

Linguistic and behavioural measures for ranking basic colour  
terms<sup>1</sup>

Greville G. Corbett and Ian R. L. Davies

*University of Surrey*

Pre-print of paper which appeared in

***STUDIES IN LANGUAGE***

correspondence address:

Prof. Greville G Corbett

Surrey Morphology Group

Department of English

University of Surrey

Guildford, Surrey, GU2 7XH, GB.

---

<sup>1</sup> The support of the Economic and Social Research Council (ESRC) is gratefully acknowledged (award number R000 23 3978). We are also very grateful to Sean Hammond for useful discussions on statistical matters. A version of the paper was read at the Conference on Color Categorization held at Asilomar, October 1992; we wish to thank the National Science Foundation for funds enabling us to attend, and the convenor, Professor C. L. Hardin.

## **Abstract**

We report an attempt to find more objective measures for identifying basic colour terms. We investigate the types of measure available, both linguistic and behavioural, and the statistical techniques for establishing the closeness of fit with the predictions derived from Berlin and Kay. This leads to an investigation of the interrelation between the measures; having examined consistency across investigators and across languages we establish that certain measures give considerably better results than others. While the indicators point in the same direction, supporting Berlin and Kay to varying degrees, different measures serve different functions.

## 1. Introduction

The work of Berlin and Kay (1969) on basic colour terms continues to provoke interest and research.<sup>2</sup> Doubts have remained about their criteria for identifying basic colour terms (see, for instance, the strong criticism in Cromer 1991:138–40)<sup>3</sup> and so there have been various attempts to find more objective measures. It has been found that languages which have a full set of basic colour terms, may still preserve a colour hierarchy. That is to say there is not simply a division between basic and non-basic colour terms but rather the terms higher on the hierarchy can still be seen to be 'more basic' than those lower on the hierarchy. In establishing the 'survival' of the hierarchy we have used various testing techniques and it is on these that we focus in this paper. We will contrast the different measures in terms of the degree to which they match the predictions of the hierarchy.

First we shall discuss briefly the Berlin and Kay hierarchy. We shall also present the current inventory of basic colour terms in Russian (section 2). This is an important language for our purposes, since it seems to be an exception to the hierarchy, and as a result a considerable body of research has been done on it. Then we discuss the types of measure available and the statistical techniques for establishing the closeness of fit with the predictions derived from Berlin and Kay (section 3). We describe each test in turn, referring briefly to the results obtained and the degree of correspondence with the hierarchy (sections 4 and 5). This leads naturally to an

---

2 See, for example, the work of Senft (1987) on Kilivila colour terms, the interpretation of Berlin and Kay's work by Wierzbicka (1990) and extensive empirical and theoretical work by MacLaury (for which, see, for instance MacLaury 1987, 1991a, 1991b, 1992). The most recent work by the originators of this line of research is reported in Kay, Berlin and Merrifield (1991).

3 A general critique is offered too by Saunders and van Brakel (1988).



directly relevant for our purposes. It is also stated that grey may occur as a 'wild card', rather earlier than was originally thought, though this is believed to happen only infrequently. We shall not include this relaxation of the hierarchy in what follows for two reasons: first, if grey can occur at various points this makes it easier to find data matching the order of the hierarchy;<sup>5</sup> and second we wish to draw comparisons with earlier analyses which retained grey only as one of the four terms at the right end of the hierarchy.

A lower-level distinction separates out the *primary* basics, comprising the first six terms on the hierarchy (Kay and McDaniel 1978: 626; see also MacLaury 1991b: 42). These correspond to the labels for the landmark colours of Miller and Johnson-Laird (1976: 344). The remaining five basic terms we shall call '*secondary* basic'.

We should now consider the basic colour terms of Russian. A good deal of effort has been expended on Russian, in view of its exceptional nature. It has two basic terms for blue,<sup>6</sup> which gives an inventory of 12 basic terms (instead of the normal maximum of eleven):

belyj 'white', černyj 'black', krasnyj 'red', zelenyj 'green', želtyj 'yellow', sinij 'dark blue', goluboj 'light blue', koričnevij 'brown', fioletovij 'purple', rozovij 'pink', oranževij 'orange', seryj 'grey'.

While several of these are straightforward, others require comment.

---

5 Bolton (1978: 310) also retained the earlier formulation because of the increased indeterminacy of the revision, which makes it 'difficult to find an appropriate statistical technique for correlating the sequence with other variables.'

6 This possibility was noted by Berlin and Kay (1969: 99). Bolton, Curtis and Thomas (1980:317) point out that in their work on Nepali akāshi 'sky, light blue' was 'the most commonly elicited secondary term', and that there is evidence that celeste 'light blue' may be acquiring basic status in Guatemalan Spanish (Harkness 1973: 177) and Peruvian Spanish (Bolton 1978: 293-4).

Our research suggests very strongly that both terms for blue are indeed basic (see, for example, Corbett and Morgan 1988;<sup>7</sup> for instrumental data on referents of the two terms see Morgan and Moss 1988/89).<sup>8</sup> Our list varies in two respects from that provided by Slobin for Berlin and Kay (1969: 98-99): first the basic term for orange is certainly oranževyj, and not that given by Slobin. The second difference is more contentious: Slobin gives purpurnyj for purple. Corbett and Morgan (1988) showed that there are considerable problems with purpurnyj and that fioletovyj has a better claim to be considered a basic term. At that stage we were unsure whether to claim it was basic; the list experiment reported in Morgan and Corbett (1989), discussed in section 4.4.1 below, suggests very strongly that fioletovyj is indeed basic. For confirmation, and for instrumental data on the referent of fioletovyj see Moss (1989).

### 3. Assessing the tests

Tests of basicness may vary in their resolving power. The fundamental requirement is that the test distinguishes between the basic and non-basic terms; we call this the level 1 criterion (section 3.1). In addition a test may be able to establish distinctions within the basic terms. Firstly the primary basic terms may be distinguished from the secondary basic terms; the level 2 criterion (section 3.2). Secondly, rather than just dividing the basic terms into two groups, a test may reveal regularities in the orderings of the basic terms - the level 3

---

7 Unknown to Corbett and Morgan (1988), Vamling (1986) claimed that Russian has two basic terms for blue, on the basis of frequency in texts. She quoted the list of basic colour terms proposed by Kulick and Vamling (1984) which corresponds exactly to the twelve given above, having been established independently. She noted, however, (Vamling 1986:226) that fioletovyj 'purple' and oranževyj 'orange' 'seem to have a less certain status as basic colour terms.

8 Whenever we correlate Russian data with the hierarchy or with other languages we take the mean of the values for the two blue terms. And when checking whether a measure distinguishes primary from secondary basics we treat the Russian blues (mean value) as a primary term. Differences between the two blue terms are considered from the perspective of translation by Alimpieva (1982a) and from a diachronic perspective by Alimpieva (1982b). Examples from early texts are given in Baxilina (1975: 174-207).

criterion. This ordering can be compared to the Berlin and Kay hierarchy (section 3.3).

It seems likely that these three levels of discrimination will be hierarchically organised. That is, if a test reveals a consistent ordering of the terms (level 3), we expect that it will also distinguish between primary basic and secondary basic terms (level 2), and between basic and non-basic terms (level 1). And a test which discriminates between primary basic and secondary basic terms (level 2) is also likely to distinguish basic from non-basic terms (level 1). Although this seems plausible it is not logically necessary. It is possible that a test might be effective at revealing a consistent order within the basic terms, but be poor at distinguishing basic from non-basic terms. If the relationship between the three levels of resolving power was not hierarchical, it would mean that different tests served different purposes, and the choice of test would depend on whether the primary aim was to establish the inventory of basic terms, or to compare the ordering of basic terms to the Berlin and Kay hierarchy or whatever.

We will next consider how to establish whether a test meets any of the three criteria described above: firstly, by considering the patterns of data which would indicate that the criteria were met, and secondly by considering what statistical tests might establish objectively whether the criteria were met.

### **3.1 The level one criterion**

If a measure met the first criterion perfectly, then there would be a sharp cut off between the basic terms and the non-basic terms. That is, there should be no overlap between the sets of scores for the two groups of terms, and the difference between the scores of the closest of the basic and non-basic terms would be large. However, failure to reach this perfect degree of separation does not mean the measure achieves nothing; it may achieve some degree of partial separation. Such partial separation can be estimated by considering the degree of

overlap, and the difference between an 'average' score for the basic and non-basic terms. There are effective statistical procedures to assist in this process but these cannot be applied to most of the data we report here, because they require the scores for each of the non-basic terms, and these have not usually been published. We therefore concentrated mostly on the degree of overlap between the basic and non-basic sets by comparing scores for the 'worst' basic terms with the 'best' non-basic terms. Where possible we give the scores for the three best non-basic terms for a given language and test to compare these with the scores for the basic terms.

### **3.2 The level two criterion**

To establish whether a test discriminates primary basic from secondary basic terms, the level 2 criterion, requires conceptually the same procedures as for distinguishing basic from non-basic terms. That is the degree of overlap between the two sets of scores, and the difference between the 'average' score for each set. In this case however, the data for each of the 11 terms is available and we are able to apply formal statistical procedures. The advantage of the formal statistical procedure, is that for cases where there is some degree of overlap, the statistics tell us whether there is a difference between the two groups on average, and whether this difference is 'statistically significant'. Tests for the difference between averages such as the t-test and the Mann Whitney U, yield a figure for the probability that the observed difference could have arisen by chance, as opposed to primary basics and secondary basics really being different on the measure. Essentially it allows us to estimate the reliability and generalizability of the data. If the probability is low (by convention, lower than .05 or 1 in 20) then on balance the observed difference is taken to be due to the effect of the variable that distinguishes the two sets; in this case the primary basic and secondary basic groups, and the difference is said to be statistically significant.

The level of significance is also affected by whether the direction of the difference is predicted or not. In our case, the prediction is that primary basic terms will outscore the secondary basic terms on each test. We are, therefore, entitled to use the 'one tailed probability', rather than the 'two tailed probability'. The one tailed value is exactly half the two tailed value. There is a possible cost to predicting the direction of the difference; if there is a difference in the other direction, no matter how large, we must treat it as statistically insignificant. We report the one tailed probabilities in each case.

There is a choice of the kind of average to use, and of the test of differences between the averages. The main considerations in choosing an average are the *scale of measurement* that the data are on and the *pattern of distribution* of the data. As far as the *scale of measurement* is concerned, in our case the data are either measured on an ordinal scale or an interval scale. The ordinal scale is the weaker of the two; it has the property that elements are ordered on the scale, that is the ranks are meaningful. We may say, for instance, that in counting English colour term in texts, white occurs most frequently and so scores 1, black scores 2, and red 3 (Table 13). Despite the fact that each score is a number we cannot assume that equal numerical differences on the scale represent equal differences on the underlying measure. This means it is not valid to apply arithmetic operations such as addition and multiplication to an ordinal scale. The Berlin and Kay hierarchy is itself an ordinal scale and the derived ranks are shown in Table 1.

An interval scale has all the properties of an ordinal scale, but in addition it has the property that the differences between values on the scale are meaningful. For example, in a count of the frequency of English colour terms in texts (Table 13 below), white occurs 247 times, black 144 and red 142. These are on an interval scale, not only does white score more than black and black more than red, but we can legitimately say that white scores 103 more than black, whereas black

scores only 2 more than red. If we are uncertain about the nature of the scale, the safest procedure is to exploit only the scale properties we are sure of; if it is uncertain whether a scale is genuinely interval we should treat the scale as though it was ordinal. In the case of the frequencies in Table 13, treating it as ordinal is legitimate, but in this case we would lose the information about the magnitude of the differences. Treating a scale as ordinal even though it is interval is sometimes necessary if we wish to compare two scales, one of which is ordinal and one interval. This is the case when comparing the Berlin and Kay hierarchy to most of the test measures which we report to see if a test meets the level 3 criterion. Most of the tests we describe yield data on an interval scale; for example, reaction time and frequency of occurrence in texts.

The second consideration in choosing the appropriate average and statistical test, is the *pattern of distribution* of the data. In our case the problem is basically that the sample sizes, that is, the number of basic and non-basic terms, are small, and so one extreme score would have a large and unwarranted effect on some kinds of average. If there were no constraints due to sample size, the arithmetic average, the mean, would be the best for our test data. It exploits the interval nature of the measures, and uses all of the available information. Alternatively, it might be safer to use the median,<sup>9</sup> as this uses only the ordinal properties of the scale, and is not influenced by the odd extreme score. If we decide to use the mean, then the appropriate test of difference would be the t-test for independent samples; whereas if we decided to use the median then the Mann Whitney U test would be appropriate. Both tell us whether the difference is statistically significant, but the t-test is more powerful, provided the assumptions of interval data and symmetrical distributions are met.

For each test we report, we compared the primary basic and the

---

<sup>9</sup> If all scores are ordered by magnitude, the median is the middle score.

secondary basic terms using both kinds of average, and both tests. It made essentially no difference in terms of which tests distinguished primary from secondary terms, and which were statistically significant. We therefore in most cases report the means, and the results of the t-tests. In one case (number of first places in a list experiment, Table 8) we use a weaker test because the distributions of the two samples do not warrant the use of t. We report the results of the Mann Whitney U test and, as with t, we report the significance level. In reporting the results of the t tests, we normally give two figures: t, and the significance level (p). The larger the value of t, the more likely it is to be significant for a given sample size. The larger the sample size, the smaller t has to be to reach significance. The degrees of freedom (d.f.) reflect the sample size; for the independent groups t test that we use, the degrees of freedom are the sum of the sample sizes of the two groups (landmark and secondary) minus two; that is  $(6 + 5) - 2 = 9$ . As this value is the same for most tests we do not normally report it. However, on some occasions we do report the degrees of freedom. This is because we carry out a 'pre-test' of the assumptions underlying the t test; if it appears that the assumptions are not met in our data set, we use a more conservative version of the t-test (separate variance estimates) and in accord with caution, the degrees of freedom are reduced; this makes it harder to achieve statistical significance. And it is on these occasions that we report the degrees of freedom.

### **3.3 The level three criterion**

Finally, if a test satisfied the level 3 criterion perfectly, the rank orders of the terms on the test would correspond exactly to the rank order of the terms on the Berlin and Kay hierarchy. However, such perfect correspondence is rare, and we need some measure of degree of correspondence or association, to tell us how close the association is to perfection, and whether it is statistically significant.

In general the appropriate statistical assessment of the strength of an index of basicness is some kind of measure of association between the rank order of colour terms on the index with the Berlin and Kay rank order. We consider three such measures. All are correlation coefficients: Pearson's  $r$ , Spearman's  $\rho$  and Kendall's  $\tau$ . For example, Hays, Margolis, Naroll and Perkins (1972: 1113) report Spearman's rank order correlation ( $\rho$ ) between the frequency of occurrence of colour terms (a measure discussed in detail in section 5.2) and the Berlin and Kay rank order for a number of languages. To make this clearer, Table 1 shows their figures for the rank orders of English colour terms together with the Berlin and Kay ranks for the same terms.

Colour terms	Berlin and Kay rank	Hays et al. rank
<i>primary</i>		
white	1.5	1
black	1.5	2
red	3	3
green	4.5	6
yellow	4.5	8
blue	6	5
<i>secondary</i>		
brown	7	4
purple	9.5	11
pink	9.5	9.5
orange	9.5	9.5
grey	9.5	7
rho	.866 <sup>10</sup>	
p	.0031	

Table 1: Rank orders of English colour terms (frequency)

When two terms tie, as in the case of black and white, which Berlin and Kay have as first equal, they are assigned the average of the ranks they jointly occupy, hence  $1 + 2 \div 2 = 1.5$ . Similarly, the last four

---

<sup>10</sup> There appears to be a minor miscalculation here. Rho is actually .84, and the significance level is higher than they say. This does not affect the point they make.

terms occupy ranks 8, 9, 10 and 11, of which the average is 9.5. It can be seen that there is a reasonable tendency for the two sets of rank orders to correspond; white and black have the highest ranks on both measures and red is third on both. At the other end of the scale, purple, orange, pink and grey, although not corresponding perfectly, have low ranks on both measures.

Spearman's rho may be thought of as an index of how close the rank orders are to perfect correspondence. A correlation of zero would indicate minimum correspondence, and a correlation of 1 would indicate perfect correspondence. As well as varying in strength, an association can vary in direction. Two variables can either co-vary so that low ranks on one tend to go with low ranks on the other, as is the case in Table 1. Or, they could co-vary so that low ranks on one tend to go with high ranks on the other. Both pairings if sufficiently pronounced would produce high valued correlations but they would differ in sign. The value would be positive in the first case, and negative in the second. In this case the correlation is positive and is .866, which seems to indicate high correspondence.

However, as with the difference between means, the statistical significance should be considered. It is possible that any observed coefficient could result from chance, as opposed to reflecting a genuine association. The larger the correlation the less likely it is that it could just be due to chance; also the larger the sample size (in this case the number of colour terms), the less likely it is that any given coefficient could be just due to chance. In Hays et al.'s case they give the probability (p) as .0031, which is much less than .05, and therefore statistically significant. The level of significance is also affected by whether the *direction* of the relationship is predicted or not. We expect the rank on the Berlin and Kay hierarchy and the index to co-vary in the same direction; that is low ranks on the index would tend to go with low Berlin and Kay ranks. In this case, as with the t-

test discussed above, we are entitled to use the 'one tailed probability' value rather than the 'two tailed' value, and it is this value we give throughout this paper.

One drawback of using Spearman's test as a measure of association is that the reliability of this statistic is reduced when there are a large number of tied scores. The Berlin and Kay hierarchy compresses the 11 colour terms into 6 positions, and hence poses some problems for the usefulness of rho. Kendall's tau, another rank order correlation is more robust in the face of tied scores and is perhaps therefore more appropriate. Tau for the data in Table 1 is .69,  $p < .002$ . It is lower than rho, reflecting the large number of ties, but is still significant.

There is, however, a more serious problem which applies to both measures of correlation. Suppose we have a candidate index which differentiates only the primary basics (the first six) from the secondary basics, and fails to make any distinctions within these two groups. Illustrative scores and the corresponding ranks are shown in Table 2.

Colour terms	Berlin and Kay rank	Candidate score	index 1 rank	Candidate score	index 2 rank
<i>primary</i>					
white	1.5	5	3.5	4	5.5
black	1.5	5	3.5	4	5.5
red	3	5	3.5	3	4
green	4.5	5	3.5	2	2.5
yellow	4.5	5	3.5	2	2.5
blue	6	5	3.5	1	1
<i>secondary</i>					
brown	7	10	9	6	11
purple	9.5	10	9	5	8.5
pink	9.5	10	9	5	8.5
orange	9.5	10	9	5	8.5
grey	9.5	10	9	5	8.5
primary ( $\bar{X}$ )		5	3.5	2.7	
secondary ( $\bar{X}$ )		10	9	5.2	

Table 2: Illustration of two candidate indexes

Consider first candidate index 1. Both statistical measures yield large correlation coefficients: Spearman's rho = .89 and Kendall's tau = .80. Both are statistically significant; the probability (p) of a correlation as high as .89 occurring by chance is less than .0009, and of .80 less than .002. So tau yields a lower correlation which is less statistically significant.

In this case it would have been appropriate to test for a difference between primary and secondary basic terms; the level 2 criterion. In this case we will assume that the data is on an interval scale. The mean for the landmark terms is 5 and for the secondary terms is 10, and there is no overlap between the two categories. In this case t would in fact equal infinity and be infinitely significant; this arises because there is no variation in the two sets of scores. However, these artificial scores were contrived to make the point about differences masquerading as correlations. Such lack of variation

is highly improbable in real data.

Let us now consider an even less successful candidate index. Suppose that a candidate index discriminated successfully between primary and secondary terms, but within these two categories got the ranks in the reverse order to the Berlin and Kay hierarchy, as shown in the final column of Table 2. Even then Spearman's correlation still yields a large and significant outcome:  $\rho = .59$ ,  $p < .03$ . In this case, Kendall's correlation produces a figure of  $\tau = .28$ , which is non-significant:  $p < .15$ . This is more in accord with our intuition that the second candidate index is not well related to the Berlin and Kay hierarchy.

In this case comparing the mean scores of the primary and secondary terms for a significant difference yields a  $t$  value of 5.2;  $p < .001$ . The difference is less significant than for the first measures because, although there is no overlap between the two categories, the difference between the two means is smaller, and there is greater variability within a category.

In summary, neither of these two correlations has ideal properties, since in the case of our first candidate index they both yield large correlations which are to some degree illusory (based on the successful discrimination of two groups only). Kendall's correlation is probably more useful than Spearman's, but inspection of the values of the candidate measure for each colour term should be performed in order to detect whether the measure is indicating a seemingly strong correspondence which is just due to its ability to discriminate the primary terms from the secondary terms. In this paper we report  $\tau$  as the preferred correlation, but where appropriate we also report  $\rho$  in order to facilitate comparison with previous work.

Both  $\tau$  and  $\rho$  assume that the data are only measured on an

ordinal scale. In the case of the hierarchy this is clearly so, and we are obliged to use these procedures, even though the test scores are on an interval scale. However, if we wish to consider the inter-relationships between measures of basicness, without at the same time considering their relationship to the Berlin and Kay scale, then we are entitled to use a more powerful measure of association, the Pearson correlation coefficient,  $r$ . As with tau and rho,  $r$  varies between -1 and +1, but it is more powerful because it uses all the information available in an interval scale; that is, it uses the magnitude of the difference between items as well as the ordinal properties of rank. It is also more useful because the coefficient can be interpreted precisely in terms of the 'proportion of variance accounted for'; that is, the amount of variation in one set of scores that can be 'explained' by the association with the other variable.

This also allows a precise comparison of the strength of coefficients, whereas tau and rho can only be interpreted ordinally; that is we can say whether one coefficient is larger than another but we cannot easily extract a precise comparative measure. The proportion of explained variance is equal to the square of  $r$  ( $r^2$ ). This squared coefficient should also be used to compare the strength of coefficients. For example, if one coefficient  $r_1 = .6$ , and a second  $r_2 = .4$ , then the ratio of their squared values is  $.36/.16 = 1.5$ ; so  $r_1$  is actually more than twice as powerful as  $r_2$ , rather than one and a half times more powerful as the raw coefficients might have suggested. Thus in looking for consistent orderings of colour terms across different languages we report the Pearson correlation coefficients; in each section dealing with a particular measure, we report the inter-correlations between the measures given in that section.

In addition in section 6 we again exploit the interval nature of the scales by using correspondence analysis (Greenacre 1984; Weller and Romney 1990) to look for possible structures underlying the patterns of associations between different measures and languages.

## 4. Behavioural tests

The behavioural measures we describe in this section are all predicated on Berlin and Kay's notion of psychological salience (1969: 6).<sup>11</sup> We describe five measures: first, in section 4.1, reaction time (RT) as recommended by Hays et al. (1972: 1120)<sup>12</sup>; the frequency with which terms are used in naming experiments (section 4.2); the consistency of use of colour terms (section 4.3); and the frequency of occurrence and the order of occurrence of terms in elicited lists, as suggested in Berlin and Kay (1969 : 6), both discussed in section 4.4.

The first three of these measures were all derived as part of four recent studies: Boynton and Olson (1987, 1990) using English speakers, Uchikawa and Boynton (1987) for Japanese speakers, and Moss et al. (1990) for Russian speakers. We outline the common experimental procedure here, but describe the details of each measure in the relevant section.

The four studies measured the time taken to name a large sample of colours: each of the 424 OSA colours<sup>13</sup> in Boynton and his co-workers' studies and the 219 colour samples of the Color Aid Corporation's range<sup>14</sup> in the Russian study. Each of the colours was presented singly and the subject was required to name the colour with a mono-lexemic term as quickly as possible. The time from the onset of the colour display to the beginning of the vocalization was recorded and this was the RT for that trial. The entire set of colours was presented twice in different random sequences and this permitted the

---

11 Note the recent discussion by MacLaury (1991b: 56).

12 It has also been used by Rosch Heider (1972: 13-15).

13 The OSA (Optical Society of America) sample attempts to sample colour space in equal perceptual steps.

14 Based on the Ostwald Color Solid; for a technical description see Foss, Nickerson and Granville (1944). and for a more accessible account see Smith (1965).

derivation of the measure of consistency described in section 4.3.

#### 4.1. Reaction time (RT)

The data we show in Table 3 are the mean RTs across every use of a term; that is across both repetitions of the set of colours and across respondents. Thus 1949 m.s. (1.949 seconds) in Table 3, is the mean RT across all occurrences of the term

Although all four studies used equivalent procedures, there is some variation in the measures reported. We have derived comparable measures where possible from each study and these are shown in Table 3. This gives the means across subjects and colour samples for:

- (a) each of the basic terms, except black for Boynton and Olson (1987) - American sample 1, and black, white and grey for Boynton and Olson (1990) - American sample 2, where these data are not given. The scores for the three non-basic terms with the fastest RT in each study are also shown, except for Boynton and Olson (1987) where data for individual non-basic terms are not given.
- (b) the mean scores ( $\bar{X}$ ) for basic and non-basic terms
- (c) the mean scores ( $\bar{X}$ ) for primary and secondary basic terms.<sup>15</sup>

---

15 The sources for our data are as follows. For the individual terms, the Russian data are taken from Moss et al. (1990: 319, Table 2); the score for 'blue' is the mean of sinij 'dark blue', and goloboj 'light blue'. The Japanese data are from Uchikawa and Boynton (1987: 1828, Table 4). The first American English sample is based on Boynton and Olson (1987: 96-7, Figure 1) which gives the RT's to individual colour samples which were given the same name unanimously (consensus scores). This was because Boynton and Olson (1987) do not report the overall RT's for each term. Our scores based on the consensus items therefore almost certainly underestimate what the overall means would have been, but probably reflect their relative speeds reasonably well. The second American-English sample scores are derived from the graph in Boynton and Olson (1990: 1315, Figure 3). The Russian means are based on Moss et al. (1990: 322, Table 4) for basic and non-basic terms, and on Moss et al. (1990: 319, Table 2) for primary and secondary terms using unweighted means. The Japanese scores are based on Uchikawa and Boynton (1987: 1829, Table 5) using weighted means for basic and non-basic, and on (1987: 1828, Table 4) for primary and secondary, using unweighted means. The first American-English figures are taken from Boynton and Olson (1987: 95, Table 1) using weighted means for basic and non-basic terms, and are derived Boynton and Olson (1987: 96-7, Figure 1) just using consensus terms for primary and secondary terms.

	Russian	Japanese	American 1	American 2
<i>primary</i>				
white	1949	1900	1900	-
black	1892	2250	-	-
red	1776	2140	1440	2550
green	1596	2010	1450	1780
yellow	1503	2190	1360	1920
blue	1520	2180	1420	1940
<i>secondary</i>				
brown	1628	2210	1470	2360
purple	1916	2000	1500	2220
pink	1991	2310	1570	2140
orange	1958	2260	1540	1970
grey	1578	2510	1490	-
three highest	2024	2350		2690
non-basic	2059	2350		2880
terms	2061	2360		2920
basic ( $\bar{X}$ )	1646	2150	1736	2110
non-basic ( $\bar{X}$ )	2169	2480	2380	3170
primary ( $\bar{X}$ )	1706	2112	1510	2050
secondary ( $\bar{X}$ )	1814	2258	1510	2170

Table 3: Response times (milliseconds) to colour stimuli in Russian, Japanese and American English

It can be seen that for each language the mean RT for basic terms is less than for non-basic terms; in fact with the single exception of grey for the Japanese, the slowest basic term is faster than the fastest non-basic term for each language. Furthermore, within the basic terms, the mean score for the primary terms is faster than the mean for secondary terms for Russian, Japanese and the second American sample, and they are equal for the first American sample (but remember the first American sample scores are just for consensus scores, and the mean for primary terms does not include the score for black). However, the separation between the primary and secondary

---

The second set of American-English scores are derived from the graph in Boynton and Olson (1990: 1315, Figure 3) for both sets of means.

terms is much less clear cut than than for basic and non-basic terms; neither difference is significant ( $t = .92$ ,  $p = .38$  for Russian;  $t = 1.5$ ,  $p = .16$  for Japanese). For each language three of the primary terms tend to be the fastest overall, whereas some of the secondary terms are faster than the remaining primary terms in each language.

Kendall's tau was calculated between the RTs and the Berlin and Kay rank order, for Russian and Japanese only, because of the missing data for the American samples. The correlation was significant for Japanese ( $\tau = .41$ ,  $p < .05$ ), but not for Russian. The Pearson correlation ( $r$ ) was calculated for the RTs between the two languages; it was statistically non-significant.

In summary, response time serves to discriminate basic from non-basic terms well; but it does not correspond to the full Berlin and Kay hierarchy very well, nor distinguish primary from secondary terms.

Some caution is required in interpreting these results. In general within an experiment the more probable or frequent a response, the faster the response will be. In the studies reported here, there were far more basic responses than non-basic responses, which could be due to greater availability of basic names, or to the nature of the colour samples or both. It is therefore possible that the lower RTs for basic terms were due to response frequency effects.

However, some recent data from Laws and Davies (forthcoming) suggests that the difference in RT is not just a response frequency effect. They presented subjects with non-focal exemplars of basic terms, which some subjects spontaneously named with basic terms and some with non-basic terms. The RTs from these subjects using basic terms were significantly faster than those using non-basic terms. In this case, the response frequencies are matched across subjects, and even a poor exemplar of a basic term was named more quickly

with a basic term than with a non-basic term.

## 4.2. Frequency of occurrence

All four papers listed at the beginning of section 4.1 report to varying degrees the frequency with which colour terms were used by their sample. That is, the number of occasions a term or group of terms was used, summed across colour samples and respondents.<sup>16</sup>

The first part of Table 4 shows the percentage frequency with which each of the basic terms occurred for Russian and the first American-English sample.<sup>17</sup> (Figures for the other samples were not provided.) In addition the frequencies for the three most frequent non-basic terms are also shown for the Russian sample.

The second part of Table 4 shows these frequencies for the basic and non-basic terms for each of the four samples, expressed as percentages of the total responses. (A gap indicates that the data are not available.) It can be seen that basic terms are used more frequently than non-basic terms, and that the proportions are reasonably similar for the Russian, Japanese and second American-English sample at about two thirds basic terms, but that the proportion of basic terms for the first American-English sample is much higher at about 90%.

---

16 While the studies cited involved large numbers of colour stimuli (at least 200), interesting results can be obtained with relatively small numbers of stimuli, as shown by Jernudd and White (1983).

17 The figures for basic and non-basic terms were derived from Moss et al. (1990: 322, Table 4) for the Russian subjects; the frequency for blue is the mean of *goloboj* 'light blue' and *sinij* 'dark blue'. Those for the Japanese subjects are from Uchikawa and Boynton (1987: 1829, Table 5), which includes some combinations of equivalent terms. Boynton and Olson (1987: 95, Table 1) provide the data for the first American-English sample, and Boynton and Olson (1990: 1313) for the second. The figures for primary and secondary terms for Russian subjects were derived from Moss et al. (1990: 322, Table 4), and from Boynton and Olson (1987: 98, Table 2) for the first American-English sample. The figures for the individual terms were also derived from these tables.

The third part of Table 4 shows the mean percentage occurrence of primary terms and secondary terms. It can be seen that there is no marked difference between the measures for either sample, and the differences are non-significant.

	Russian	Japanese	American 1	American 2
<i>primary</i>				
white	0.9		2.0	
black	0.9		0.3	
red	3.6		4.0	
green	9.7		24.5	
yellow	5.1		7.0	
blue	8.2		14.8	
<i>secondary</i>				
brown	4.5		6.6	
purple	8.4		10.2	
pink	7.2		9.1	
orange	4.5		9.1	
grey	6.7		2.5	
three highest	6.5			
non-basic	5.2			
terms	2.6			
basic total	67.7	62.1	89.9	64.4
non-basic total	32.2	37.9	10.1	35.6
primary ( $\bar{X}$ )	4.7		8.7	
secondary ( $\bar{X}$ )	6.3		7.5	

Table 4: Percentage frequency of occurrence in four samples

It is clear that frequency of response does not perfectly separate the basic terms from the non-basic terms for Russian as 'black', 'white', 'orange' and 'brown' occur less frequently than the two most frequent non-basic terms. Comparison of the frequencies with the Berlin and Kay rank order shows that the correlation is negative ( $\tau = -.44$ ) for Russian, that is, in the opposite direction to that predicted by the hierarchy. It is therefore non-significant, since we are using the one-tailed criterion (see section 3). The correlation with the hierarchy is also non-significant for the American-English. However, the

frequencies in the two languages are strongly correlated:  $r = .80$ ,  $p < .002$ .

To a degree, as Boynton and Olson point out, this lack of correspondence between observed and predicted frequencies, may be a function of inadequate sampling of colour space, but it is clear that the volume of colour space designated by particular terms varies; and that the magnitude of the volumes does not correspond with the Berlin and Kay sequence. For example, red occupies a relatively small region, and green a relatively large one.

It is possible that the frequency with which the basic terms are evoked or selected in normal use does correspond more closely to the Berlin and Kay sequence; frequency in texts, which we discuss below (section 5.2.), may reflect this.

#### **4.3. Consistency of use**

Boynton and Olson (1987) defined a term as being used consistently if an individual used the same term for the same stimulus for the two instances when it was presented. All four studies report at least one measure based on the frequency of consistent use. Moss et al. (1990) report frequency of inconsistent use; Boynton and Olson (1987) report both frequency of consistent and of inconsistent use; Uchikawa and Boynton (1987) report the percentage of consistent use, and the ratio of consistent to inconsistent use; and finally Boynton and Olson (1990) report the ratio of consistent to inconsistent use. It is relatively simple to convert any of these measures into any of the others, and we have chosen to convert them into the percentage of consistent use. It is worth noting that it is not totally arbitrary which of these measures are used. The use of the ratio of consistent to inconsistent use, which Boynton and Olson seem to prefer, changes the relative differences within a measure compared to either the

percentage of consistent or inconsistent use. For example, imagine four terms are used with 90%, 60%, 30% and 10% consistency. Conversion of these to the ratio of consistent to inconsistent use yields figures of 9, 1.67, 0.43, 0.11 respectively. It is now tempting to say that the first term is about 5 times 'better' than the second term, and getting on for 1000 times better than the final term. In comparison with the percentage consistency measures the range of comparisons is about  $1^{1/2}$  times for the first two terms, and a ratio of nine for the first and last term. In a sense the ratio measure has spread the scale out, but by taking 'two bites at the cherry'. We believe that the first measure is closer to the raw data, makes the least assumptions, and is the figure we choose to offer. For those who see merit in the ratio measure there is sufficient information provided to perform the conversion.

Table 5 shows the percentage consistent use for each sample for:- (a) basic and non-basic terms (b) primary and secondary terms (c) for each basic term. In addition the figures for the three non-basic terms with the highest consistent use for Russian, Japanese and the second American-English sample are shown. (The equivalent data for non-basic terms is not included in Boynton and Olson 1987.)<sup>18</sup> It can be seen that the proportion of consistent use for basic terms is markedly higher than for non-basic terms for all samples, although some non-basic terms were used more consistently than some basic terms. And as the frequency of use of basic terms was greater than

---

<sup>18</sup> The figures for basic and non-basic terms were derived from Moss et al. (1990: 322, Table 4) for Russian subjects, Uchikawa and Boynton (1987: 1829, Table 5) for Japanese, Boynton and Olson (1987: 95, Table 1) for the first American-English sample and Boynton and Olson (1990: 1314, Table 1) for the second American-English sample. In each case the figure is the total number of consistent uses divided by the total number of uses expressed as a percentage. The figures for primary and secondary terms are taken from Moss et al. (1990: 319, Table 2) for Russian, Uchikawa and Boynton (1987: 1828, Table 4) for Japanese, Boynton and Olson (1987: 98, Table 2) for the first American sample, and Boynton and Olson (1990: 1314, Table 1) for the second American-English sample. In each case the figures are weighted means, except for Japanese where it is an unweighted mean as the frequency of occurrence of each term was not given. The figures for the individual terms were taken from the same tables as for primary and secondary terms.

for non-basic terms, basic terms account for a large proportion of consistent use. The figures are very similar for the four samples with the main variation due to the first American-English sample, who used basic terms more consistently than the others.

	Russian	Japanese	American 1	American 2
<i>primary</i>				
white	46.6	65.2	65.5	51.9
black	38.5	66.7	66.7	70.6
red	66.7	64.5	71.8	63.9
green	79.2	80.9	90.2	79.9
yellow	74.8	61.2	78.2	73.8
blue	73.8	61.5	89.7	79.3
<i>secondary</i>				
brown	73.0	74.2	74.9	72.5
purple	73.5	80.8	80.5	75.7
pink	73.0	70.9	73.3	71.9
orange	77.4	67.9	78.0	78.1
grey	73.5	72.6	65.3	61.1
Highest three	68.7	67.5	-	61.3
non-basic	61.9	64.0	-	56.5
terms	61.1	61.9	-	54.8
basic	71.7	72.2	81.9	75.0
non-basic	45.6	56.0	45.9	45.0
primary ( $\bar{X}$ )	63.3	66.7	77.0	69.9
secondary ( $\bar{X}$ )	74.1	73.3	74.4	71.9

Table 5: Percentage consistent responses

Comparison of consistent use for primary and secondary terms suggests that there is no reliable difference between them. In Russian and Japanese, and for the second American-English group, secondary terms are used more consistently than the primary terms, and although the first American-English sample do use the primary terms more consistently than the secondary terms, the difference is small. The figures for the individual terms confirm this picture; whilst the majority of basic terms are used more consistently than the non-basic

terms with the highest consistency, there is no trend for the primary terms to be used more consistently than the secondary terms. Further it appears that there is little correspondence between the rank orders of the individual terms and the Berlin and Kay hierarchy; in fact none of the correlations are significant; tau ranges from .14 to .32. Finally, considering the inter-relationships between the measures across samples, the Russian and the two American samples significantly intercorrelate; Pearson's correlation ( $r$ ) ranges from .57 to .82, the lowest level of significance was .05. But the Japanese data do not significantly correlate with any of the other samples;  $r$  ranges from .22 to .33.

#### **4.4 Elicited lists**

In describing their fourth criterion, psychological salience, Berlin and Kay suggest three indices (though these are not exclusive); the first is 'a tendency to occur at the beginning of elicited lists of colour terms' (1969:6). In this section we report data from list experiments in three languages: Russian (section 4.4.1), English (4.4.2) and French (4.4.3), and the inter-relationship between the list measures across languages (4.4.4).

##### **4.4.1 Russian**

Morgan and Corbett (1989) report an experiment in which 31 native speakers of Russian were asked, in Russian, to write down as many colour terms as they could think of within five minutes. After each minute subjects were asked to draw a line across their paper before continuing to add colour terms so that results could be calculated for each successive minute. Note that the measure is not exactly that envisaged by Berlin and Kay. We compare across subjects, expecting that basic terms will occur on more lists than non-basics.

Table 6 shows the number of respondents who had produced each

term after 1 minute, the number who put each term first in their lists, the rank order of these frequencies, the frequencies after 5 minutes together with their ranks. It can be seen that first, with the exception of rozovyj 'pink' and seryj 'grey', after 1 minute, both frequency measures separate the basic from non-basic terms. Second, that both frequency measures and the number of first places, are higher for the primary than for the secondary basic terms ( $t = 3.09$ ,  $p < 2.01$  for frequency after 1 minute;  $t = 2.2$ ,  $d.f. = 5.1$ ,  $p < .03$  for the first places;  $t = 2.1$   $p < .04$  for frequency after 5 minutes).

term	1st minute		rank	5 <sup>th</sup> minute	
	occurrences	(first)	(within basics)	occurrences	rank (within basics)
<i>primary</i>					
belyj 'white'	18	(9)	9	28	7
černyj 'black'	21	(1)	5.5	30	2.5
krasnyj 'red'	24	(11)	3.5	31	1
zelenyj 'green'	25	(2)	2	30	2.5
želtij 'yellow'	21		5.5	27	10
sinij 'dark blue'	24	(4)	3.5	28	7
goluboj 'light blue'	27	(1)	1	29	4.5
<i>secondary</i>					
koričnevij 'brown'	17		10	29	4.5
fioletovij 'purple'	20		7	27	10
rozovij 'pink'	12	(1)	11	26	12
oranževij 'orange'	19		8	28	7
seryj 'grey'	10		12	27	10
<i>highest non-basic<sup>19</sup></i>					
beževij 'beige'	13			22	
bordovij 'claret'	13			20	
sirenevij 'mauve'	12			20	
primary ( $\bar{X}$ )	22.4	(4.9)		29.1	
secondary ( $\bar{X}$ )	15.6	(0.2)		27.4	

Table 6: Number of Russian informants offering each term in the list experiment (31 subjects)

The degree of correlation with the hierarchy after 1 minute is  $\tau = .32$  (not significant) and that after 5 minutes is  $\tau = .52$ ,  $p < .02$ . The data on first position (given in parentheses) significantly correlate with the hierarchy ( $\tau = .51$ ,  $p < .025$ ), but it seems this is primarily due to a clear difference between primary and secondary terms; only rozovij 'pink' of the secondary terms ever appears first on a list, and only želtij 'yellow' of the primary terms never appears first on any

<sup>19</sup> The highest non-basics are given after one minute (for better comparison with studies for which only data from shorter time limits are available). If taken after five minutes they would be:

	occurrences after one minute	occurrences after five minutes
beževij 'beige'	13	22
malinovij 'raspberry'	11	22
salatovij 'light green'	10	22

lists.<sup>20</sup> However, the precise value of tau needs treating with caution because of the relatively small sample, and the large number of tied scores.

#### 4.4.2 English

For cross language comparison we turn to Battig and Montague's classic work (1969). They took two groups of students, from the Universities of Maryland and Illinois, 442 subjects in all. Their procedure included a total of 56 subject headings, for which subjects were asked to list items for 30 seconds each. Colour was one of the categories investigated. The procedure proved to be a good indicator of basicness; the 11 basic terms of English occurred within the first 12 places (violet occurred in eleventh place, so that gray was in fact twelfth). The results relating to the eleven basic terms are given in Table 7.

---

<sup>20</sup> The other first choices on the list were be\ evyj 'beige' and kremovyj 'cream', each put first by a single subject.

term	total		rank (within basics)	Illinois occurrences	Maryland occurrences
	occurrences	(first)			
<i>primary</i>					
white	272	(2)	8	86	186
black	314	(8)	6	111	203
red	435	(319)	2	166	269
green	431	(9)	3	169	262
yellow	387	(9)	4	152	235
blue	438	(77)	1	169	269
<i>secondary</i>					
brown	217	(2)	10	80	137
purple	282		7	97	185
pink	224	(8)	9	94	130
orange	382	(6)	5	154	228
gray	94		11	29	65
<i>highest non-basics</i>					
violet	153	(1)		62	91
turquoise	71	(1)		27	44
gold	56			19	37
primary ( $\bar{X}$ )	379.5	(70.7)		142.2	237.3
secondary ( $\bar{X}$ )	239.8	(3.2)		90.8	149.0

Table 7: Number of American informants offering each term in the list experiment (total 442 subjects)

The large samples used by Battig and Montague show the strengths and weaknesses of the list technique. It proves efficient at separating basic from non-basic terms, with only one 'outsider' namely violet scoring higher than a basic term. Similarly the primary basic terms score significantly higher than the secondary basic terms on all measures except the number of first places (t ranges from 2.13 to 2.98, p ranges from .01 to .03).

However, none of the frequency measures correlates significantly with the hierarchy (tau ranges from .16 to .36). The obvious problem is the low position of white and black; it seems that English speakers are uncertain as to whether they should be counted as colours or not<sup>21</sup>

---

21 Bolton (1978: 294, 308) reports a similar discrepancy involving white and black with Spanish speakers in Peru, as do Bolton et al. (1980: 318) for speakers of Nepali. For data on Luganda, where white and black also had low positions in a listing task, see Pollnac (1975:92).

(the same explanation probably holds for why gray occurred less frequently than violet).

The position in informants' lists might seem intuitively to correspond better to degree of salience, than the total frequency, which reflects mere occurrence. However, there is a problem in carrying out this comparison, as Battig and Montague's data does not seem internally consistent. The mean positions given for blue (1.6) and for red (2.7), do not seem consistent with the number of first places shown for red (319 out of a possible 442), and blue (77 out of a possible 442).

We have similar data on a British sample, consisting of 47 policewomen and members of the Psychology Department volunteer panel. They were asked individually to write down as many colour terms as they could; there was no absolute time limit, but if they paused for more than about 20 seconds the procedure was concluded. Table 8 shows the frequency, the mean rank, and the number of first places for each of the eleven basic terms, and the three most frequent non-basic terms.

Term	Frequency	Mean rank	Number of first places
<i>primary</i>			
white	42	9.9	1
black	42	8.0	1
red	43	2.5	25
green	45	5.0	2
yellow	43	4.5	2
blue	45	2.9	10
<i>secondary</i>			
brown	36	11.5	0
purple	39	8.8	1
pink	36	8.8	0
orange	44	9.1	0
grey	37	12.4	0
<i>non basics</i>			
turquoise	30	9.5	2
mauve	26	10.9	0
beige	20	14.1	0
primary ( $\bar{X}$ )	43.3	5.5	7.0
secondary ( $\bar{X}$ )	38.4	10.1	0.2

Table 8: Frequency, mean rank and number of first places for the British sample (n = 47)

It can be seen that the frequency measure separates the basic from the non-basic terms effectively, and that with the exception of orange, all the secondary basics have lower frequencies than the primary basics ( $t = 3.3$ ,  $p < .005$ ). However, the correlation with the hierarchy is non-significant.

The mean rank does not separate the basics from the non basics so clearly; turquoise and mauve have lower means than grey and brown; turquoise is also lower than white. However, with the exception of black and white, the primary terms have lower means than the secondary terms ( $t = 3.1$ ,  $p < .006$ ). Again the correlation with hierarchy is non significant.

The number of first places for the primary basic is higher than for the secondary basics (only purple ever occurs first on a list),  $z = 2.7$ ,  $p$

< .003; but turquoise has more firsts than any of the secondary basics. In this case, however the correlation with the hierarchy is significant ( $\tau = .5$ ,  $p < .03$ ). Again the figure should be treated with caution because of the small sample, and large number of ties.

Overall, the frequency, and the mean rank measure seem to serve slightly different functions. Frequency separates basic from non basic terms effectively, but mean rank discriminates within the basic terms best; the non significant correlation with the hierarchy is mainly because of the low means of black and white.

#### **4.4.3 French**

We also have data on a list experiment for French (Morgan 1993). 74 first-year undergraduate students of English at the University of Toulouse took part. The results after one minute and after five minutes are given in Table 9:

term	one minute		five minutes		
	occurrences	(first) rank (within basics)	occurrences	rank (within basics)	
<i>primary</i>					
blanc 'white'	65	(8)	6	71	7
noir 'black'	71	(4)	2.5	74	1
rouge 'red'	67	(11)	4.5	72	4
vert 'green'	72	(5)	1	72	4
jaune 'yellow'	67	(6)	4.5	73	2
bleu 'blue'	71	(30)	2.5	72	4
<i>secondary</i>					
marron 'brown'	49		11	68	10
violet 'purple'	56		9	70	9
rose 'pink'	59	(6)	7	71	7
orange 'orange'	57	(3)	8	71	7
gris 'grey'	51	(1)	10	67	11
<i>highest non-basic<sup>22</sup></i>					
beige 'beige'	32		-	67	
mauve 'mauve'	24		-	43	
(bleu) indigo 'indigo'	21		-	34	
primary ( $\bar{X}$ )	68.8	(10.7)		72.3	
secondary ( $\bar{X}$ )	54.4	(2.0)		69.4	

Table 9: Number of French informants offering each term in the list experiment (74 subjects)<sup>23</sup>

After one minute there was a clear gap between the basic terms and non-basics (the next term beige 'beige' appeared on only 32 lists).

22 The highest non-basics are given after one minute . If taken after five minutes they would be:

	occurrences after one minute	occurrences after five minutes
beige 'beige'	32	67
(bleu) marine 'navy blue'	20	56
fuchsia 'fuchsia'	13	44
(bleu) ciel 'sky blue'	6	44

23 Forbes (1976: 71; 1986: 89-90) reports an experiment in which she asked for 'les noms des couleurs principales en français' ('the principal colour terms in French'); she asked 100 informants, and the terms were mentioned by the following numbers of informants: blanc 'white': 95; noir 'black' 93; rouge 'red' 92; vert 'green' 86; jaune 'yellow' 85; bleu 'blue' 92; brun 'brown' 53; marron 'brown' 50; violet 'purple' 63; rose 'pink' 60; orange 'orange' 73; gris 'grey' 51.

After five minutes, although the overall correspondence to the Berlin and Kay hierarchy was better, the gap between basics and non-basics had closed (marron 'brown' appeared 68 times and beige 'beige' 67, equal with gris 'grey').<sup>24</sup> It is interesting to note the figures for first occurrence on list (in brackets); as with English the two leaders in this regard are bleu 'blue' and rouge 'red', though in the reverse order to English.<sup>25</sup> When we look at the rankings after five minutes we see that the basic terms appear on nearly all lists (and, as mentioned, some non-basic terms are 'catching up') so that there is little separating them.

All three measures significantly distinguish between the primary and secondary basic terms. (t ranges from 1.91 to 6.8, p ranges from .009 to .05; d.f. for number of first places is 5.8); the frequency after one minute does so most effectively. Both frequency measures correlate significantly with the hierarchy (tau = .42 after 1 minute and .43 after 5 minutes; p <.05) but this correlation is again largely due to the difference between primary and secondary basic terms.

#### **4.4.4 Inter-relationship between the list measures across languages**

Although the correlations between the list measures and the hierarchy are generally in the expected direction (12 out of 13 are positive), only five reach statistical significance; these are for Russian after 5 minutes, the number of Russian first places, English mean rank, and French after both 1 minute and 5 minutes. These correlations range from .43 to .52, p < .05 in all cases. In general the correlations are mainly due to distinguishing between primary basic and secondary basic terms.

---

24 Note that brun 'brown' appeared on only 4 lists in the first minute and on 9 by the end of the five minutes.

25 For a limited list experiment with Ukrainian subjects see Priestly (1981–83). And for a related type of experiment, in which speakers were asked to specify the colour names necessary for a minimum colour lexicon, see von Wattenwyl and Zollinger (1979).

In contrast the correlations across languages on equivalent measures are generally strong; the correlations between the French and Russian frequencies after 1 minute, the American total, and the English frequency (the most equivalent measures) range from  $r = .75$  to  $.92$ ,  $p < .004$  in all cases. This suggests that there is a consistent pattern across languages over and above that due to the hierarchy (as noted previously by Bolton et al. 1980:318).

## 5. Linguistic tests

We now turn to the linguistic tests. First we look at derivational morphology (section 5.1), which relates to one of Berlin and Kay's subsidiary criteria. Then we look at frequency in texts (section 5.2), and finally at word length (section 5.3).

### 5.1. Derivational morphology

One of Berlin and Kay's subsidiary criteria for basicness, number (v), refers to 'distributional potential' (1969:6) and the example given involves derivational morphology. Dixon (1982: 23-4) made a stronger claim:

the colour term hierarchy is rather strict and it is unlikely that a word low in the hierarchy would have morphological/syntactic possibilities that a higher word lacks.

This claim is too strong; when we examine Russian data we find that there are numerous cases of particular derivational possibilities being open to a term low on the hierarchy but not to all those above it. Nevertheless, terms high on the hierarchy tend to have greater derivational possibilities than those below. Using the derivational dictionary of Worth, Kozak and Johnson (1970), Corbett and Morgan

established the derivational possibilities for each of the basic terms. Thus belyj 'white', has derivatives such as belovatyj 'whitish', snežnobelyj 'white as snow', belit' 'to whiten' and so on, a total of 135 derived forms. The details have been presented elsewhere (Corbett and Morgan 1988: 36-42) and so just the total of derived forms will be given here (Table 10). Note that krasnyj 'red' was difficult to analyse since it originally meant 'beautiful'; forms with this meaning (like modern prekrasnyj 'beautiful') were not included in the count.

colour term	number of derived forms	rank (basic terms)
<i>primary</i>		
belyj 'white'	135	1
černyj 'black'	67	3
krasnyj 'red'	39	7
zelenyj 'green'	63	4
želtyj 'yellow'	56	5
sinij 'dark blue'	71	2
goluboj 'light blue'	43	6
<i>secondary</i>		
koričnevyyj 'brown'	3	11.5
fioletovyyj 'purple'	6	10
rozovyyj 'pink'	13	9
oranževyyj 'orange'	3	11.5
seryj 'grey'	20	8
<i>highest non-basics<sup>26</sup></i>		
buryj 'brown'	18	
lilovyyj 'lilac'	7	
bagrovyyj 'crimson'	3	
primary ( $\bar{X}$ )	69.5	
secondary ( $\bar{X}$ )	9.0	

Table 10: Number of derivatives of colour terms in Russian.

It can be seen that this test separates basic terms from non-basic only

---

26 The highest non-basics quoted are those identified in the study by Corbett and Morgan (1988: 38); since they investigated only terms which had some possibility of being considered basic, it is possible that some other term has more derivatives. Note that ranks are shown only for basic terms, to allow for consistent comparison with sets of data where there is information on basic terms only.

to a limited degree; there are non-basics like buryj 'brown' which have more derived forms than basics. There is good evidence that koričnevij is indeed the basic term for 'brown', and not buryj: this is shown by, for instance, data on elicited lists (Table 6). This is an instance of an interesting effect noted in Corbett and Morgan (1988: 56-8): when a new basic term displaces an older term,<sup>27</sup> the new term (koričnevij in this case) will lag behind the old in number of derivations, even when other measures confirm it as basic.

The test does separate the primary and secondary basic terms; there is no overlap between the two sets and the difference is highly significant ( $t = 4.3$ ,  $d.f. = 5.6$ ,  $p < .003$ ). Similarly, the degree of correlation with the hierarchy is  $\tau = .64$ ,  $p < .005$ .

For comparison we turn to French. Lacroix (1978: 183) provides data on just 23 derivational possibilities for French colour terms according to Littré 's dictionary (see Table 11). Note that there is doubt as to the basic term for brown; marron appears to be in the process of replacing brun (Forbes 1979; 1986). We use marron here for consistency with the statistics cited in sections 4.4.3 and 5.2. Note, however, that brun 'brown' had six derived forms in the count by Lacroix, which would give it rank 8. As noted in the discussion of Russian, when a basic term is replaced, the number of derivatives takes longer to adjust than does, for instance, frequency in texts. Given these problems the data must be treated with considerable caution.

---

27 See Baxilina (1975:208-34) for details of the development of the terms for brown in Russian.

colour term	number of derived forms	rank (basic terms)
<i>primary</i>		
blanc 'white'	14	1.5
noir 'black'	9	6
rouge 'red'	10	5
vert 'green'	14	1.5
jaune 'yellow'	13	3
bleu 'blue'	8	7
<i>secondary</i>		
marron 'brown'	0	11
violet 'purple'	3	9
rose 'pink'	5	8
orange 'orange'	2	10
gris 'grey'	11	4
<i>highest non-basics</i>		
roux 'russet'	14	
blond 'blond'	11	
brun 'brown'	6	
primary ( $\bar{X}$ )	11.3	4.0
secondary ( $\bar{X}$ )	4.2	8.4

Table 11: Number of derivatives of colour terms in French (out of 23 possibles)

There is no clear-cut separation of the basic and non-basic terms; the highest non-basic terms; the same order of derivatives as the highest basic terms. However, the primary basic terms have significantly more derivatives than the secondary basic terms ( $t = 3.4, p < .004$ ), and there is a modest correlation with the hierarchy ( $\tau = .44, p < .04$ ). Note that this is one example where the measure fails as a level 1 test, but succeeds as a level 2 test.

## 5.2. Word frequency in texts

Hays et al. (1972) used the frequency of colour terms in continuous text as an index of psychological salience. We will first consider the languages for which the best data are available (Russian and English), and then consider others for comparison (French, Hebrew and

Spanish).

Russian was one of the languages considered by Hays et al.; since that study, better sources of frequency data have become available, notably Zazorina (1977), an investigation of over one million words of twentieth century Russian. From this we can extract the data given in Table 12.

The three highest non-basic terms outscore four of the basic terms, and so the measure does not reliably pass the level 1 criterion. However, the difference between primary and secondary basic is almost complete; the exception is *seryj* 'grey' ( $t = 3.8$ , d.f. = 5.8,  $p < .005$ ).

term	number of occurrences	rank (basic terms)
<i>primary</i>		
belyj 'white'	471	2
černyj 'black'	473	1
krasnyj 'red'	371	3
zelenyj 'green'	216	4
želtyj 'yellow'	109	8
sinij 'dark blue'	180	5
goluboj 'light blue'	137	6
<i>secondary</i>		
koričnevij 'brown'	23	10
fioletovij 'purple'	22	11
rozovij 'pink'	49	9
oranževij 'orange'	15	12
seryj 'grey'	116	7
<i>highest non-basics</i>		
belosnežnyj 'snow-white'	67	
ryžyj 'ginger'	59	
buryj 'brown'	31	
primary ( $\bar{X}$ )	299.8	
secondary ( $\bar{X}$ )	45.0	

Table 12: Frequency of basic colour terms in Russian texts

The degree of correlation with the hierarchy is high,  $\tau = .77$ ,  $p < .001$ , but three basic colours are 'out of order' and deserve comment. First želtyj 'yellow' has fewer occurrences than would be suggested by its position on the hierarchy. We have no explanation for this, but it seems to be a common problem; Hays et al. (1972: 1112) found yellow lower in frequency than expected in English, French, German and Spanish, as well as Russian. Koričnevij 'brown' is also low; this is because it is in the process of taking over from an earlier term for brown, namely buryj (Corbett and Morgan 1988: 45, 48, 51-2). As can be seen, in the count reported here, buryj is actually more frequent than koričnevij. This may well result from the prevalence of certain set expressions (like buryj ugol') in this corpus. In other counts of twentieth century Russian, reported in Corbett and Morgan (1988: 47), koričnevij is the more frequent. Seryj 'grey' is higher than expected; this perhaps should be connected with its status as a 'wild card'.

There are other sources for frequency data available for Russian, though from smaller corpora. Corbett and Morgan calculated average ranks, taking four smaller corpora and counting Zsorina's four corpora separately. The ranking remains as in Table 12 above and so further details would be superfluous here (see Corbett and Morgan 1988: 50).<sup>28</sup>

Good data on English are available for comparison in Johansson and Hofland's (1989) analysis of the LOB corpus.<sup>29</sup> The Lancaster-

---

28 For data from earlier periods of Russian see Vamling (1986:229-30). Unfortunately Vakar's (1966) word count gives too short a list for most of the colour terms to figure.

29 Moskovič (1969:18) analysed a somewhat larger corpus (1,452,000 words) of English prose, mainly recent; he does not give the actual number of occurrences, but does give the rank order, which shows a remarkable fit with Berlin and Kay, the more impressive since he was unaware of their work. It is: 1 white, 2 black, 3 red, 4 green, 5 yellow, 6 blue, 7 grey, 8 brown, 9 purple, 10 pink, (11 scarlet), 12 orange. This also illustrates that the frequency test performs well in separating basics from non-basics (in the LOB corpus too, scarlet occurs more frequently than orange, being found 14 times as compared to 7 for orange). If we again consider just the basic terms, the degree of correlation of the rank order found by

Oslo/Bergen corpus consists of a million words of modern British English. It has been tagged, so that frequencies of colour terms functioning as adjectives can be separated from other uses (thus allowing more reliable comparison with other languages). The data are as in Table 13.

term	occurrences	rank (basic terms)
<i>primary</i>		
white	247	1
black	144	2
red	142	3
green	85	4
yellow	31	8
blue	83	5
<i>secondary</i>		
brown	61	7
purple	16	10
pink	28	9
orange	7	11
grey	80	6
<i>highest non-basics</i>		
silver <sup>30</sup>	22	
bronze	14	
scarlet	14	
primary ( $\bar{X}$ )	122.0	
secondary ( $\bar{X}$ )	38.4	

Table 13: Frequency of basic colour terms in English texts

As with Russian, the three most frequent non-basic terms have scores higher than at least one basic term, orange; and the highest non-basic term, silver, also outscores purple. The frequencies of silver and bronze are probably over-estimates of their use as colour

---

Moskovič with Berlin and Kay is  $\tau = .89$ ,  $p < .0009$ . Moskovič's study provides several interesting comparisons with the present work - for example he considers the number of phraseologisms into which the different terms enter, and also the translation equivalents between English, French and Russian. Note, however, that Kikuchi and Lichtenberg (1983) argue, on the basis of data from 29 languages, that the development of figurative use of colour terms is not constrained by the Berlin and Kay hierarchy.

30 Though the information on tagging enables us to select only adjectival usage, we do not know how many examples of silver and bronze relate strictly to colour.

adjectives, however. It appears that frequency in texts does better in English than in Russian in terms of the level one criterion. With the exception of yellow, the primary terms all score higher than the secondary terms, ( $t = 2.5$ ,  $p < .03$ ) and the correlation with the hierarchy is high:  $\tau = .77$ ,  $p < .001$ .

There are some striking points of comparison with Russian. First the number of occurrences of colour terms in a corpus of comparable size is markedly lower in English.<sup>31</sup> This deserves investigation, but probably results from the composition of the corpora, literature having a larger place in the Russian corpus. When we turn to the ranking of the individual terms, two surprising similarities are found. First there is the unexpectedly low position of yellow as with Russian eltyj (a recurring problem as mentioned earlier). And second, grey, like seryj again shows evidence of its 'wild card' status, occurring much higher than would have been expected from the original hierarchy.

For comparison we turn to French. The best data to hand are provided by Forbes (1976: 43), extracted from *Le Dictionnaire des fréquences du T.L.F.*, based on a corpus of over 37 million words for the twentieth century, mainly from literary texts. Relevant data are given in Table 14. Note that there is a problem with the three terms which are homographs (indicated '(H)'): there is no way of isolating the examples of, for instance, rose the colour 'pink' from rose the flower.

---

31 This is particularly interesting since in early Russian writing, of the 11th–12th centuries but even as late as the 17th century, the use of colour terms seems to be largely avoided (Baxilina 1975: 14, 56).

term	occurrences	rank
<i>primary</i>		
blanc 'white'	12758	2
noir 'black'	14198	1
rouge 'red'	7914	3
vert 'green'	4449	6
jaune 'yellow'	3121	8
bleu 'blue'	6328	4
<i>secondary</i>		
marron 'brown' <sup>32</sup>	100	11
violet 'purple'	413	10
rose 'pink' (H)	5413	5
orange 'orange' (H)	443	9
gris 'grey' (H)	3890	7
primary ( $\bar{X}$ )	8128.0	
secondary ( $\bar{X}$ )	2051.0	

Table 14: Frequency of basic colour terms in French texts

The primary terms in general have higher scores than the secondary terms, but rose 'pink' and gris 'grey', score higher than some primary terms. Never the less the difference is significant ( $t = 2.7, p < .02$ ). The correlation with the Berlin and Kay hierarchy is  $\tau = .57, p < .01$ .

At this point it is worth comparing data on other languages from (Hays et al. 1972: 1112-13) particularly for those languages for which they have data for all or almost all terms on the hierarchy (they do not report on non-basics). They report data on Hebrew supplied by Asa Kasher (from a basic word list by Balgur). These data are given in Table 15.

---

32 Recall the problem with 'brown' in French, discussed in section 5.1; the other term for 'brown', brun, occurred 1734 times in this corpus, which - other things being equal - would put it at rank 9. Forbes gives data at this point only on the eleven basics and the additional problematic term for brown.

term	occurrences	rank
<i>primary</i>		
lavan 'white'	118	1
shahor 'black'	117	2
adom 'red'	94	3
yaroq 'green'	57	4
tzahov 'yellow'	35	5
kahol 'blue'	31	7
<i>secondary</i>		
hum 'brown'	16	8
argaman 'purple'	6	10
varod 'pink'	10	9
katom 'orange'	0	11
afor 'grey'	33	6
primary ( $\bar{X}$ )	75.3	
secondary ( $\bar{X}$ )	13.0	

Table 15: Frequency of basic colour terms in Hebrew texts

This measure both separates the primary and secondary terms with the exception of afor 'grey' ( $t = 3.6$ ,  $d.f. = 6.2$ ,  $p < .005$ ), and correlates highly with the hierarchy ( $\tau = .85$ ,  $p < .0008$ ).

The other language for which Hays et al. give data on all the terms is Castilian Spanish (from Buchanan 1927), reported in Table 16.

term	occurrences	rank
<i>primary</i>		
blanco 'white'	420	1
negro 'black'	290	2
rojo 'red'	109	5
verde 'green'	120	4
amarillo 'yellow'	36	6
azul 'blue'	126	3
<i>secondary</i>		
moreno 'brown'	31	7
cardeno 'purple'	16	10
rosado 'pink'	19	9
naranja 'orange'	28	8
gris 'grey'	15	11
primary ( $\bar{X}$ )	183.5	
secondary ( $\bar{X}$ )	21.8	

Table 16: Frequency of basic colour terms in Castilian Spanish texts

Hays et al. do not discuss the relative claims of the competing terms for basic status; we shall simply take the more frequent:<sup>33</sup> unlike them we shall not use data from a second count in which not all the terms are represented.

The primary and secondary terms are clearly separated ( $t = 2.8$ ,  $d.f. = 5.0$ ,  $p < .02$ ), and the degree of correlation between the Spanish data and the hierarchy is  $\tau = .77$ ,  $p < .001$ .

As well as correlating strongly with the hierarchy, all the frequencies in texts discussed in this section strongly inter-correlate. The ten correlations vary between  $r = .84$  and  $r = .99$ ; the range of probabilities is  $<.001$  to  $<.0009$ . The lowest value is for the correlation between English and French, and the highest for Hebrew and Russian.

---

33 Except for grey, where the first form they give is rather surprising and so we take the more obvious choice. An informant questioned the status of the terms for 'brown' and 'purple': note that the source Hays et al. were using dates from 1927.

### 5.3. Word length

This measure was proposed by Durbin (1972: 270-2).<sup>34</sup> His argument ran as follows: Zipf claimed that as the relative frequency of a word increases so it will tend to diminish in length. If all languages have passed through Stage I (having terms for black and white), then those two terms would have had a high frequency of usage since they would have been the only basic colour terms. As more terms were added, all colour terms would have had a relatively lower relative frequency since the semantic domain of colour would have been covered by more terms. Hence:

- '(a) Black and white should be about the same length in all languages and all stages.
  
- (b) There should be an increase in length running from black and white to purple, pink, orange and grey.
  
- (c) A dependent corollary of (b) above would be that the average word length of Stage I languages should be less than later stages, with Stage VII having the greatest average word length.' (1972: 271).

These hypotheses are partially confirmed; the main effect, however, is to draw a distinction between the stage of development where black, white, green, yellow and blue are found as opposed to that with the full set of eleven terms (Hays et al. 1972: 1111).

Let us apply this measure to some of the languages we have already investigated (see Table 17). The Table shows two measures of word length; first the number of phonemes, and second (after the slash) the number of syllables for each basic term, for four languages:

---

<sup>34</sup> See also Rosch Heider (1972: 13-15).

English, French, Russian and Spanish. In addition, for the first three languages, it also gives the word lengths for the shortest non-basic terms which we have identified from the list experiments (section 4.4).<sup>35</sup>

It is clear that word length fails the level one criterion. For both French and Russian, there is at least one non-basic term shorter than the shortest basic term, and two other non-basic terms equal in length to the shortest basic terms. For English, the three shortest non-basics are as short as the shortest basics, and shorter than six of the basic terms.

---

35 From the data available from the list experiments, starting from the 'near misses', we identified three terms which are shorter than, or as short as, the basic terms; this is sufficient to demonstrate that word-length fails as a level one criterion. It does not guarantee, of course, to find the shortest colour terms.

term	English	French	Russian	Spanish
<i>primary</i>				
white	3/1	3/1	3/1	6/2
black	4/1	4/1	4/1	5/2
red	3/1	3/1	5/1	4/2
green	4/1	3/1	5/2	5/2
yellow	4/2	3/1	4/1	7/4
blue	3/1	3/1	***	4/2
<i>secondary</i>				
brown	4/1	4/2	8/3	6/3
purple	4/1	5/2	8/4	7/3
pink	4/1	3/1	5/2	6/3
orange	4/2	4/2	7/3	7/3
grey	3/1	3/1	3/1	4/1
<i>non-basics</i>				
first <sup>36</sup>	3/1	2/1	2/1	
second	3/1	3/1	3/1	
third	3/1	3/1	3/1	

primary ( $\bar{X}$ ) 3.5/1.2 3.2/1.0 4.2/1.3 5.2/2.3

secondary ( $\bar{X}$ ) 3.8/1.2 3.8/1.6 6.2/2.6 6.0/2.6

Table 17: Length of basic colour terms (phonemes / syllables)

NOTES:

- (i) The terms counted for each language are as in Table 14 for French, Table 12 for Russian, and Table 16 for Spanish.<sup>37</sup>
- (ii) The citation form of Russian adjectives is the nominative singular masculine form. Since, however, the length of the ending varies for other agreement forms it seems more appropriate to count here the length of the stem. (The length of the nominative singular masculine can be obtained by adding two phonemes/one syllable in each case.)<sup>38</sup>

---

36 The terms identified were:

English: tan (3/1), beige (3/1), rose (3/1), and so on.

French: or 'gold' (2/1), beige 'beige' (3/1), mauve 'mauve' (3/1), and so on.

Russian: al-yj 'scarlet' (2/1), ryž-yj 'ginger' (3/1), sed-oj 'white (of hair)' (3/1), and so on.

37 We are grateful to Carol Sanders for help with Romance data.

38 The Russian data draw attention to a problem with Durbin's approach, of which he is fully aware, namely that a new term may replace an older one by 'borrowing, diffusion, or internal restructuring'. In Russian, krasnyj 'red' has replaced earlier terms based on the root [erv- (Baxilina 1975: 134-73), and kori[nevyj 'brown' has supplanted buryj (as discussed above). It is interesting to note, however, that the term which is probably the most recent addition

- (iii) For Russian blue (\*\*\*) , the figures are: sin' - 'dark blue' (3/1), golub- 'light blue' (5/2).

However, turning to the other criteria, we find that primary basic terms are significantly shorter than secondary basic terms for French syllables ( $t = 2.7, p < .02$ ), Russian phonemes ( $t = 2.0, d.f. = 4.8, p < .05$ ), and Russian syllables ( $t = 2.5, d.f = 4, p < .03$ ). These two languages are also weakly correlated with the hierarchy (French syllables  $\tau = .48, p < .05$ , Russian syllables  $\tau = .58, p < .02$ ) but these correlations are almost entirely due to the differences between primary and secondary terms. In general, word length - whether measured by phonemes or by syllables - is only weakly associated with the hierarchy. The correlations are all positive, that is in the expected direction, but the only significant values are for French syllables ( $\tau = .48; p < .05$ ), and for Russian syllables ( $\tau = .58 p < .02$ ).

The measures of word length are reasonably correlated across languages. Of the 28 inter-correlations of the measures shown in Table 17, all are positive, and 15 of them are significant ( $r$  ranged from .52 to .93, the range of probabilities is from .05 to  $<.009$ ). The majority of the low correlations are between English and Spanish syllables and the other measures, although they are significantly correlated themselves. Every language is significantly correlated with every other language on one or other measure. In fact the correlation between these related languages in respect of word length is in general higher than the correlation of word length with the hierarchy.

## 6. The inter-relationship between measures and languages

In earlier sections we discussed a range of measures in turn, in relation to different languages. It was clear that for three kinds of measure, frequency in texts, word length, and frequency in elicited lists, there

---

to the inventory of basic colour terms, namely *fioletovyj* 'purple', is also the longest (equal to *kori[nevyj* 'brown').

were marked consistencies across languages. In some cases, this consistency was over and above that due to a common relationship with the hierarchy.

What we have not done is consider the relationships *between* measures. One way of doing this would be to produce the correlation matrix between all the measures we have discussed. This would yield over 600 correlations, thus presenting enormous problems in detecting patterns of interrelationships, and undermining the usefulness of statistical significance. In 600 correlations, 30 correlations would appear to be genuine significant associations, when in fact they might be the products of chance; and there is no reliable way of distinguishing the genuine from the chance correlations.

There are powerful multi-variate techniques such as factor analysis for detecting structural patterns underlying large correlation matrices, but they require a large sample (in our case of colour terms) to be valid. A sample of 11 is much too small for such techniques. Correspondence analysis (Greenacre 1984; Weller and Romney 1990) is a technique for exploring patterns of association in large data matrices and it is less stringent in its requirements. In brief, it attempts to discover a number of 'vectors' - structures underlying the patterns of co-variation - and each case (colour) and variable (measure) has a value (weight) on each vector. The analysis provides a measure of 'percentage of inertia' for each vector, which in essence is a measure of the strength, or degree of influence, of that vector,<sup>39</sup> and a measure of the degree of statistical significance of each vector. There is a spatial metaphor underlying the product of the analysis. Each element (colours and measures) is plotted in a graph whose axes are the vectors which are statistically significant. The more similar a pair of elements, the closer together they appear to be in the graph.

---

39 This is analogous to the percentage of variance accounted for in regression or factor analysis.

Moreover, the distance apart of two elements can be seen to be due to component distances on each of the vectors. By inspecting which elements differ on which vectors, it is sometimes possible to interpret a vector in terms of some meaningful property of the elements.

The data matrix of the 11 basic terms and the 36 measures reported in Tables 3 to 17 (including all cases where there are data for every basic term) were subjected to correspondence analysis. The analysis revealed one major vector which accounted for 85% of the inertia ( $p < .00009$ ), a second vector which accounted for 7.6% of the residual inertia ( $p < .00009$ ), and six other vectors which, although statistically significant, had a more or less negligible magnitude of effect on the structure. Figure 1 shows each measure plotted in the two dimensional space of the first two vectors; the major vector forms the horizontal axis, and the minor vector the vertical.<sup>40</sup> It can readily be seen that the variables form clusters; these clusters are generally groups of equivalent measures across languages. Thus, as would be expected from the patterns of correlations already discussed, measures of frequency in texts from different languages tend to cluster together. At the other end of vector 1 (the right hand side in Figure 1) is a second large cluster, consisting of the measures of word length for each language and the behavioural measures (response time, frequency of naming, consistency and elicited lists). However, first places in elicited lists surprisingly do not fit here, but appear as a dispersed group in the bottom left hand quadrant. Within the large cluster on the right hand side, word length tends to occupy the rightmost area, and the behavioural measures fall closer to the centre, although there is some overlap between these two regions. The separation of word length measures from word frequency measures is interesting; it apparently is inconsistent with Zipf's Law. On the other hand, the two sets of measures have similar values on vector 2;

---

<sup>40</sup> Appendix 1 gives the values of the two vectors for each measure and the numbers of the tables from which they were taken.

although this is less important than vector 1, it indicates that there is some partial relationship between the two sets of measures. The major vector seems to be characterized by the difference between frequency in texts and word length, and the minor vector something to do with the difference between the number of first places in lists and the remaining measures.

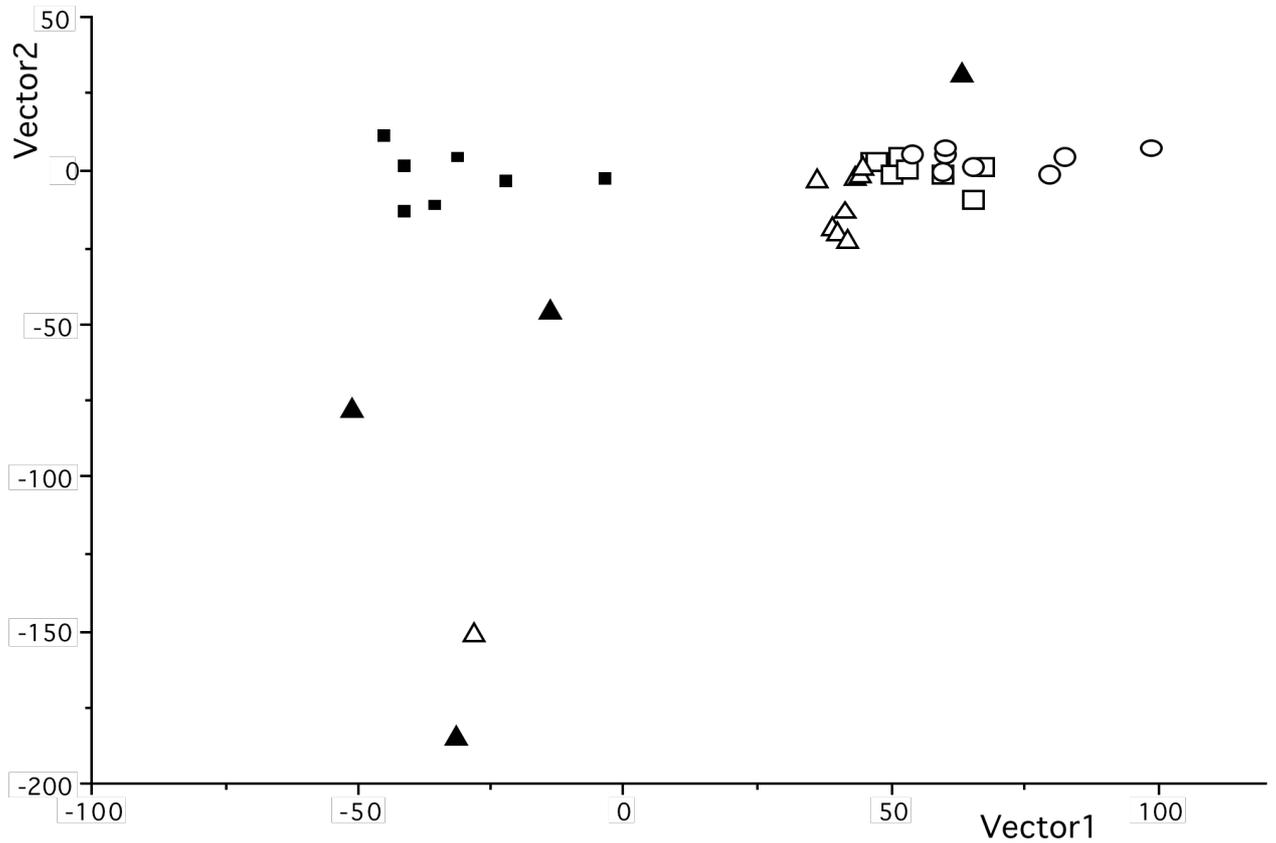


Figure1

Correspondence Analysis for each measure and each language

Behavioural measures

□ = RT, frequency and consistency

△ = Lists - frequency and mean position

▲ = Lists - number of first places

Linguistic measures

■ = Frequency in texts and no. of derived forms

○ = Word length

Figure 2 shows the colour terms plotted in the same space as the measures are in Figure 1. There is a degree of association between the pattern of terms in the space, and the hierarchy; the six primary basic terms form an ordered group, varying primarily on the first vector. At

the other extreme, vector 1 has three of the secondary basic terms, brown, purple and orange. The remaining two secondary terms, are distant from the other three, and differ from the primary terms on vector 2.

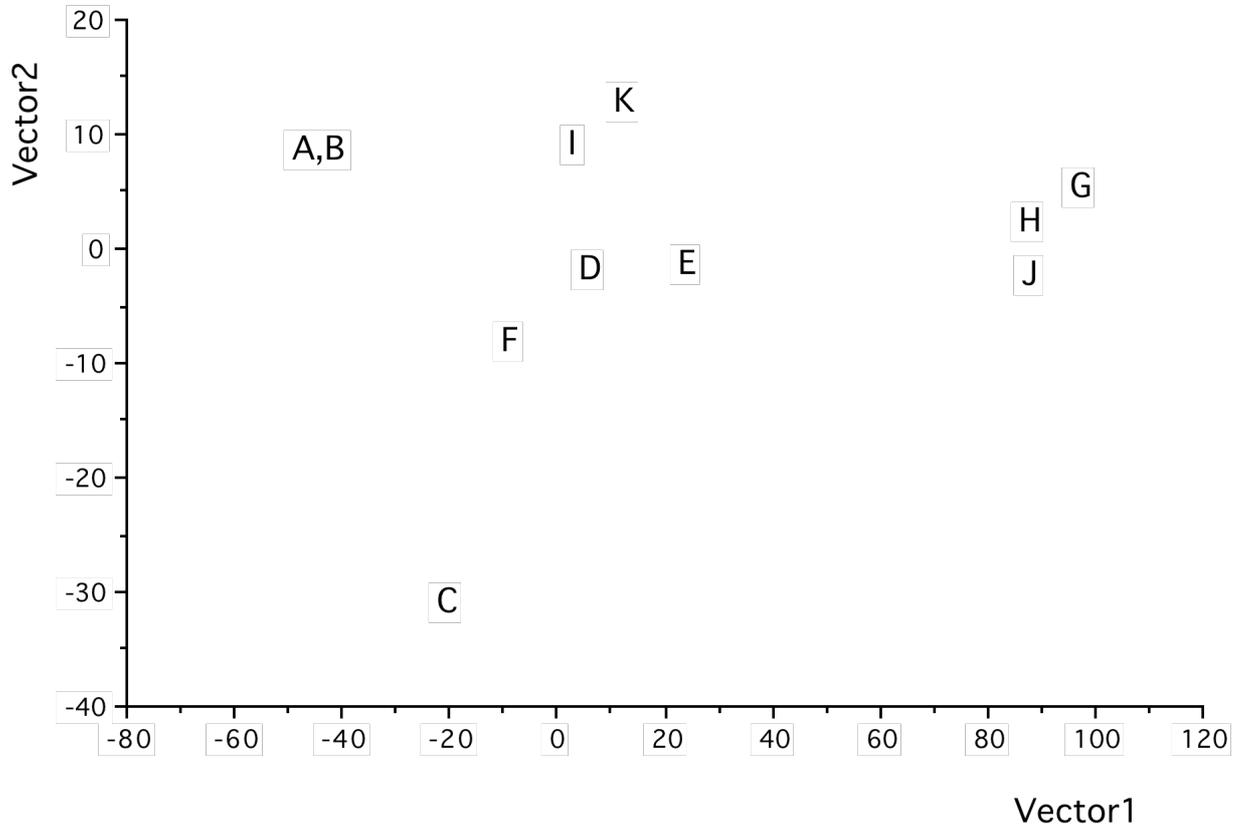


Figure 2

### Correspondence Analysis of the 11 basic terms

- |           |           |
|-----------|-----------|
| A: White  | G: Brown  |
| B: Black  | H: Purple |
| C: Red    | I: Pink   |
| D: Green  | J: Orange |
| E: Yellow | K: Grey   |
| F: Blue   |           |

Could either of the two vectors be a good measure of the hierarchy? Vector 2 is clearly poorly related in that purple and orange have similar values to white, black and red. The order on vector 1 is white, black,

red, blue, pink, green, grey, yellow, purple, orange and brown. Given the wild card status of grey, the primary deviation from the ranks of the hierarchy is the relatively high position of pink. The correlation with the hierarchy is  $t = .51$ ;  $p < .05$ , but although this is a reasonably strong value, several of the correlations of the single measures are greater.

## **7. Combinations of measures within languages**

We saw in the previous section that there were strong relationships between measures across languages. In this section we consider whether, within a language, the measures might be combined to provide a better measure of basicness than any measure on its own. We do this for Russian, English (British and American data) and French; it is only for these three that we have a reasonable spread of the measures (thus the number of combinations is lower than in Appendix 2).

	RT	Frequency of occurrence	Consistency	List frequency	List first places	Derived terms	Frequency in texts	Number of phonemes	Number of syllables	Total hurdles	Mean rank	Rank of weighted mean
<i>belyj</i> 'white'	+	-	-	+	+	+	+	-	-	5	5.1	1
	9	10.5	10	6.5	2	1	2	1.5	3			
<i>černyj</i> 'black'	+	-	-	+	+	+	+	-	-	5	5.2	2
	7	10.5	11	2.5	5.5	2	1	4	3			
<i>krasnyj</i> 'red'	+	-	-	+	+	+	+	-	-	5	5	3
	6	9	9	1	1	6	3	7	3			
<i>zelenyj</i> 'green'	+	+	+	+	+	+	+	-	-	7	3.8	4
	4	1	1	2.5	4	3	4	7	7.5			
<i>želtyj</i> 'yellow'	+	-	+	+	-	+	+	-	-	5	5.2	6
	1	6	3	8.5	9	5	7	4	3			
'blue'	+	+	+	+	+	+	+	-	-	7	4.0	5
	2	3	4	5	3	4	5	4	6			
<i>koričnevyyj</i> 'brown'	+	-	+	+	-	-	-	-	-	3	7.9	9
	5	7.5	7.5	4	9	10.5	9	9.5	9.5			
<i>fioletovyyj</i> 'purple'	+	+	+	+	-	-	-	-	-	4	8.1	10
	8	2	5.5	8.5	9	9	10	9.5	11			
<i>rozovyyj</i> 'pink'	+	+	+	+	+	-	-	-	-	5	8.2	8
	11	4	7.5	11	5.5	8	8	11	7.5			
<i>oranževyyj</i> 'orange'	-	-	+	+	-	-	-	-	-	3	8.1	11
	10	7.5	2	6.5	9	10.5	11	7	9.5			
<i>seryj</i> 'grey'	+	+	+	+	-	+	+	-	-	6	5.6	7
	3	5	5.5	10	9	7	6	1.5	3			

Table 5: Summary of all tests for Russian terms

	RT	Frequency of occurrence	Consistency	List frequency	List first places	Frequency in texts	Number of phonemes	Total hurdles	Mean rank	Rank of weighted mean
<i>white</i>	10	10	- 11	+ 1.5	- 6	+ 1	- 2.5	2	6.7	1
<i>black</i>	?	11	+ 8	+ 6.5	- 6	+ 2	- 8	3	6.9	3
<i>red</i>	3	8	+ 9	+ 4.5	+ 1	+ 3	- 2.5	4	4.4	2
<i>green</i>	4	1	+ 1	+ 1.5	+ 3.5	+ 4	- 8	4	3.3	6
<i>yellow</i>	1	6	+ 5	+ 4.5	+ 3.5	+ 8	- 8	4	5.1	8
<i>blue</i>	2	2	+ 2	+ 1.5	+ 2	+ 5	- 2.5	4	2.4	5
<i>brown</i>	5	7	+ 6	+ 10.5	- 9.5	+ 7	- 8	3	7.6	7
<i>purple</i>	7	3	+ 4	+ 8	- 6	- 10	- 8	2	6.6	10
<i>pink</i>	9	4.5	+ 7	+ 10.5	- 9.5	+ 9	- 8	3	8.2	9
<i>orange</i>	8	4.5	+ 3	+ 3	- 9.5	- 11	- 8	2	6.7	11
<i>grey</i>	6	9	- 10	+ 9	- 9.5	+ 6	- 2.5	2	7.4	4

Table 6: Summary of all tests for English terms

	List frequency	List first places	Derived terms	Frequency in texts	Number of phonemes	Number of syllables	Total hurdles	Mean rank	Rank of weighted mean
<i>blanc</i> 'white'	+ 7	+ 3	+ 1.5	2	- 3.5	- 4.5	3	3.4	1.5
<i>noir</i> 'black'	+ 1	+ 7	- 6	1	- 9	- 4.5	2	4.8	1.5
<i>rouge</i> 'red'	+ 4	+ 2	- 5	3	- 3.5	- 4.5	2	3.5	3
<i>vert</i> 'green'	+ 4	+ 6	+ 1.5	6	- 3.5	- 4.5	3	4.3	6
<i>jaune</i> 'yellow'	+ 2	+ 4.5	- 3	8	- 3.5	- 4.5	2	4.3	8
<i>bleu</i> 'blue'	+ 4	+ 1	- 7	4	- 3.5	- 4.5	2	4.0	4.5
<i>marron</i> 'brown'	+ 10	- 10.5	- 11	11	- 9	- 10	1	10.3	11
<i>violet</i> 'purple'	+ 9	- 10.5	- 9	10	- 11	- 10	1	9.9	10
<i>rose</i> 'pink'	+ 7	+ 4.5	- 8	5	- 3.5	- 4.5	2	5.3	4.5
<i>orange</i> 'orange'	+ 7	+ 8	- 10	9	- 9	- 10	2	8.8	9
<i>gris</i> 'grey'	+ 11	+ 9	- 4	7	- 3.5	- 4.5	1	6.5	7

Table 7: Summary of all tests for French terms

Tables 18, 19, 20, summarize the performance of each term on each test in two ways. First, whether on a given measure that term scores better than the best non-basic term. If it passes this test, it is shown as +; if it fails as -; and if the data is not available the column is left blank. Second, we give the rank within the eleven basic terms, for each basic term on the measures for which we have data. The pass-fail data, and the ranks are all taken from the tables in earlier sections. In the case of the English, we have used the first American sample for RT (Table 3), and for frequency of occurrence (Table 4), as there was least missing data for this sample; but we have used the second American sample for the consistency measure, as the data for the highest non-basic terms was available. We have not used the syllable version of word length for the English sample because there was so little variation in the scores.

Finally, we combine the data within each language in three ways:

- (a) the hurdles model: here we simply add the number of tests for which a given term scores better than the strongest non-basic term. That is, we sum the number of + results in each row, and the total is shown in the column labelled 'hurdles model'.
- (b) the mean rank: the mean of the ranks shown in a row is calculated and is shown in the next column.
- (c) the weighted mean: this score is derived from a separate correspondence analysis for each of the three languages; this is described more fully later (see also Davies et al. 1992).

### The Hurdles Model

There is a relatively weak trend for the primary basic terms to 'jump' more hurdles than the secondary basic terms, but there are exceptions in each language. In Russian, seryj 'grey' does better than černyj 'black', belyj 'white', želtyj 'yellow' and krasnyj 'red'; in English, pink and brown do better than white, and in French, violet 'purple', rose 'pink' and orange 'orange' do better than bleu 'blue'. None of the

correlations with the hierarchy is significant (tau ranges from -.18 to -.37).

### Mean Rank

In this case, the separation of primary basics from secondary basics is complete for Russian and French. It is almost complete for English: the only exceptions are purple, which has a marginally higher rank than black and white, and orange, which scores the same as white and is marginally higher than black. The correlations with the hierarchy are significant for Russian, tau = .52 p <.02, and for French, tau = .54 p <.02, but not for English. However, these correlations are mostly attributable to the difference between primary and secondary basics.

### Weighted Mean

For each language, the original test data that appeared in tables in earlier sections, and is summarized in Tables 18-20, was subjected to correspondence analyses. In each case, there was one major vector which accounted for a substantial proportion of the inertia. The minimum was 78%. We have used the scores for each colour term on this vector as weighted scores in that they reflect the contributions of each measure in accord with how much each measure contributes to that vector score. We have then ranked the vector scores and these are shown in the final columns of the tables.

On this measure, the primary basics are well separated from the secondary basics in Russian and English; the only exception is that grey in English does better than several primary basic terms. The separation in French is partial in that jaune 'yellow' does worse than rose 'pink' and gris 'grey', and bleu 'blue' is equal to rose 'pink'. The correlations with the hierarchy are significant for all three languages: tau = .81, p < .0009 for Russian; tau = .56 p <.02 for English and French.

In summary, the mean rank and the weighted rank do well at separating the primary basics from the secondary basics; better in general than any measure on its own. However, although the correlations with the hierarchy are generally significant, they are of the same order as for the best single measure; the weighted rank correlates  $\tau = .81$  for Russian compared to  $\tau = .77$  for frequency in tests, which is the best single measure for Russian; and  $\tau = .57$  for French compared to  $\tau = .56$  for frequency in texts, the best measure for French.

## 8. Discussion

Several clear patterns emerge from the data we have reviewed. First, there is clear consistency over languages. By and large the efficacy and range of a test is independent of the language it is applied to. This contention is supported by the pattern of intercorrelation between the same measure over different languages reported at the end of each 'test' section (4.1 to 4.4 and 5.1 to 5.3); the correlations are generally large and positive. And also by the correspondence analysis (Figure 1), where the uses of the same measure across languages tend to cluster together.

Second, there are broadly just three superordinate groups of measures with perhaps a fourth less clear-cut group. This is seen most easily in the correspondence analysis (Figure 1). The groups are:

- (1) Frequency in texts and number of derived forms.
- (2) Word length
- (3) The behavioural measures - RT, frequency of naming, consistency of naming, and frequency in elicited lists (but not the number of first places in elicited lists).

The possible fourth group is the number of first places in elicited lists. This group is rather more diffuse than the others; but it is clear

that this measure is very different to frequency in elicited lists and the other behavioural measures.

The three groups fall on both sides of the linguistic - behavioural distinction that we drew in the introduction and the results section. The first two are linguistic, but are also very different to each other in terms of what they appear to be measuring. They are maximally different on vector 1 in the correspondence analysis. The third group consists of all the behavioural measures, except the number of first places on elicited lists, and is closer to word length than to frequency in texts.

This superordinate structure can also be seen when we consider what each measure achieves. We have summarised the main patterns in the results in Table 21. There, for each measure and language, we have considered whether a measure achieves each of our three measurement criteria. That is, if a measure helps distinguish basic from non-basic terms, this is indicated by a plus in column one; if it fails the level 1 criterion it receives a minus. If no data is available then we indicate this with a question mark. Similarly, if a measure helps distinguish primary from derived terms it receives a plus in column 2, otherwise a minus. Finally if it helps discriminate the ranking within the basic terms over and above the level 2 criterion it receives a plus, otherwise a minus. We have adopted this relatively coarse pass-or-fail procedure in order to summarise a large set of results succinctly, and to help reveal general patterns.

The rules we have adopted for passing and failing the three criteria are as follows:

(1) Level 1 if no more than one non basic term ranks higher than some basic terms or no more than two non basic terms rank higher than just one basic term, then that measure passes the level one criterion.

(2) Level 2. If either no more than one secondary basic term is higher than some primary basic terms, or the difference between the mean scores is significant at more than the .025 level (that is,  $p < .025$ ), then it passes the level 2 criterion.

(3) Level 3. If the correlation is (a) statistically significant at  $< .025$ , and (b) this is not just due to the level 2 criterion, then it passes the level 3 criterion.

Measure	Language	Criterion		
		Level 1 Basic/ non basic	Level 2 Primary/ secondary	Level 3 Correlation with Hierarchy
<b>BEHAVIOURAL</b>				
RT	Russian	+	-	-
	Japanese	+	-	-
Frequency of occurrence:	Russian	-	-	-
	Am. Eng.	?	-	-
Consistency of naming:	Russian	-	-	-
	Japanese	-	-	-
	Am. Eng. 1	?	-	-
	Am. Eng. 2	-	-	-
Elicited lists:				
a) Frequency:	Russian (5 mins)	+	-	-
	Am. Eng.	+	+	-
	Brit. Eng.	+	+	-
	French (5 mins)	+	+	-
b) First Places	Russian	+	+	-
	Am. Eng.	+	+	-
	Brit. Eng.	-	+	-
	French (5 mins)	+	-	-
<b>LINGUISTIC</b>				
Number of derived forms:	Russian	-	+	+
	French	-	+	-
Frequency in texts:	Russian	-	+	+
	English	?	+	+
	French	?	+	+
	Hebrew	?	+	+
	Spanish	?	+	+
Word length:				
a) Phonemes:	Russian	?	+	-
	English	?	-	-
	French	?	-	-
	Spanish	?	-	-
b) Syllables:	Russian	?	+	-
	English	?	-	-
	French	?	+	-
	Spanish	?	-	-

Table 21: Assessment of each test according to the three criteria

Table 21 shows that only measures from groups 1 and 3 identified by the correspondence analysis are useful for establishing basicness or

making discriminations within basic terms. Group 1 measures, that is, frequency in texts and numbers of derived forms, generally pass the level 2 and 3 criteria. That is they are useful for discriminating within the basic terms. But their level 1 status is less clear. There are three cases for which we have data (for non-basic as well as basic terms). These are the number of derived forms for Russian and French, and frequency in texts for Russian. All three instances fail the level 1 criterion; that is they do not reliably separate the basic and non basic terms. It would be useful to see if frequency in texts did serve this function in other languages.

By and large, the elicited list measure succeeds at both levels 1 and 2 but fails level 3. That is it is effective in establishing the basic terms, and further it can differentiate between the primary and secondary terms, but it does not correlate with the Berlin and Kay hierarchy. Of the rest of the behavioural measures, only RT appears at all useful, and then only for distinguishing basic and non basic terms. Finally word length in general appears to have only limited use. In some languages, Russian and French, the primary and secondary terms do differ on one or both measures of word length, but none of the measures correlate with the hierarchy.

What Table 21 implies for future work is that the elicited list procedure, and frequency in texts are the two most useful measures. Further, there is some suggestion that they complement each other. Only the elicited list measure reliably distinguishes basic from non basic terms; and only frequency in texts discriminates within the basic term inventory and correlates with the hierarchy. The list procedure has the further advantage that it is quick to perform, whether respondents write down the forms themselves or give oral responses. The frequency in text measure is more laborious but given a corpus of texts, spoken or written, there are various software packages which will produce the required concordance information.

## APPENDIX 1

<u>Source</u>	<u>Measure</u>	<u>Table</u>	<u>Vector1</u>	<u>Vector2</u>
Russian	RT	3	46.17	3.25
Japanese	RT	3	47.16	3.32
Russian	Frequency of naming	4	67.12	0.88
American 1	Frequency of naming	4	65.33	-9.55
Russian	Consistency	5	59.53	-1.39
Japanese	Consistency	5	51.40	4.50
American 1	Consistency	5	50.00	-1.24
American 2	Consistency	5	52.91	0.40
Russian	Frequency in lists (1 min)	6	41.69	-12.90
Russian	No. first places in lists	6	-51.22	-78.13
Russian	Frequency in lists (5 min)	6	44.58	-1.01
Illinois	Frequency in lists	7	42.17	-22.46
Maryland	Frequency in lists	7	39.23	-18.89
Ill. + Maryland	Frequency in lists	7	40.34	-20.23
Ill. + Maryland	No. first places in lists	7	-30.97	-184.07
Brit. English	Frequency in lists	8	43.43	-2.04
Brit. English	No. first places in lists	8	63.67	31.74
Brit. English	Mean position in lists	8	-27.56	-151.14
French	Frequency in lists	9	36.36	-2.71
French	No. first places in lists	9	-13.49	-45.55
French	Frequency in lists (5 min)	9	44.77	0.91
Russian	No. of derived forms	10	-31.26	4.27
French	No. of derived forms	11	-03.55	-2.52
Russian	Frequency in texts	12	-40.87	-13.14
English	Frequency in texts	13	-22.16	-2.93
French	Frequency in texts	14	-40.90	1.89
Hebrew	Frequency in texts	15	-35.19	-11.28
Spanish	Frequency in texts	16	-45.12	11.30
English	No. of phonemes	17	54.09	5.37
English	No. of syllables	17	59.80	-0.63
French	No. of phonemes	17	60.08	4.92
French	No. of syllables	17	82.35	4.61
Russian	No. of phonemes	17	79.50	-1.33
Russian	No. of syllables	17	98.67	7.74
Spanish	No. of phonemes	17	60.13	7.64
Spanish	No. of syllables	17	65.15	1.27

Correspondence Analysis: Co-ordinates for each measure and source tables

## References

- Alimpieva, R. V. 1982a. Slova sinij, goluboj v proizvedenijax S. Esenina i ix èkvivalenty v pol'skix perevodax. Materialy po rusko-slavjanskomu jazykoznaniju, 8-14. Voronež.
- Alimpieva, R. V. 1982b. Stanovlenie leksiko-semantičeskix grupp cvetovyx prilagatel'nyx v russkom jazyke pervoj poloviny XIX v. Voprosy semantiki: Issledovanija po istoričeskoj semantike, 49-60. Kaliningrad.
- Battig, William F. and Montague, William E. 1969. Category norms for verbal items in 56 categories: a replication and extension of the Connecticut category norms. Journal of Experimental Psychology: Monograph, vol. 80, no. 3, part 2
- Baxilina, N. B. 1975. Istorija cvetooboznačenij v russkom jazyke. Moscow: Nauka.
- Berlin, Brent and Paul Kay 1969. Basic color terms: their universality and evolution. Baltimore: University of California Press.
- Bolton, Ralph. 1978. "Black, white, and red all over: the riddle of color term salience." Ethnology 17.287-311.
- Bolton, Ralph, Curtis, Anne T. and Thomas, Lynn L. 1980. "Nepali color terms: salience on a listing task." Journal of the Steward Anthropological Society 12.309-22.
- Boynton, Robert M. and Olson, Conrad. X. 1987. "Locating basic colors in the OSA space." Colour Research and Application 12.94-105.
- Boynton, Robert M. and Olson, Conrad X. 1990. "Salience of chromatic basic color terms confirmed by three measures." Vision Research 30 no 9:1311-17.
- Buchanan, M. A. 1927. A Graded Spanish Word Book. Toronto: University of Toronto Press. [Quoted from Hays et al. 1972:1121.]
- Corbett, Greville G. and Morgan, Gerry. 1988. "Colour terms in Russian: reflections of typological constraints in a single language." Journal of Linguistics 24.31-64.

- Cromer, Richard F. 1991. Language and Thought in Normal and Handicapped Children. Oxford: Blackwell.
- Davies, I. R. L., MacDermid, C., Corbett, G. G., McGurk, H., Jerrett, H., Jerrett, T. and Sowden, P. 1992. "Colour terms in Setswana: A linguistic and perceptual approach." Linguistics 30.1065-1103.
- Dixon, R. M. W. 1982. Where have all the adjectives gone?: and other essays in semantics and syntax. Berlin: Mouton.
- Durbin, Marshall 1972. "Basic terms - off-color?" Semiotica 6.257-78.
- Forbes, Isabel 1976. Structural Semantics with Particular Reference to the Colour Vocabulary of Modern Standard French. Unpublished PhD thesis, University of Edinburgh.
- Forbes, Isabel 1979. "The terms brun and marron in modern standard French." Journal of Linguistics 15.295-305.
- Forbes, Isabel 1986. "Variation and change in the basic colour vocabulary of French." SIGMA 10.81-103. (Publication du C.E.L.A.M., Université de Provence, Aix-en-Provence & Université Paul Valéry, Montpellier)
- Foss, Carl E., Nickerson D. and Granville, Walter C. 1944. "Analysis of the OSTWALD color system." Journal of the Optical Society of America 34 no. 7: 361-81.
- Greenacre, M. J. 1984. Theory and Application of Correspondence Analysis. London: Academic Press.
- Harkness, Sara 1973. "Universal aspects of learning color codes: a study in two cultures." Ethos 1.175-200.
- Hays, D. G., Margolis, E., Naroll, R. and Perkins, D. R. 1972. "Color term salience." American Anthropologist 74.1107-1121.
- Jernudd, Björn H. and White, Geoffrey M. 1983. "The concept of basic color terms: variability in For and Arabic." Anthropological Linguistics 25.61-81.
- Johansson, S. and Hofland, K. 1989. Frequency Analysis of English Vocabulary and Grammar: based on the LOB Corpus: Vol. I: Tag Frequencies and Word Frequencies. Oxford: Clarendon.

- Kay, Paul 1975. "Synchronic variability and diachronic change in basic color terms." Language in Society 4.257-70.
- Kay, Paul, Berlin, Brent and Merrifield, William 1991. "Biocultural implications of systems of color naming." Journal of Linguistic Anthropology 1.12-25
- Kay, Paul and McDaniel, Chad K. 1978. "The linguistic significance of the meanings of basic color terms." Language 54. 610-46.
- Kikuchi, Atsuko and Lichtenberk, Frantisek 1983. "Semantic extension in the colour lexicon." Studies in Language 7.25-64.
- Kulick, D. and Vamling, K. 1984. Ryska. In T. Pettersson (ed.) Färgterminologi: Seminarieuppsatser i allmän språkvetenskap (=Praktisk lingvistik 9), 79-109. Lund. [Quoted from Vamling (1986).]
- Lacroix, J.-C. 1978. Notes sur l'emploi des terms de couleur en français. In S. Tornay (ed.), Voir et nommer les couleurs. Nanterre: Laboratoire d'ethnologie et de sociologie comparative. 181-206.
- Laws, Glynis and Davies, Ian. Forthcoming. The Stroop effect, linguistic basicness and colour focality.
- MacLaury, Robert E. 1987. Co-extensive semantic ranges: different names for distinct variants of one category. In Barbara Need, Eric Schiller and Anna Bosch (eds) CLS 23: Papers from the 23rd Annual Regional Meeting of the Chicago Linguistic Society: I: The General Session, 268-82. Chicago: Chicago Linguistic Society.
- MacLaury, Robert E. 1991a. "Exotic color categories: linguistic relativity to what extent?" Journal of Linguistic Anthropology 1.26-51.
- MacLaury, Robert E. 1991b. "Social and cognitive motivations of change: measuring variability in color semantics." Language 67.34-62.
- MacLaury, Robert E. 1992. "From brightness to hue: an explanatory model of color-category evolution." Current Anthropology 33 (2).137-186.
- Miller, George A. and Johnson-Laird, Phillip N. 1976. Language and

- Perception. Cambridge, Mass.: Harvard University Press.
- Morgan, Gerry 1993. "Basic colour terms: comparative results for French and Russian." Journal of French Language Studies 3.1-17.
- Morgan, Gerry and Corbett, Greville G. 1989. "Russian colour term salience." Russian Linguistics 13.125-141.
- Morgan, Gerry and Moss, A. E. St. G. 1988/89. "The two blues of Russian: the referents of sinij and goluboj." Die Farbe 35/36.353-7.
- Moskovič, V. A. 1969. Statistika i semantika: Opyt statističeskogo analiza semantičeskogo polja. Moscow: Nauka.
- Moss, A. E. 1989. Basic colour terms: problems and hypotheses Lingua 78.313-320.
- Moss, A. E., Davies, I. R. L., Corbett, G. G. and Laws, G. 1990. "Mapping Russian basic colour terms using behavioural measures." Lingua 82: 313-32.
- Pollnac, Richard B. 1975. "Intra-cultural variability in the structure of the subjective color lexicon in Buganda." American Ethnologist 2.89-109.
- Priestly, Tom M. S. 1981-83. "On basic color terms in early Slavic and Ukrainian." The Annals of the Ukrainian Academy of Arts and Sciences in the United States 15.243-51.
- Rosch Heider, Eleanor 1972. "Universals in color naming and memory." Journal of Experimental Psychology 93.10-20.
- Saunders, B. A. C. and van Brakel, J. 1988. "Re-evaluating basic colour terms." Cultural Dynamics (Leiden) 1.359-78.
- Senft, Gunther 1987. "Kilivila colour terms." Studies in Language 11.313-46.
- Smith, C. N. 1965. Student Handbook of Color. New York: Reinhold.
- Uchikawa, K. and Boynton, R. M. 1987. "Categorical color perception of Japanese observers: comparisons with that of Americans." Vision Research 27 (10).1825-33.
- Vakar, N. P. 1966. A Word Count of Spoken Russian: the Soviet Usage. Ohio: Ohio State University Press.

- Vamling, Karina 1986. "A note on Russian blues." Slavica Lundensia 10.225-33.
- von Wattenwyl, André and Zollinger, Heinrich 1979. "Color-term salience and neurophysiology of color vision." American Anthropologist 81.279-88.
- Weller, Susan C. and Romney, A. Kimball 1990. Metric Scaling: Correspondence Analysis Newbury Park, CA: Sage. (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-075).
- Wierzbicka, Anna 1990. "The meaning of color terms: semantics, culture, and cognition." Cognitive Linguistics 1.99-150.
- Worth, D. S., Kozak, A. S. and Johnson, D. B. 1970. Russian Derivational Dictionary. New York: American Elsevier.
- Zasorina, L. N. (ed.) 1977. Častotnyj slovar' russkogo jazyka: okolo 40 000 slov. Moscow: Russkij jazyk.