0.82 (mean = 0.77, $SD = 0.07$), indicating controls agree with themselves to about the same degree that they agree with each other (see Table 1). Importantly, JB's self-consistency scores were within the normal range even under conditions where a complete loss of the language of colour was observed in the intervening time between sorts (i.e., between Sessions 2 and 3).

4. Study 3

In the previous sections, we evaluated the extent to which (a) JB agrees with controls, (b) JB agrees with himself and (c) controls agree with each other, in freesort. What these analyses do not demonstrate is whether these levels of agreement are greater than would be expected by chance responding or indeed greater than would be expected by grouping similar chips together. To evaluate these two questions we compared the self-consistency and agreement scores reported in the previous sections to the expected values from two models of freesorting behaviour: random categories and arbitrary coherent categories (grouping by similarity).

4.1. Models

4.1.1. Random categories model

Under a hypothesis that two participants sort randomly, the expected value for φ_C is given by:

$$\phi_{\text{random}} = \sqrt{\frac{2(r-1)(c-1)-1}{2N(k-1)}}$$

where r and c are the number of groups used by the two participants (or the same participant on two occasions), k is the smaller of r and c and N is the number of stimuli (i.e., 60 in this case). The derivation is provided in Wills and McLaren (1998, pp. 239–240).

4.1.2. Arbitrary coherent categories model

This model considers a category to be coherent if each member of the category is adjacent to at least one other member. Adjacency is defined here in terms of the stimulus dimensions hue and brightness. Two stimuli are adjacent if they are identical on one dimension and maximally similar (within the presented set) on the other. To produce coherent categories reliably a participant must be sensitive to hue and lightness, but need not know the location of the category boundaries or how many members each category should contain. Accordingly, the classifications generated are coherent but arbitrary with respect to category boundary and size. The arbitrary coherent categories model can be seen as one way of formalizing the grouping-by-similarity principle that Roberson et al. (2000) suggest is the primary constraint on colour naming across languages.

Under the arbitrary coherent categories model, the expected value for Φ_C for a pair of participants, one of whom uses r groups whilst the other uses c groups, is the mean value of Φ_C across all pairs of arbitrary coherent categorisations where one categorisation has r groups and the other has c groups. There are a very large number of distinct arbitrary coherent categorisations, which makes the calculation of an exact value impractical. We therefore estimated the expected value from a sample of 100 arbitrary coherent r-group classifications and 100 arbitrary coherent c-group classifications. The process by which samples of arbitrary coherent classifications were produced is described in the Appendix A.

4.2. Results and discussion

Perhaps unsurprisingly, the random categories model predicted lower levels of agreement than the arbitrary coherent categories model. For brevity, we therefore only report the levels of agreement predicted by the latter model. Results are reported separately for comparisons of this model with controls and comparisons of this model with JB.

4.2.1. Controls

The mean level of agreement expected between the Experiment 1 control participants under the arbitrary coherent categories model is 0.62 ($SD = 0.10$). As Table 1 illustrates, controls agree with each other to a greater extent than this model predicts. The same applies to control participants' self-consistency in Experiment 2 (see Table 5). Thus, our analyses support the view that there is more to control freesort performance than creating a set of arbitrary coherent groupings.

4.2.2. JB

The mean levels of agreement expected between JB and the Study 1 control participants under the arbitrary coherent categories model are shown in Table 4 (bottom panel). JB agrees with controls to a significantly greater extent than the model predicts, $t(26) = 11.25$, 12.87 and 11.15 for Sessions 1, 2 and 3, respectively.[6] Comparable results

[6] All p values in this section are $<.01$.

were found when only age-matched controls were considered, $t(8) = 5.82$, 5.83 and 5.34, respectively.

Table 5 (bottom panel) shows the mean levels of self-consistency predicted for JB under the arbitrary coherent categories model, together with the standard deviations. For each of the three self-consistency scores, JB's level of agreement is at least 1.6 standard deviations above the mean level of agreement of the arbitrary coherent categories model. The model's expected level of agreement was either equal to, or exceeded that of, JB on no more than 5% of occasions.

In summary, the agreement between JB and controls was consistently and significantly higher than would be expected if JB were generating arbitrary coherent classifications. There is also reasonable evidence to suggest that JB's self-consistency exceeds that which would be expected if he were generating arbitrary coherent classifications.

5. General discussion

This paper reports investigations of colour categorisation in healthy controls and in a neurological patient using two quantitative measures, the meta-contrast co-efficient (MCC) and Cramer's phi (Φ_C), to provide objective indicators of the coherence and consistency of the categories formed. In our first study, we found evidence of substantial variability in the performance of healthy controls, such that neither the number of categories nor the content of groupings generated were as consistent as previously reported. Our second study involved longitudinal examination of colour categorisation in a neurological patient, JB, suffering from a progressive deterioration in language. We found that the patient's freesort categorisation performance remained relatively stable on our measures of coherence and consistency, despite a profound and near-complete loss of colour language. Finally, we compared the self-consistency and between-participant consistency scores of our participants with those of two models of free sorting and found that the categories generated by JB were more self-consistent, and more consistent with the categories of healthy controls, than would be expected if he were creating coherent but otherwise arbitrary classifications.

The variability in free classification of colour that we observed in healthy adults, while consistent with that reported in some free classification tasks (e.g., Wills & McLaren, 1998), was inconsistent with the three control participants reported in Roberson et al. (1999). Importantly, this suggests there is nothing particularly special about the eight categories created by Roberson et al.'s controls. How might we account for this difference? One possibility may be that Roberson's sample of three controls is unrepresentative. Subsequent work by Roberson and colleagues (e.g., Roberson et al., 2000; Roberson, Davies, et al., 2005) was based on larger samples of healthy controls, but they were either not presented with a freesort task or were presented with a very different set of colour chips and so do not provide directly comparable evidence. It is also possible that for our controls, the initial freesort task with a small set of items increased the variability in sorting the second larger set. We did not pursue this issue for two reasons. First, the available evidence suggests that the variability of subsequent freesorts is reduced by an initial freesort (Milton & Wills, submitted). Second, and most importantly, JB also sorted the small set before the large set, and thus it was critical to retain this feature in the control conditions.

If maintenance of colour categorisation is dependent upon the ability to produce and comprehend colour vocabulary, then deterioration in the ability to use and recognise colour terms should severely impair categorisation ability. We investigated this hypothesis in Study 2 and found that JB's categorisation strategy, as assessed by our coherence and consistency measures, was both consistent with that of controls and remained stable despite the catastrophic decline in his colour language. Although some difference was observed in JB's self-consistency scores over time, this was neither commensurate with his profound language decline nor outside the range of control self-consistency scores. Given evidence of a mild decline in JB's self-consistency scores over time, it appears that the ability to use and recognise colour terms may play a limited role in categorisation. However, this is clearly far from deterministic as JB's performance remained well within the normal range despite his complete loss of colour vocabulary.

In the final part of our investigation, we compared the levels of self-consistency and between-subject consistency produced by JB, and by controls, to the levels predicted by two models of freesort behaviour: random categories and arbitrary coherent categories. There are several critical findings. First, controls agree with themselves and each other to a greater extent than predicted by either model. Second, the level of agreement between JB and controls is consistently and significantly higher than would be expected under the arbitrary coherent categories model. From this we can conclude that JB and controls were not randomly grouping chips nor were they simply grouping perceptually similar chips together. A related point is made by Kay & Regier (2007) in their re-evaluation of the Berinmo colour naming data. The authors argued that if colour naming were based solely on grouping by similarity, then rotations of the Berinmo colour boundaries (preserving the coherence and shape of actual categories) should align equally well with those of other languages. In fact, the actual boundaries produced by the Berinmo resulted in groupings that were more similar to other languages in the World Colour Survey than any of the rotations, suggesting some universal constraints in the placement of colour boundaries.

Whilst the levels of agreement predicted by our arbitrary coherent categories model are significantly lower than the levels of agreement seen in our participants, in absolute terms these two levels of agreement are not massively different. Thus, while JB and our controls performed better than would be expected if they were responding solely on

the basis of perceptual similarity, the relatively large Φ_C scores predicted by the arbitrary coherent categories model suggests that perceptual similarity may play an important role in freesort categorisation. Such a claim is consistent with many formal models of free classification (e.g., Pothos & Chater, 2005), but is at odds with Roberson et al.'s statement that "sorting colours into categories solely by observation is something that cannot be done" (Roberson et al., 1999, p. 27).

Roberson et al.'s (1999) study of their anomic patient LEW has previously been used to support the idea that freesort behaviour depends on language (Davidoff, 2004). In contrast, our study of patient JB seems to indicate that freesort behaviour is largely unaffected by a catastrophic loss of the ability to produce and comprehend colour vocabulary. As the neuropsychological evidence from LEW and JB appears to support opposing views, one might ask whether differences in pathology might account for the discrepancy. JB was diagnosed with semantic dementia and LEW suffered a left hemisphere stroke. In both conditions the left hemisphere is compromised and recent findings suggest that the influence of language in colour perception is mediated by this hemisphere (Gilbert, Regier, Kay, & Ivry, 2006). Accordingly, left-hemisphere damage may reduce the degree to which language influences performance in tests of colour perception. However, this does not explain why apparently only one of these patients could perform the freesort task. It is more likely that particular regions in the left hemisphere are implicated, though it is impossible to be more specific without further information on the nature and extent of LEW's lesion.

Closer examination of LEW's freesort may assist in resolving this disparity. Roberson et al. (1999) acknowledge that LEW's performance was not random, but claim his performance revealed *no* effect of category boundary. While this conclusion is supported by some evidence (e.g., unlike the Roberson et al. controls, LEW grouped together 5PB/7 through 5Y/7) it is inconsistent with other evidence (e.g., LEW placed the boundary between green and blue in the same place as Roberson et al. controls). In this context, we thought it would be informative to compare LEW's freesorting to our control participants. Using our Φ_C measure, we obtained a score of 0.67 which is only one standard deviation below the performance of our controls. Furthermore, LEW's performance was reliably higher than that of the arbitrary coherent categories model ($\Phi_{coherent} = 0.60$, $t(26) = 7.72$, for all controls; $\Phi_{coherent} = 0.59$, $t(8) = 3.45$, for older control participants). Thus, although LEW is clearly impaired in his knowledge of colour terms and his freesort performance is below average, it is probably within the normal range as assessed by Φ_C and is certainly better than would be predicted if he were simply creating coherent but arbitrary categories.

We have argued that poor performance on tests of colour naming and pointing indicates impairment of colour vocabulary. This seems uncontroversial, but a more complex issue is whether impairment of colour vocabulary implies an impairment of linguistic colour knowledge. One might argue, for example, that our naming and pointing tests measure linguistic competence rather linguistic knowledge per se. This argument, in turn, raises the question of how one distinguishes linguistic knowledge from non-linguistic knowledge without reference to linguistic competence. One answer is to define linguistic knowledge as the subset of knowledge whose acquisition was affected by language (as appears to be the definition in some of the cross-cultural and developmental studies previously discussed). However, our studies focused on maintenance rather than acquisition, asking whether the maintenance of colour categories relies on the ability to access colour terms. Our studies cannot, and were not intended to, address the issue of whether JB's acquisition of colour knowledge was affected by his culture or language.

In some ways, the maintenance question we pose is similar to the one addressed by studies of the effects of verbal interference on categorical perception (Gilbert et al., 2006; Roberson & Davidoff, 2000; Witthoft et al., 2003). Interestingly, that work suggests that disruption of access to colour terms disrupts colour perception, whilst the current study suggests that disruption of access to colour terms is largely irrelevant to colour perception. It may be that the effects of language on colour categorisation are substantially modulated by task specific demands. Whatever the merits of this position, JB's data demonstrates at a minimum that colour categorisation is not dependent on the linguistic knowledge that supports colour naming and pointing as has been previously argued by Roberson, Davidoff and their colleagues.

Our focus in this paper has been on the neuropsychological evidence used to support the case for the dependent role that language plays in colour cognition. Closer inspection of the existing neuropsychological evidence and our own longitudinal examination of another case raises questions about the nature of this relationship. This neuropsychological evidence does not speak directly to the adult and developmental cross-cultural work that has also been used to support a strong Whorfian position. Nevertheless, our neuropsychological evidence is consistent with developmental studies which have failed to find a dependent relationship between colour language and cognition (e.g., Franklin et al., 2005; Pitchford & Mullen, 2001). In summary, then, this work contributes to a body of research in multiple domains which raises doubts about at least some of the stronger versions of the Whorfian viewpoint.

In investigating the unique contribution of language to colour categorisation, we argue that our neuropsychological methodology provides a more comprehensive examination of the role of language in colour categorisation than previously attempted. In this respect, the approach we have taken also serves as a powerful demonstration of the capacity for single case investigations to shed light on core issues in mainstream psychological theory that prove hard to resolve using traditional experimental methodology (e.g., Marshall & Newcombe, 1984; Shallice, 1988). As with all

single case studies, replication is important; particularly when conflicting arguments are made on the basis of data from individual cases. Nevertheless, on the basis of the data presented here, it would appear that the role of language is not as fundamental as classical and recent formulations of the Whorfian hypothesis suggest. In short, language may assist colour categorisation, and it might be involved in acquiring this skill. However, once acquired, language is not essential for the maintenance of colour categorisation.

Acknowledgments

We would like to thank JB and his wife for taking part in this research and Debi Roberson for sharing her materials. This work was partially supported by EC Framework 6 project Grant 516542 (NEST) and BBSRC Grant 9/S17109 to A.J. Wills.

Appendix A. Sampling the set of coherent classifications

To sample the set of coherent classifications, one can sample the set of all possible classifications and retain only coherent samples. In a classification of sixty items, the probability of a sampled classification being coherent is very small. This means that, for the approach to be of practical use, one must be able to select coherent classifications automatically. The algorithm we used to sample the set of coherent classifications is described below.

The algorithm represents classifications as the presence and absence of boundary elements between adjacent stimuli. If, and only if, two adjacent stimuli belong to different groups does a boundary element exist between them. Boundary elements also exist above the brightest, and below the least bright, stimuli in the set.

The algorithm creates a classification by creating a collection of boundary elements. These boundary elements are generated randomly, under the constraint that no stimulus has four boundary elements (a stimulus with four boundary elements has no adjacent within-category items and hence the category it forms is not coherent). Under this constraint, a necessary and sufficient condition of coherence is that both ends of every boundary element must adjoin at least one other boundary element end. Non-coherent classifications can therefore be identified (and rejected) by the detection of an unconnected boundary element. A set of boundary elements that have no unconnected ends uniquely describes a particular coherent freesort of the stimulus set.

References

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

Bornstein, M. H., Kessen, H., & Weiskopf, S. (1976). Color vision and hue categorisation in young human infants. *Journal of Experimental Psychology: Human Perception and Performance, 2*, 115–129.

Bruner, J. S. (1957). On perceptual readiness. *Psychological Review, 64*, 123–152.

Campbell, D. T. (1958). Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioural Science, 3*, 14–25.

Cappelletti, M., Kopelman, M. D., Morton, J., & Butterworth, B. (2005). Dissociations in numerical abilities revealed by progressive cognitive decline in a patient with semantic dementia. *Cognitive Neuropsychology, 22*, 771–793.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Education and Psychological measurement, 10*, 37–46.

Cramér, H. (1946). *Mathematical models of statistics*. Princeton, NJ: Princeton University Press.

Davidoff, J. (2004). Coloured thinking. *The Psychologist, 17*, 570–572.

Davidoff, J. (2001). Language and perceptual categorisation. *Trends in Cognitive Science, 5*, 382–387.

Davidoff, J., & Roberson, D. (2004). Preserved thematic and impaired taxonomic categorisation: A case study. *Language and Cognitive Processes, 19*, 173–174.

Druks, J., & Shallice, T. (2000). Selective preservation of naming from description and the 'restricted preverbal message'. *Brain and Language, 72*, 100–128.

Franklin, A., Clifford, A., Williamson, E., & Davies, I. R. L. (2005). Color term knowledge does not affect categorical perception of color in toddlers. *Journal of Experimental Child Psychology, 90*, 114–141.

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

Goldstone, R. L. (1995). Effects of categorisation on color perception. *Psychological Science, 6*, 298–304.

Haslam, S. A., & Turner, J. C. (1992). Context-dependent variation in social stereotyping 2: The relationship between frame of reference, self-categorisation and accentuation. *European Journal of Social Psychology, 22*, 251–277.

Hock, R. R. (1992). *Forty studies that changed psychology: Explorations into the history of psychological research*. New Jersey: Prentice-Hall.

Hodges, J. R., Patterson, K., Oxbury, S., & Funnell, E. (1992). Semantic dementia: Progressive fluent aphasia with temporal lobe atrophy. *Brain, 115*, 1783–1806.

Ishihara, S. (1992). *Ishihara's Tests of colour blindness*. Tokyo: Kanehara Shuppan Co.

Jefferies, E., Bateman, D., & Lambon Ralph, M. A. (2005). The role of the temporal lobe semantic system in number knowledge: Evidence from late-stage semantic dementia. *Neuropsychologia, 43*, 887–905.

Kay, P., & Regier, T. (2007). Colour naming universals: The case of the Berinmo. *Cognition, 102*, 289–298.

Kay, P., & Kempton, W. (1984). What is the Sapir-Whorf hypothesis? *American Anthropological Association, 86*, 65–78.

Lucy, J. A., & Shweder, R. A. (1979). Whorf and his critics: Linguistic and non-linguistic influences on color memory. *American Anthropologist, 81*, 581–605.

Lyons, F., Hanley, J. R., & Kay, J. (2002). Anomia for common names and geographical names with preserved retrieval of names of people: A semantic memory disorder. *Cortex, 38*, 23–35.

McGarty, C. (1999). *The categorisation process in social psychology*. London: Sage.

Marshall, J. C., & Newcombe, F. (1984). Putative problems and pure progress in neuropsychological single case studies. *Journal of Clinical Neuropsychology, 6*, 65–70.

Milton, F. N., & Wills, A. J. (2004). The influence of stimulus properties on category construction. *Journal of Experimental Psychology. Learning, Memory, and Cognition., 30*(2), 407–415.

Milton, F. N., & Wills, A. J. (submitted for publication). Perseveration of sort strategy in free classification.

Munsell Colour Science Laboratory (2005). *RIT Munsell Colour Science Laboratory*, Retrieved May 8, 2006. Available from http://mcsl.rit.edu/.

Newhall, S. M., Nickerson, D., & Judd, D. B. (1943). Final report of the O.S.A subcommittee on the spacing of the Munsell colors. *Journal of the Optical Society of America, 33*, 385–418.

Pitchford, N. J., & Mullen, K. T. (2001). Conceptualization of perceptual attributes: A special case for color? *Journal of Experimental Child Psychology, 80*, 289–314.

Pothos, E. M., & Chater, N. (2005). Unsupervised categorisation and category learning. *Quarterly Journal of Experimental Psychology, 58A*(4), 733–752.

Roberson, D., & Davidoff, J. (2000). The categorical perception of colors and facial expressions: The effect of verbal interference. *Memory & Cognition, 28*, 977–986.

Roberson, D., Davidoff, J., & Braisby, N. (1999). Similarity and categorisation: Neuropsychological evidence for a dissociation in explicit categorisation tasks. *Cognition, 71*, 1–42.

Roberson, D., Davies, I., & Davidoff, J. (2000). Colour categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology. General, 129*, 369–398.

Roberson, D., Davidoff, J., Davies, I., & Shapiro, L. (2004). The development of colour categories in two languages: A longitudinal study. *Journal of Experimental Psychology. General, 133*, 554–571.

Roberson, D., Davidoff, J., Davies, I., & Shapiro, L. (2005). Colour categories: Evidence for the cultural relativity hypothesis. *Cognitive Psychology, 50*, 378–411.

Roberson, D., Davies, I. R. L., Corbett, G. G., & Vandervyver, M. (2005). *Free-sorting of colors across cultures: Are there universal grounds for grouping? Journal of Cognition and Culture, 5*(3–4), 87–124.

Robinson, G., & Cipolotti, L. (2001). The selective preservation of colour naming in semantic dementia. *Neurocase, 7*, 65–75.

Rosch Heider, E., & Olivier, D. C. (1972). The structure of the colour space in naming and memory for two languages. *Cognitive Psychology, 3*, 337–354.

Rosch, E. (1973). Natural categories. *Cognitive Psychology, 4*, 328–350.

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge: Cambridge University Press.

Snowden, J. S., Neary, D., & Mann, D. M. A. (1996). *Frontotemporal lobar degeneration: Frontotemporal dementia, progressive aphasia, semantic dementia*. London: Churchill Livingstone.

Turner, J. C. (1985). Social categorisation and the self-concept: A social cognitive theory of group behaviour. In E. J. Lawler (Ed.). *Advances in group processes* (Vol. 2, pp. 77–122). Greenwich, CT: JAI Press.

Weigl, E. (1941). On the psychology of so-called processes of abstraction. *Journal of Abnormal and Social Psychology, 36*, 3–33.

Whorf, B. L. (1956). *Language, thought and reality*. Cambridge, Mass: MIT Press.

Wills, A. J., & McLaren, I. P. L. (1998). Perceptual learning and free classification. *The Quarterly Journal of Experimental Psychology, 51B*, 235–270.

Witthoft, N., Winawer, J., Wu, L., Frank, M., Wade, A., & Boroditsky, L. (2003). Effects of language on colour discriminability. In R. Alterman & D. Kirsch (Eds.), *Proceedings of the 25th annual meeting of the cognitive science society* (pp. 1247–1252). Mahwah, NJ: Lawrence Erlbaum.