Memory & Cognition
2004, 32 (1), 83-95

Linear separability in superordinate natural language concepts

WIM RUTS and GERT STORMS
University of Leuven, Leuven, Belgium

and

JAMES HAMPTON
City University, London, England

Two experiments are reported in which linear separability was investigated in superordinate natural language concept pairs (e.g., *toiletry–sewing gear*). Representations of the exemplars of semantically related concept pairs were derived in two to five dimensions using multidimensional scaling (MDS) of similarities based on possession of the concept features. Next, category membership, obtained from an exemplar generation study (in Experiment 1) and from a forced-choice classification task (in Experiment 2) was predicted from the coordinates of the MDS representation using log linear analysis. The results showed that all *natural kind* concept pairs were perfectly linearly separable, whereas *artifact* concept pairs showed several violations. Clear linear separability of natural language concept pairs is in line with independent cue models. The violations in the artifact pairs, however, yield clear evidence against the independent cue models.

Categories and concepts are used to organize our knowledge of the objects in the world around us (Estes, 1994). The structure of natural concepts has been the subject of many studies because it is such a fundamental issue in human cognition. A question of crucial importance, in this regard, concerns the constraints by which natural concepts are characterized. It would be hard to believe that human culture has passed on categories with a structure that is not coordinated with the constraints of human information processing (Medin & Schwanenflugel, 1981; Sperber, 2000).

The classical view states that semantic concepts can be described in terms of defining features that are singly necessary and jointly sufficient (e.g., Sutcliffe, 1993). Ample evidence has been provided against this view (for overviews see, e.g., Hampton, 1997; Komatsu, 1992; E. E. Smith & Medin, 1981). Two major classes of theories have been formulated on the basis of probabilistic models. First, exemplar models (Medin & Schaffer, 1978; Nosofsky, 1984, 1986) state that categories are represented by stored traces of particular exemplars that have been previously encountered. New items are assumed to be judged as an instance of a category to the extent that they are sufficiently similar to one or more of the instance representations stored in memory. Second, prototype models (Hampton, 1979, 1998; Rosch & Mervis, 1975) assume that a category is represented by an abstract summary representation. More specifically, a prototype is defined by a set of characteristic features that all carry more or less weight in the definition of the concept. In this view, categorization is based on whether an item possesses enough of the characteristic features.

Classifying stimuli on the basis of similarity to prototypes involves a summing of evidence (e.g., matching characteristic features) against some criterion. Stimuli are accepted as members of a category if the summed evidence exceeds the criterion; otherwise, they are rejected (Medin & Schwanenflugel, 1981). Categories defined in this way fulfill the constraint of linear separability (Sebestyen, 1962). That is, for a category considered on its own, a linear function of attributes exists that perfectly separates members from nonmembers. Similarly, for any pair of categories within a given domain, there exists a linear function, defined over the set of attributes for the domain, that perfectly separates the two categories. In the case of vague category boundaries, as are commonly found in many natural concepts, the linear separability constraint would simply be that there exists a linear function of attributes that has a monotonic relation with the relative degree of category membership in each class. Gärdenfors (2000) has indeed proposed that a criterion for defining a *natural property* is that it should form a convex region of a domain in a conceptual space, thus obeying this constraint.

Prototype models are not the only models that incorporate the assumption of linear category separability. Related models, like the average distance model (Reed, 1972), versions of cue validity and frequency models (Medin & Schaffer, 1978), and some versions of Ashby's decision-boundary model (Ashby & Maddox, 1992) also make the same assumption. The term *independent cue models* refers to a collection of models that all obey linear separability

Part of the reported research was sponsored by Grant G.0266.02 from the F.W.O.-Flanders. Correspondence should be addressed to W. Ruts, Psychology Department, University of Leuven, Tiensestraat 102, B-3000 Leuven, Belgium (e-mail: wim.ruts@psy.kuleuven.ac.be).

(Franks & Bransford, 1971; Hayes-Roth & Hayes-Roth, 1977). Most exemplar models, like Medin and Schaffer's (1978) context model or Nosofsky's (1984, 1986) generalized version of that model, do not assume that linear separability constrains category representation, because no role is played in the models by any measure of proximity to the center of the category. These models are called *relational coding models*. Consequently, finding out whether linear separability constrains natural categories brings evidence to bear on the advantages of independent cue models versus exemplar models. More specifically, should it turn out that natural categories do *not* obey the constraint of linear separability, then one would have good reason to prefer relational coding models. In such an event, there would also be reason to consider more complex theory-based representations (e.g., Rips, 1989) for concepts, in which similarity per se is not the key factor in determining category membership.

Because linear separability is such an important constraint in formal models of categorization, different studies have investigated linear separability in category learning experiments, in which participants learned new categories of artificial stimuli. Most of these studies failed to yield evidence that linearly separable categories can be learned more rapidly than categories that are not linearly separable (Medin & Schaffer, 1978; Medin & Schwanenflugel, 1981; Wattenmaker, Dewey, Murphy, & Medin, 1986). Recently, however, J. D. Smith, Murray, and Minda (1997) and Blair and Homa (2001) reported results in favor of independent cue models when they used better differentiated categories with many exemplars and in experiments where participants had to classify stimuli in four categories (instead of two, as in most category learning experiments). In conclusion, the results from category learning experiments in which linear separability has been manipulated do not unanimously favor or argue against independent cue models, although there is evidence that in at least some conditions it is easier to learn linearly discriminable categories.

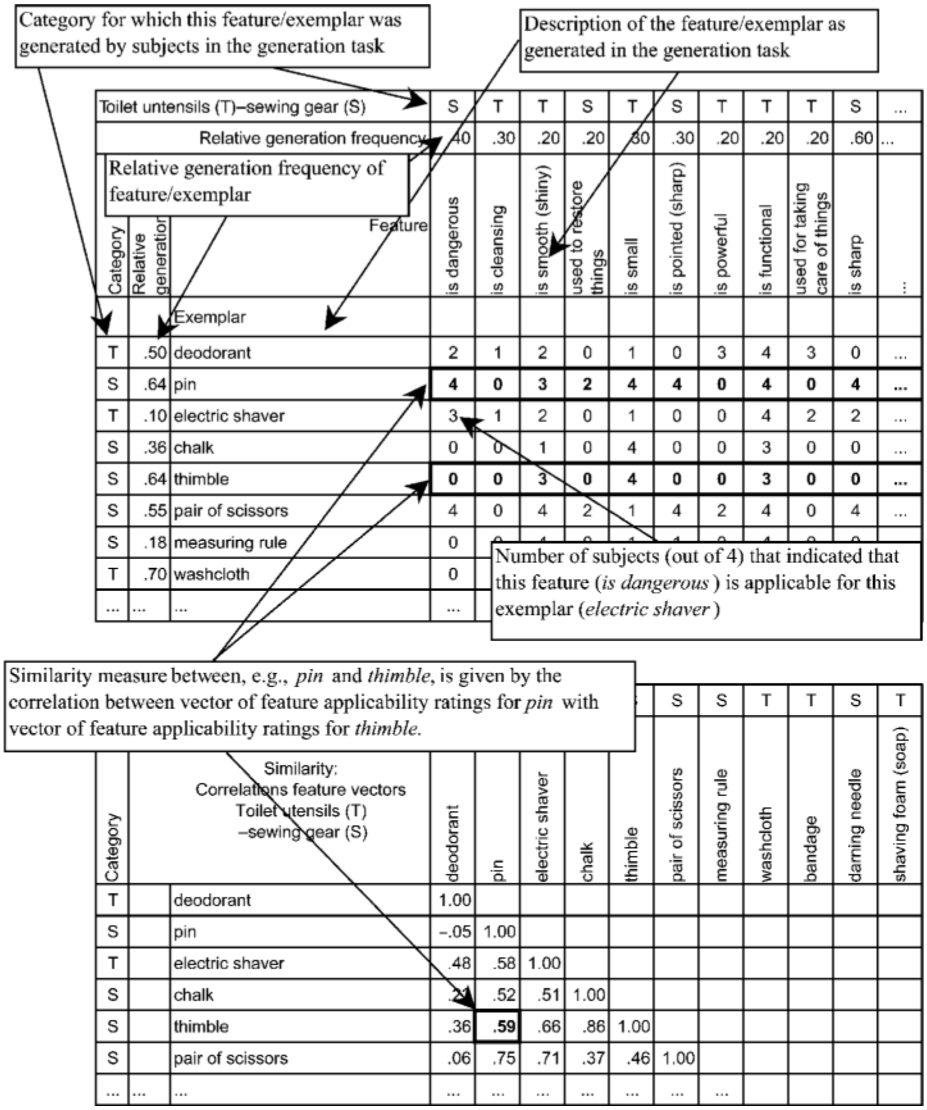
Given the importance of linear separability in formal models of classification learning, as well as the importance of testing the relevance of such models for the representation of natural concepts in semantic memory, it is surprising how little attention has been paid to the question of whether natural language concepts are themselves linearly separable. The lack of such studies may perhaps be owing to the difficulties associated with the selection of the attributes or information that is to be taken into account in the evaluation of linear separability in this context. A rare exception to this lack of interest can be found in Malt, Sloman, Gennari, Shi, and Wang (1999). In their study, artifact concepts of containers (such as can, bottle, jar, etc.) were studied in different languages. The authors showed that pictured stimuli that were labeled with the same word were not linearly separable on the basis of proximity in a two-dimensional similarity space (measured in different ways). In particular, whereas differences in similarity relations between languages were small, the differences in naming categories between languages were very large. However, because the study restricted itself to names for basic-level artifacts, and only solutions in two dimensions were investigated, further exploration of the issue is clearly called for.

In this article, we focus on the extent to which natural language superordinate categories are linearly separable. By *natural language category*, we mean the grouping of objects, and object classes, that people consider a lexical term or commonly understood lexicalized phrase. Because the central question is whether there exists a linear function that perfectly separates two such categories, semantically related contrast pairs of superordinate concepts (e.g., *toiletry–sewing gear*) were used. Testing linear separability of unrelated categories (e.g., *vehicles–sewing gear*) would be pointless because the possibility of two categories not being discriminable will arise only when they share the same semantic domain and are close to each other within that domain. Following up on the study of Malt et al. (1999), we also concentrate on the possible difference between superordinate *natural kind* (e.g., *fish–mammals*) concept pairs and superordinate *artifact concept* pairs (e.g., *toiletry–sewing gear*) regarding the linear separability issue. (For convenience, we use the term *natural kind* in a loose sense to refer to categories within the general domain of biological or living things, in contrast to the domain of man-made artifacts.) To address the problem of finding the appropriate similarity space in which to test for linear separability, we adopt a different approach from that of Malt et al. We compute proximities between category items on the basis of a large set of characteristic features of the categories themselves, based on a generation task. This approach has been successfully used before in several studies (e.g., Hampton, 1979; Storms, De Boeck, & Ruts, 2000; Storms, De Boeck, Van Mechelen, & Ruts, 1996; Verbeemen, Vanoverberghe, Storms, & Ruts, 2001) to predict a number of category-related measures. Three major tasks are used to gather the necessary data. First, category members of the contrasting categories are taken from an exemplar generation task. Second, category features are taken from a feature generation task, and third, similarities are calculated from feature profiles of the exemplars taken from a feature applicability rating task. The data gathered from these tasks are illustrated in Figure 1 and the tasks are explained in detail in the Method section. By submitting the resulting proximity matrix to multidimensional scaling (MDS), a formalization of the semantic relations between category exemplars is obtained based on the features that are considered to be relevant to the category itself. The question of whether the categories are linearly separable can then be answered by using log linear analysis to look for a linear function that perfectly separates the exemplars of the two categories.

EXPERIMENT 1

Method

Participants. Sixty participants between the ages of 18 and 61 years took part. Of these, 23 females and 26 males, participat-


Category for which this feature/exemplar was generated by subjects in the generation task
Description of the feature/exemplar as generated in the generation task
Relative generation frequency of feature/exemplar
Number of subjects (out of 4) that indicated that this feature (*is dangerous*) is applicable for this exemplar (*electric shaver*)
Similarity measure between, e.g., *pin* and *thimble*, is given by the correlation between vector of feature applicability ratings for *pin* with vector of feature applicability ratings for *thimble*.

		Toilet untensils (T)–sewing gear (S)	S	T	T	S	T	S	T	T	T	S	...
		Relative generation frequency	.40	.30	.20	.20	.30	.30	.20	.20	.20	.60	...
Category	Relative generation	Feature	is dangerous	is cleansing	is smooth (shiny)	used to restore things	is small	is pointed (sharp)	is powerful	is functional	used for taking care of things	is sharp	...
		Exemplar											
T	.50	deodorant	2	1	2	0	1	0	3	4	3	0	...
S	.64	pin	**4**	**0**	**3**	**2**	**4**	**4**	**0**	**4**	**0**	**4**	**...**
T	.10	electric shaver	3	1	2	0	1	0	0	4	2	2	...
S	.36	chalk	0	0	1	0	4	0	0	3	0	0	...
S	.64	thimble	**0**	**0**	**3**	**0**	**4**	**0**	**0**	**3**	**0**	**0**	**...**
S	.55	pair of scissors	4	0	4	2	1	4	2	4	0	4	...
S	.18	measuring rule	0										
T	.70	washcloth	0										
...	...	...	...										

								S	S	T	T	S	T
Category		Similarity: Correlations feature vectors Toilet utensils (T) –sewing gear (S)	deodorant	pin	electric shaver	chalk	thimble	pair of scissors	measuring rule	washcloth	bandage	darning needle	shaving foam (soap)
T		deodorant	1.00										
S		pin	–.05	1.00									
T		electric shaver	.48	.58	1.00								
S		chalk	[illegible]	.52	.51	1.00							
S		thimble	.36	**.59**	.66	.86	1.00						
S		pair of scissors	.06	.75	.71	.37	.46	1.00					
...	...	...	...	...	...	...	...	...	...				

Figure 1. Part of the feature applicability matrix and the exemplar intercorrelation matrix for the *toiletry–sewing gear* concept pair.

ing voluntarily, generated exemplars and features of the superordinate categories studied. Eleven other participants completed the item × feature applicability (matrix-filling) task. Of these, 4 were male research assistants who participated voluntarily and the remaining 7 were female psychology students at the University of Leuven who were each paid the equivalent of $10 US for their participation. For practical reasons, and in particular the length of some of the tasks, tasks could not be evenly divided among participants.

Materials. All materials in the experiments were collected and presented in Flemish. The closest English equivalents have been used throughout the present text. In this experiment, eight superordinate natural kind concepts and 10 superordinate artifact concepts were used. Because of the large effort involved in collecting and analyzing exemplar and feature norms, we used existing normative data for all natural kind and for one artifact concept (*vehicles*). These data were used in Storms et al. (2000) and in Verbeemen et al. (2001). For the remaining nine artifact concepts, new data were collected using identical procedures.

The natural kinds taken from the earlier research were *insects*, *fish*, *birds*, *mammals*, *trees*, *flowers*, *fruits*, and *vegetables*. To investigate the linear separability of category boundaries, these eight concepts were combined into the following eight semantically related contrast pairs: *insects–fish*, *insects–birds*, *insects–mammals*, *fish–birds*, *fish–mammals*, *birds–mammals*, *trees–flowers*, and *fruits–vegetables*. For the matrix-filling task, all exemplars generated in the generation studies from Storms et al. (2000) and from Verbeemen et al. (2001) were selected. Furthermore, all features that were generated by at least 2 (out of 10) participants from the feature generation studies described in the same articles were selected. After duplicates (i.e., exemplars and features that were selected for both concepts within a contrast pair) were removed, this procedure resulted in sets of 48–83 exemplars and 20–51 features per pair of concepts. Note that to provide a strong test of linear separability, it was important to have as complete a sample of category members as possible for each category. For the selection of features, however, a limit was set to make the completion of the applicability matrix possible within less than 120 min. All stimuli were single words or were familiar lexicalized compound noun phrases in Flemish (see Figure 1 for examples).

The 10 superordinate artifact concepts were arranged in the following contrasting pairs: *toiletry–sewing gear*, *kitchen utensils–tableware*, *cleaning utensils–gardening utensils*, *vehicles–construction machines*, and *clothing–accessories*. The exemplars and features of the artifact concept *vehicles* were taken from the Storms et al. (2000) study. For the 9 remaining artifact concepts, the same exemplar and feature generation tasks were conducted. All generated exemplars and the 15 most frequently generated features from each category were used in the matrix-filling task. Each artifact contrast pair had between 61 and 85 exemplars and 24 to 30 features. Note that the selection criterion for features was different from the criterion used for the natural kinds. Selecting all features that were generated by at least 2 (out of 10) participants from the feature generation task would have resulted in sets with more than 60 features. The resulting exemplar × feature matrix would have been too large for participants to complete in a reasonable time. As a check, we later reanalyzed the natural kind concept pairs using just the 15 most frequently generated features. The results and conclusions of the analyses were effectively the same.

Procedure. To generate exemplars and features for the nine new artifact concepts, participants were asked to write down 10 or more exemplars for one to three unrelated concept categories. The same participants were also asked to generate as many features as they could think of for one to three different unrelated concept categories. Thus each participant only ever saw one of the two concepts making up a contrast pair. For both tasks, participants received a sheet of paper with instructions and free space to write down a list. Instructions contained an example from a different category (*furniture*) of showing how the task had to be completed and explicitly directed participants not to generate word associations. Although no time limit was imposed, participants never needed more than 10 min for one generation task. This procedure has been successfully applied in previous studies such as Hampton (1979), Storms et al. (2000), and Storms, Ruts, and Vandenbroucke (1998). For each of the nine artifact concepts, 10 to 13 participants generated exemplars and features. None of the participants generated both features and exemplars for the same concept. None of the participants from the exemplar and feature generation task participated in the matrix-filling task. The presentation order of the tasks was randomized over participants.

Having collected full sets of exemplars and features for all contrast pairs of superordinates, participants in the exemplar × feature applicability task (matrix-filling task) were given a matrix where the rows were labeled with all exemplars of a superordinate concept pair, and the columns were labeled with the features of the same concept pair (Figure 1). All exemplars as well as all features were presented in a random order. Participants were asked to fill out all entries in the matrix with a 1 or a 0, to indicate whether a feature was considered present in the exemplar corresponding to the row of the entry. Completion of the applicability matrix took about 90 to 120 min. Two matrices (*vehicles–construction machines* and *clothing–accessories*) were completed by 8 participants (a mixture of students and research assistants). Since reliability was high and the task could test the limits of a student's motivation to respond with due attention, it was decided to use only 4 research assistants for the remaining matrices. Measured reliability of the data was still sufficiently high with this number of participants.

Results and Discussion

Matrices were summed over participants, resulting in a single exemplar × feature matrix for every concept pair. (Figure 1 shows part of the summed matrix for the *toiletry–sewing gear* concept pair.) The entries of these matrices are frequencies, corresponding to the number of participants that judged the corresponding feature (column) applicable to the corresponding exemplar (row). Thus, for every exemplar, a vector of feature applicability was obtained, which was then correlated with every other exemplar vector in the matrix, resulting in an intercorrelation matrix between all possible pairs of exemplars. This matrix then represented the similarity between all pairs of exemplars, calculated in terms of which features of the two superordinate concepts the pair of exemplars possessed (or did not possess) in common, and which features they did not.

The reliability of this intercorrelation matrix was estimated by repeating this procedure with the data from each half of the participants and then applying the Spearman–Brown formula. Reliabilities for all pairs were .90 or above, except for *toiletry–sewing gear* (.84). Given that (looking ahead) this last contrast was one of the better discriminated pairs in the subsequent analysis, the lower reliability is probably not a matter for concern. The difference in reliability between the matrices filled out by 4 participants and the ones filled out eight times was not significant.

The intercorrelation between the vectors of feature applicability ratings for every pair of exemplars is a proximity measure between the exemplars with regard to the features generated for the superordinates involved. Since this measure is but one of several possible proximity measures, we subjected the data to two other commonly used measures—Euclidean and city-block distances (see Nosofsky, 1984). The correlations between the matrices based on the three different proximity measures were all well above .90, except for the *vehicles–construction machines* pair, with a correlation of .78 between Euclidean and intercorrelation proximity measures. The high intercorrelations show that the choice of the underlying similarity measure is not crucial. In the remainder of this article, only the correlation-based similarity data were further analyzed.

Multidimensional scaling. To investigate the linear separability of the two contrasting categories in each pair, a geometric configuration of the exemplars was obtained from an MDS solution (SAS, nonmetric PROC MDS). Because some correlations had a negative value, which would be interpreted as a missing value in the MDS analysis, 1.0 was added to all correlations, yielding input similarities that could vary between 0.0 and 2.0. MDS analyses were conducted for two to five dimensions for every concept pair. As can be seen in Table 1, the stress values of the MDS solutions for all of the concept pairs were below .18 in two dimensions. (Stress 1 values provide a measure of goodness-of-fit for the scaling solution, with low values indicating good fit.) For the higher dimensional solutions, stress values were less than .12. In general, the stress values for natural kind categories were somewhat lower than those for the artifact categories. However, all values indicate that the corresponding solutions were good, especially considering the large number of stimuli scaled in a low dimensionality (Kruskal & Wish, 1978).

Table 1
Stress Values of the MDS Solutions for Contrasting Pairs in Two to Five Dimensions

	Number of Items in Category		Stress Values in Dimensionality			
Concept Pair*	1	2	2	3	4	5
Toiletry–sewing gear	35	30	.14	.09	.06	.04
Cleaning utensils–garden utensils	30	31	.08	.06	.05	.04
Kitchen utensils–tableware	38	29	.17	.11	.08	.05
Clothing–accessories	47	38	.15	.09	.06	.04
Vehicles– construction machines	40	36	.14	.09	.06	.04
Insects–fish	34	40	.05	.04	.04	.03
Insects–birds	34	43	.07	.06	.05	.04
Insects–mammals	34	32	.06	.04	.04	.04
Fish–birds	40	43	.05	.04	.04	.03
Fish–mammals	40	32	.05	.04	.03	.03
Birds–mammals	43	32	.07	.05	.04	.04
Trees–flowers	24	24	.04	.03	.02	.02
Fruits–vegetables	35	44	.16	.09	.07	.05

*Category 1–Category 2.

Regression analyses. To evaluate linear separability of the concepts in each pair, a logistic regression procedure was used. Exemplars were allocated to the category for which they were generated most frequently. The resulting dichotomous variable was used as the criterion variable, and the exemplar coordinates from the MDS solutions functioned as predictors in four separate regressions, corresponding to the MDS solutions in the four different dimensionalities with from two to five dimensions. To give an idea of how linear separability can be evaluated on the basis of MDS coordinates, Figure 2 shows, for the *toiletry–sewing gear* concept pair, the plotted MDS solution in two dimensions. The solid line, which draws the optimal boundary provided by the logistic regression, divides the group in two categories. (Due to space limitations, only one plotted MDS solution can be given.)

The proportions of correct classifications (i.e., correct category predictions) are shown in Figure 3 for all 13 concept pairs in each of the four dimensionalities. Exact

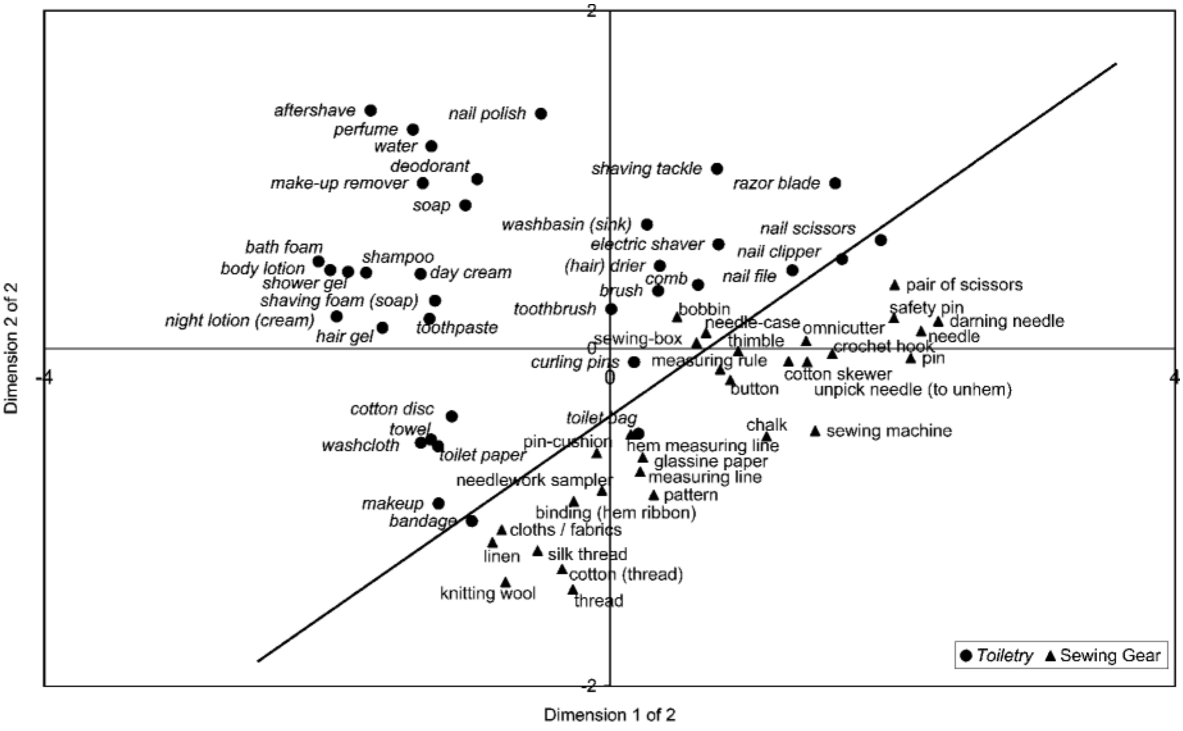


Figure 2. MDS representation of a two-dimensional solution for the contrast pair *toiletry–sewing gear*. The solid line marks the optimal boundary provided by the logistic regression.

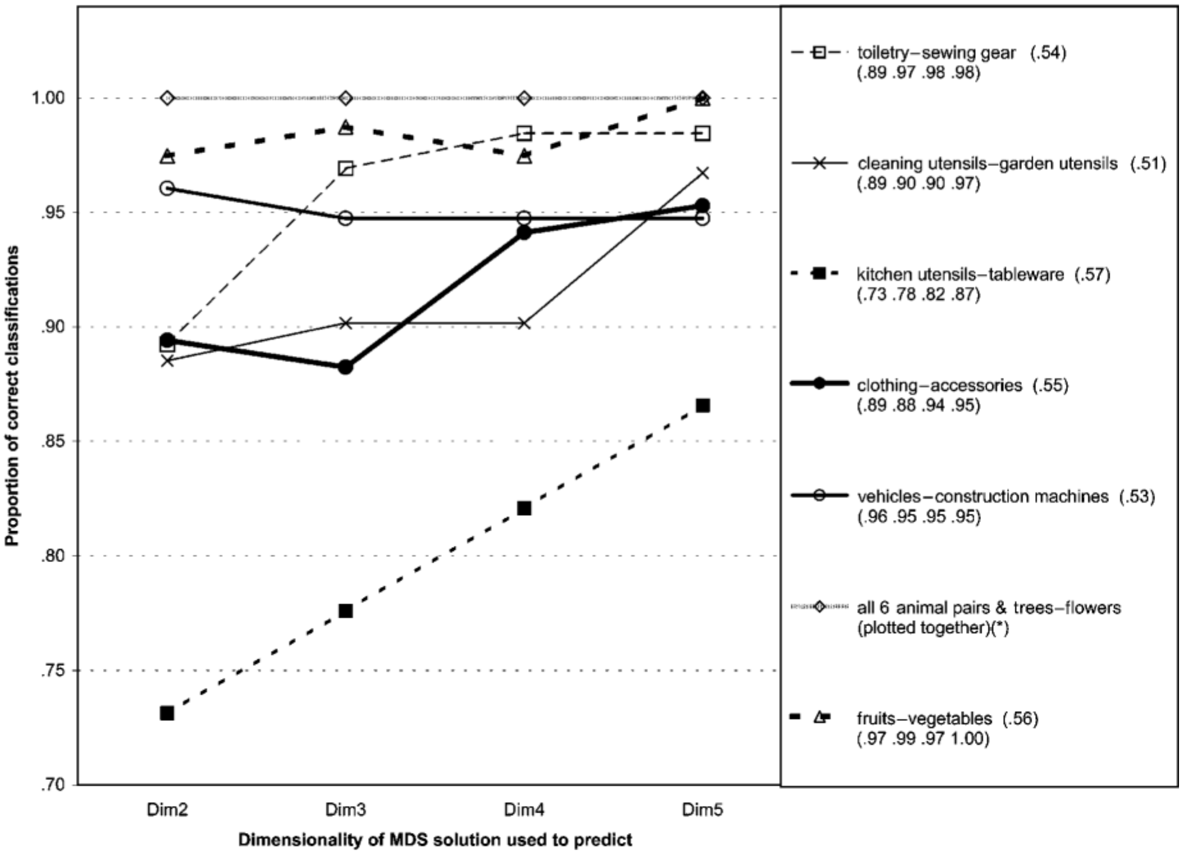


Figure 3. Proportion of correct classifications for exemplar generation task predictions in two to five dimensions. *The series *insects–fish*, *insects-birds*, *insects–mammals*, *fish–birds*, *fish–mammals*, *birds–mammals*, and *trees–flowers* all have a 100% correct classification and, for the sake of clarity, use the same legend key.

proportions for solutions in two to five dimensions, respectively, are printed in the second set of parentheses after the category names in the legend. Except for *fruits–vegetables* in two, three, and four dimensions, category membership could be perfectly predicted for all natural kind categories in every dimensionality. Proportions for the artifact categories never reached 100%, but all those except for *kitchen utensils–tableware* were above 85% in a two-dimensional solution and 95% or above in five dimensions. Thus, none of the artifact categories showed perfect linear separability in any dimensionality up to five.

In a logistic regression without predictors (i.e., without information other than the incidence of both values of the criterion), it is optimal to classify all exemplars in the category with the highest incidence. In the concept pair *kitchen utensils–tableware*, for example, 57% of the exemplars were generated as *kitchen utensils.* Therefore, 73% correct classifications for *kitchen utensils–tableware* in two dimensions is rather low if one takes into account that without any dimensional information at all, 57% could still be correctly classified. The proportion reduction in error (PRE) in a logistic regression is to be compared with the proportion of explained variance of a linear regression analysis and gives the quality of the category prediction using a logistic regression procedure. Formally,

$$\mathrm{PRE} = \frac{\mathrm{VAR}(Y) - \mathrm{VAR}(\varepsilon_i)}{\mathrm{VAR}(Y)},$$

where Y is the criterion variable and ε_i is the prediction error.

As shown in Figure 4, the difference between *kitchen utensils–tableware* and the other artifact concept pairs was more pronounced on the PRE measure, with PRE values varying between .288 (for two dimensions) and .658 (for five dimensions).

The difference between the average number of misclassifications for natural kind pairs and artifact pairs was tested with a randomization test (Edgington, 1995; Onghena & Van Damme, 1994). The difference turned out to be statistically significant, with p values below .01 for solutions in two, three, four, and five dimensions. Thus, linear separability was violated significantly more in artifacts than in natural kinds.

A more detailed overview of the misclassified exemplars can be found in Table 2. Here we see that, for many of these exemplars, the difficulty in predicting category

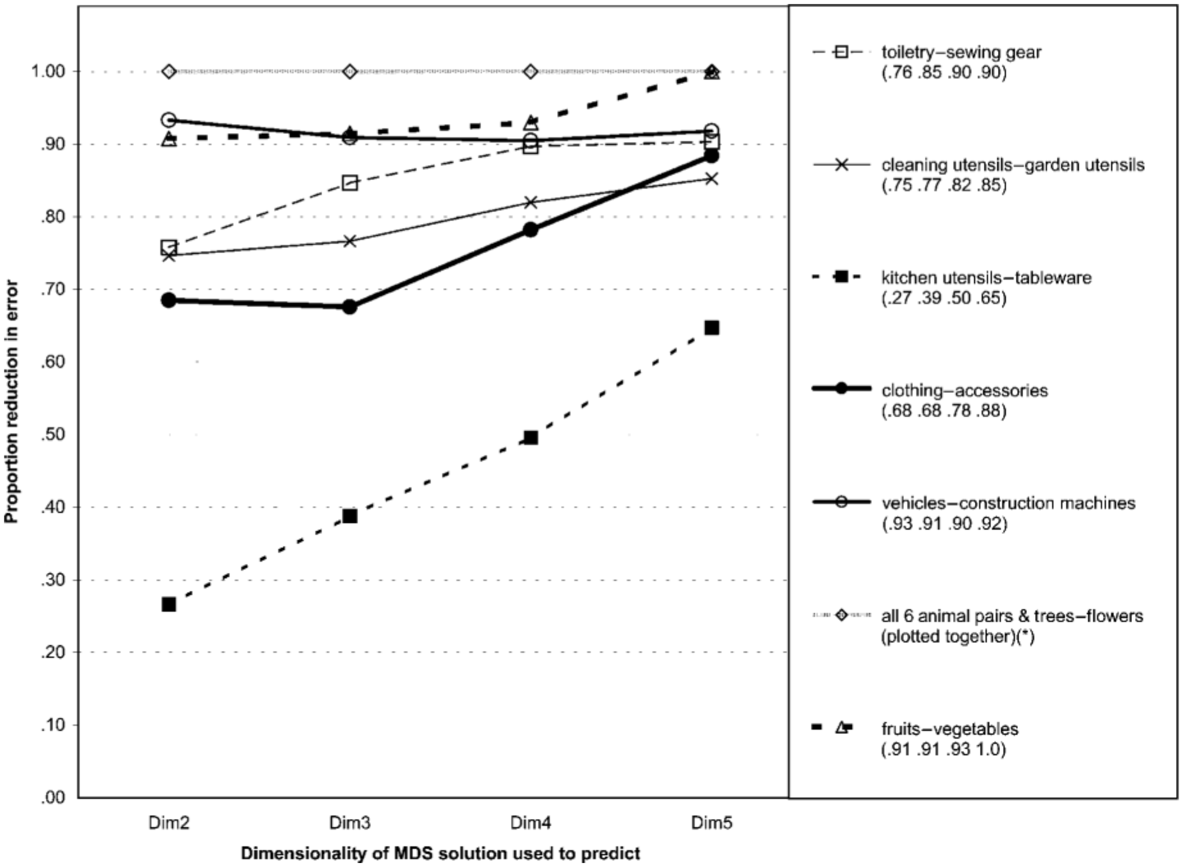


Figure 4. Proportion reduction in error (PRE) for exemplar generation task predictions in two to five dimensions. *The series *insects–fish*, *insects–birds*, *insects–mammals*, *fish–birds*, *fish–mammals*, *birds–mammals*, and *trees–flowers* all have a PRE of 1.0 in all dimensions and, for the sake of clarity, use the same legend key.

membership may be owing to the overlap of the two categories involved. For example, a bucket may serve both as a cleaning utensil and as a garden utensil. Whereas most biological kinds form a taxonomy of mutually exclusive classes, there is no such constraint on artifact kinds. A failure of linear discriminability in artifacts could therefore be attributed to two possible sources. It may be that the failure is the result of a significant region of overlap between the two categories, leading to vagueness in the allocation of an exemplar to one or the other exclusively. Alternatively, it may be that some objects are more similar to the prototype of one group, but belong to the other—a "true" violation of the assumptions of the independent cue model. One way to check these two alternatives is to retest linear separability in the data of Experiment 1 after leaving out all exemplars that were generated for both categories in a concept pair. When this was done, the results were very similar to the above-described analyses, including all exemplars, with artifact pairs still failing to show full separability even when exemplars in the overlap region were removed.

In summary, Experiment 1 showed that superordinate natural language concepts referring to categories within the animal and plant kingdoms were clearly linearly separable, even within low dimensionality representations with as few as two or three underlying dimensions. Artifact concepts, on the other hand, although reasonably well separated by a linear function, showed violations of linear separability for every pair that was studied, even when up to five underlying dimensions were used or when exemplars generated for both categories in a contrasting pair were left out.

One possible difficulty with interpreting these results is that we relied on exemplar generation to determine category membership. Where an item was generated to both members of a contrasting pair of categories, we assigned it to whichever category it was generated to most frequently. Furthermore, some items were generated only for one category (e.g., *bat* for *birds*) whereas, if choosing between the two contrasting categories (e.g., *birds* vs. *mammals*), a participant may be more inclined to categorize it in the other (e.g., *mammals*). Given a small sample size, it is therefore possible that some items may have been misassigned in Experiment 1, in terms of the category that people would judge them to be closest to on some underlying measure of category membership.

Table 2
Misclassified Exemplars by the Logistic Regression Predictions

Misclassified Exemplar	Member of Category	Dimensionality of Occurrence			
Toiletry–sewing gear					
cosmetics bag	1	D2	D3	D4	D5
nail clipper	1	D2	–	–	–
bobbin	2	D2	–	–	–
sewing box	2	D2	–	–	–
needle case	2	D2	–	–	–
nail scissors	1	D2	–	–	–
bandage	1	D2	–	–	–
curling pins	1	–	D3	–	–
Cleaning utensils–garden utensils					
working gloves	2	D2	D3	D4	D5
plant spray	2	D2	D3	D4	D5
copper/brass polish	1	D2	D3	D4	–
hand wiper (squeegee)	1	D2	D3	D4	–
bucket	1	D2	D3	D4	–
garden hose	2	D2	D3	–	–
seed	2	D2	–	D4	–
Kitchen utensils–tableware					
corkscrew	2	D2	D3	D4	D5
knife	2	D2	D3	D4	D5
oven glove	1	D2	D3	D4	D5
wooden spoon	1	D2	D3	D4	–
pot (Tupperware)	1	D2	D3	D4	–
filleting knife (fish knife)	2	D2	D3	D4	–
boiler	1	–	D3	D4	D5
ladle	1	–	D3	D4	D5
chopping board	1	–	D3	D4	–
salt	1	D2	–	–	D5
fish fork	2	D2	D3	–	–
coffee spoon	2	D2	D3	–	–
dessert spoon	2	D2	D3	–	–
dessert fork	2	D2	D3	–	–
wok	1	D2	D3	–	–
rolling pin	1	D2	D3	–	–
colander	1	D2	–	–	–
scraper	1	D2	–	–	–
spatula	1	D2	–	–	–
fork	2	D2	–	–	–
soup spoon/tablespoon	2	D2	–	–	–
bottle opener	1	–	–	D4	–
toothpick	2	–	–	D4	–
jar	2	–	–	D4	–
coffeepot	1	–	–	–	D5
bowl	2	–	–	–	D5
cutlery	1	–	–	–	D5
Clothing–accessories					
mitt(en)	2	D2	D3	D4	D5
fur coat	1	D2	D3	D4	D5
glove	2	D2	D3	D4	–
lingerie garter	1	D2	D3	D4	–
scarf	2	D2	D3	–	–
shinpad stocking	2	D2	D3	–	–
shoelace	2	D2	D3	–	–
cape	2	D2	–	D4	–
cap	2	D2	–	–	–
helmet	2	–	D3	–	D5
bonnet	1	–	–	–	D5
earmuff	2	–	D3	–	–
waistcoat (vest)	1	–	D3	–	–
Vehicles– construction machines					
skip container (truck)	2	D2	D3	D4	D5
wheelbarrow	1	D2	D3	D4	D5
tanker	1	D2	D3	D4	D5
conveyor	2	–	D3	D4	D5
Fruits–vegetables					
rhubarb	1	D2	D3	D4	–
winter radish	2	–	–	D4	–
medlar	1	D2	–	–	–

Note—The second column indicates the exemplar generation categorization. Last four columns specify the MDS dimensionality used for prediction for which misclassification occurred. Concept pair: Category 1–Category 2.

In Experiment 2, we therefore repeated the tests of linear separability using a different criterion for category assignment, based on a more direct judgment.

EXPERIMENT 2

The predictions in Experiment 1 were made for category membership on the basis of the results of an exemplar generation task, where exemplars were assigned to the category for which they reached the highest generation frequency. Taking a closer look at the particular exemplars that were wrongly categorized in Table 2, one can see that there is a prima facie case for misassignment. The astute reader familiar with the concept literature may have wondered how the MDS models were so readily able to classify notorious cases such as *whales* and *bats* in the correct categories. In fact, in our exemplar generation data, *whale* was more commonly generated as a fish, and *bat* as a bird, so that the models were correctly capturing the participants' generation of category members, even though they were not capturing the biologically correct classification of these species. (The fact that *whale* is termed *walvis* in Flemish, with *vis* being the term for fish, adds to the potential for confusion in exemplar generation.) It is therefore possible that use of generation frequency resulted in the overestimation of the linear separability of natural kind pairs and in the underestimation of artifacts. Therefore, in Experiment 2, participants were asked to classify all exemplars of a concept pair in a forced-choice task. Because a forced-choice classification task involves both categories of a concept pair, it was assumed that this would result in more reliable categorizations than those based on generation frequencies for separated categories, and so provide a better test of linear separability.

Method

Participants. Eleven participants between the ages of 19 and 42 years took part in this experiment. Three of them were male research assistants who participated voluntarily, 4 were first-year psychology students at the University of Leuven who partially fulfilled a requirement in their study curriculum, and 4 were female master's psychology students who were each paid the equivalent of $6 US for their participation.

Materials. The concepts studied, as well as the exemplar set used, were identical to those from Experiment 1.

Procedure. Thirteen lists of exemplars, corresponding to the 13 superordinate concept pairs that were used in Experiment 1, were presented to participants in a forced-choice task. The participants had to indicate to which of the two categories of a concept pair each exemplar best belonged. For half of the participants the two category names were presented in reverse order.

Results and Discussion

Reliability. The reliability of the data resulting from the forced-choice task was evaluated by split-half correlations corrected with the Spearman–Brown formula, with halves referring to the summed entries of half of the participants that completed the task. All estimates were .99 or above. Nonmodal responding (McCloskey & Glucksberg, 1978), or the proportion of responses that disagreed with the majority view, was low for both types of category, although higher for artifacts than for natural kinds. Only *kitchen utensils–tableware* (6%) showed more than 5% nonmodal responding. Agreement was thus generally high, and there was surprisingly little vagueness in the determination of the category boundaries.

Regression analyses. Categorization according to the majority view, as well as the categorizations from each of the 11 participants separately, were used as criterion variables in a prediction study analogous to the one described in Experiment 1. We will describe the analyses based on the majority view in detail. Because space limitations do not permit an elaborate description of the analyses for separate participants, we will summarize these results only briefly.

Figure 5 shows the proportions of correct classifications for the 13 concept pairs. The log linear analyses of the averaged dichotomized data of the different natural kind category pairs were again perfectly linearly separable in solutions with two to five dimensions, with the exception of *fruits–vegetables*, where five underlying dimensions were needed to reach perfect prediction. As for the artifact categories, category predictions were almost identical to those based on the generation frequencies used in Experiment 1. The only exception was the *vehicles–construction machines* pair, where perfect linear prediction could be obtained in four and five dimensions for the data from Experiment 2.

Figure 6 shows the PRE values for the 13 concept pairs. For natural kind pairs, the results were very similar to those described in Experiment 1: The difference between the PRE values for the dichotomized forced-choice data and the values calculated in Experiment 1

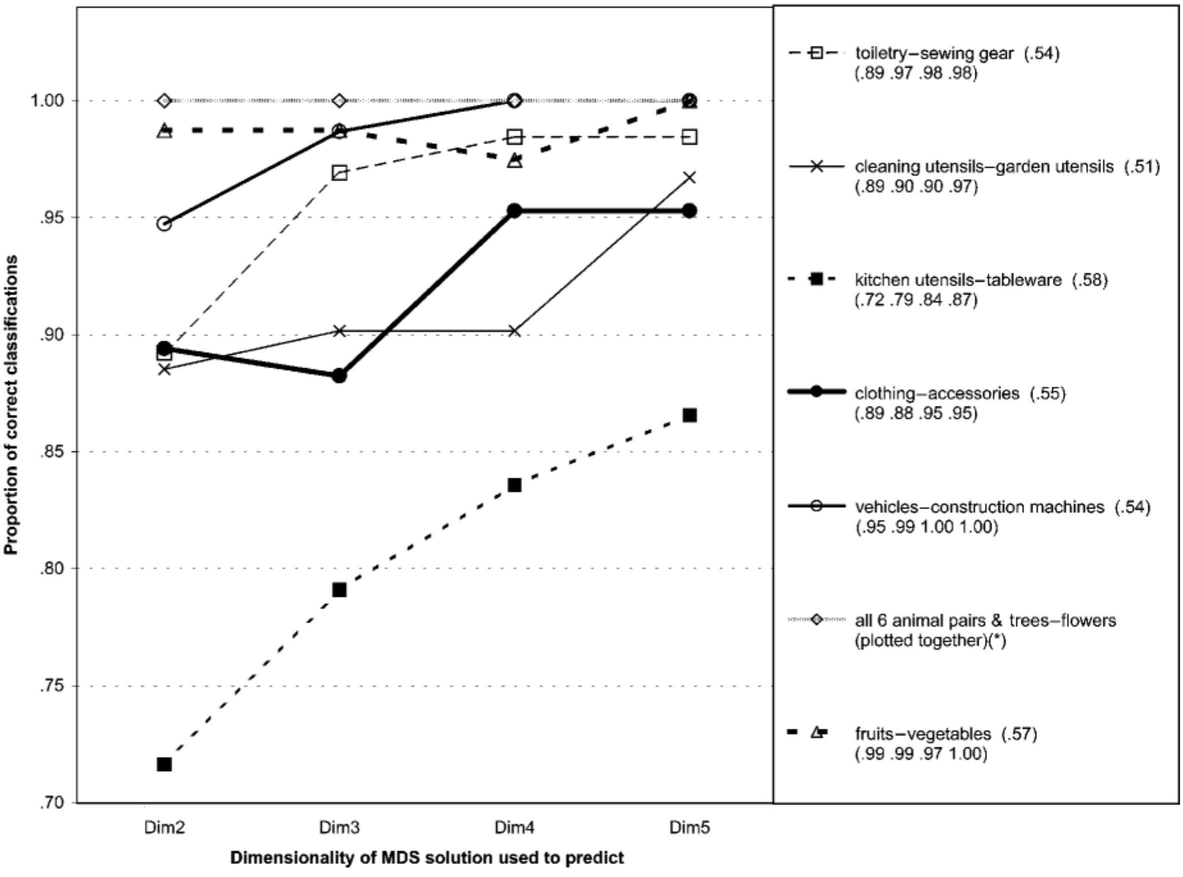


Figure 5. Proportion of correct classifications for the dichotomized forced-choice classification in two to five dimensions. *The series *insects–fish*, *insects–birds*, *insects–mammals*, *fish–birds*, *fish–mammals*, *birds–mammals*, and *trees–flowers* all have a 100% correct classification and, for the sake of clarity, use the same legend key.

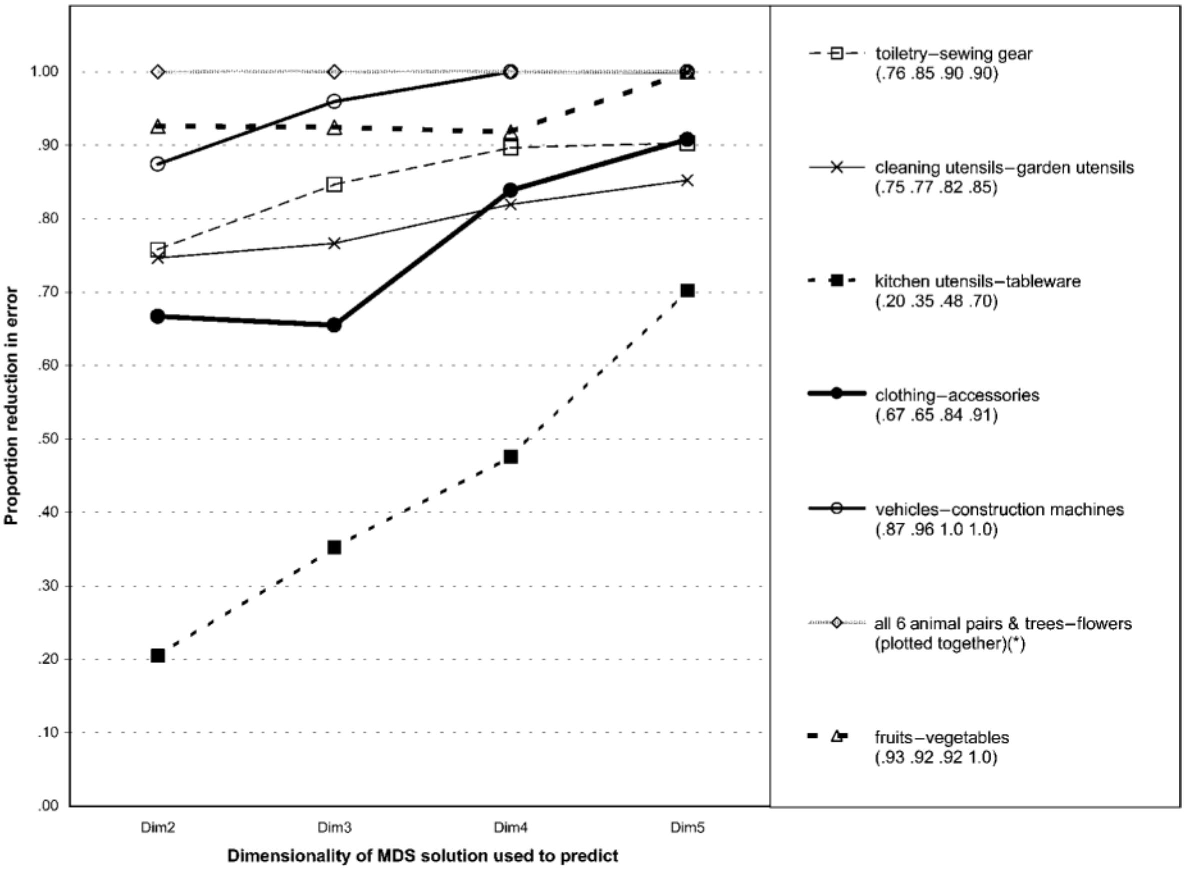


Figure 6. Proportion reduction in error (PRE) for the dichotomized forced-choice classification predictions in two to five dimensions. *The series *insects–fish*, *insects–birds*, *insects–mammals*, *fish–birds*, *fish–mammals*, *birds–mammals*, and *trees–flowers* all have a PRE of 1.0 in all dimensions and, for the sake of clarity, use the same legend key.

were never larger than .02. For artifact concepts, the results were somewhat different from those of Experiment 1. PRE values for *toiletry–sewing gear* and *cleaning utensils–gardening utensils* remained unchanged. For the three other artifact concept pairs, PRE values in Experiment 2 were somewhat higher in four and five dimensions than they were in Experiment 1. PRE values above .90 were found for the *clothing–accessories* concept pair in five dimensions and PRE values of 1.00 were found for *vehicles–construction machines* with four and five dimensions, indicating a far better prediction for these concept pairs. The difference between PRE values for *kitchen utensils–tableware* and for the other artifact concept pairs were far more pronounced than in Experiment 1, with PRE values ranging from .21 (for two dimensions) to .70 (for five dimensions).

The difference between the number of misclassifications for natural kind pairs and that for artifact pairs for the dichotomized forced-choice categorization was again evaluated with a randomization test (Edgington, 1995; Onghena & Van Damme, 1994). As in Experiment 1, there were again significantly more violations of linear separability in the artifact concept pairs than in the natural kind concept pairs ($p < .05$ for all dimensionalities).

Table 3, which shows the individual data, indicates the number of participants (out of 11) for whom violations of linear separability were observed for every concept pair and for solutions in two to five dimensions. Membership predictions for the individual forced-choice clas-

Table 3
Number of Participants Out of 11 Who Showed Violations of Linear Separability in Two to Five Dimensions

	Participants with Violations for Dimensionality			
Concept Pair	2	3	4	5
Toiletry–sewing gear	11	11	11	11
Cleaning utensils–garden utensils	11	11	11	11
Kitchen utensils–tableware	11	11	11	11
Clothing–accessories	11	11	11	10
Vehicles–construction machines	7	7	7	6
Insects–fish	3	1	1	0
Insects–birds	4	1	1	0
Insects–mammals	1	1	1	0
Fish–birds	0	0	0	0
Fish–mammals	8	6	5	4
Birds–mammals	1	0	0	0
Trees–flowers	7	7	1	1
Fruits–vegetables	10	10	8	4

sifications showed the same general picture as the majority view categorizations. More specifically, almost all participants showed perfect linear separability for most pairs of natural kinds in three to five dimensions, but for pairs of artifact concepts, linear separability was rarely obtained. The natural kind pairs were not perfectly linearly separable for every individual participant, but most of the prediction errors could be attributed to erroneous (or, more charitably, "idiosyncratic") categorizations made by individual participants with respect to the biological (for animals, trees, and flowers) or commonsense (for fruits and vegetables) categorizations. For example, 1 participant categorized *catfish* as a *mammal* in choosing between mammals and fish, whereas the model predicted it in all dimensions to be a *fish*. Similarly *aster* and *gerbera* (also *Transvaal daisy*) were occasionally categorized as *trees*, but were predicted by the model in lower dimensions to be *flowers*, and *pomegranate* and *papaya*, both predicted to be *fruits*, were categorized occasionally as *vegetables*. Most likely, such classifications reflect either unintended oversights or incomplete knowledge on the part of individual participants about the words used or the cultural norms involved.

A second class of prediction errors can be attributed to genuine cases where similarity-based categorization appears to break down. For instance, a *sperm whale* (which is called *protvis* in Flemish, with *vis* being translated as "fish") has mostly fishlike features. Not surprisingly, *sperm whale* was predicted to be a fish in every dimensionality but was categorized by some participants as a *mammal*. Similarly, *rhubarb* was often predicted (in three to four dimensions) to be a *vegetable* but was categorized as *fruit* by the participants. Note however that predictions for the other commonly confused item—*bat*—correctly matched the participants' categorization in solutions with more than two dimensions.

A randomization test (Edgington, 1995; Onghena & Van Damme, 1994) was used to evaluate the difference between the number of misclassifications for natural kind pairs and artifact pairs, for data from every participant in the forced-choice categorization task separately. The tests for every participant yielded a significantly larger number of violations for the artifact pairs ($p <$ .05) except for the analyses for 1 single participant in three, four, and five dimensions, where no significant difference was found.

The results of Experiment 2 confirmed the conclusions drawn from Experiment 1. It was again shown that the studied superordinate natural language concept pairs are nearly linearly separable. Experiment 2 again indicated that there was a difference between artifact and natural kind concept pairs. Natural kind concepts were clearly linearly separable in low-dimensionality representations—those with as few as two or three underlying dimensions—if the consensus view on category assignment was used. At an individual level, with four or five dimensions, even the *whale–fish/mammal* assignment correctly matched the majority of participants' category judgments. Artifact concepts were again found to yield many more violations against linear separability, even in five dimensions. The number of violations in the artifact pairs was significantly larger than that in the natural kind pairs.

In the context of Experiment 1, we discussed how some discriminability problems may have originated from the presence of items that belonged in both categories. In Experiment 2, this account can no longer be offered. Since participants were required to choose which of the two contrasting categories an item best belonged to, the linear discriminability constraint would clearly apply. Items in a region of overlap should be allocated in this task to the category to which they have the highest similarity (closest proximity), and so failures of linear discriminability can be attributed only to the existence of items that violate the assumptions of independent cue models, being more similar to the prototype of one concept, but being judged to be a better member of the other.

GENERAL DISCUSSION

Two experiments have been described that investigated linear separability in superordinate natural language concepts. Both experiments used the same set of contrasting superordinate concepts, eight pairs of natural kinds and five pairs of artifacts. In the first experiment, classification of exemplars was based on an exemplar generation task, and in the second experiment, classifications were based on a forced-choice classification task. The results of both experiments showed that superordinate natural language concepts are broadly speaking nearly linearly separable, but that there are significant violations of the constraint, particularly for the artifact domain.

The results of the two experiments showed a clear difference between the natural kind and artifact pairs in that, except for *fruits–vegetables*, all natural kinds were perfectly linearly separable with as few as two or three underlying dimensions. On the other hand, whereas artifact concepts approached linear separability, only one artifact category pair, *vehicles–construction machines*, showed perfect linear separability in up to five dimensions. Using randomization tests, we showed that the number of violations in the artifact concept pairs was significantly larger than in the natural kind pairs. The only exception for the natural kind pairs studied was the concept pair *fruits–vegetables*, where five underlying dimensions were needed to reach perfect linear separability. Notably, although *fruits* and *vegetables* are groupings of natural kinds, they differ from the other natural kinds in the degree to which their human use also determines their classification. For example, *rhubarb* is not biologically speaking a fruit, but presumably its use in jams and desserts leads it to be categorized in that class.

One point worth considering is that for exemplars we used lexical terms referring not to individual objects but to classes of objects. Thus the terms *spoon* or *bucket* could refer to a diverse set of individual objects, some of which may be better members of one superordinate, and some of the other. It may well be the case therefore that our data have underestimated the true extent to which linear discriminability may be violated. Malt et al. (1999) used photographs of individual objects in their study of container terms and reported a striking failure of the objects named by particular labels to form convex regions of the similarity space.

Another question that naturally arises in the context of our results is the degree to which a five-dimensional MDS solution would be able to fit any categorization data because of the number of free parameters estimated in the model. Perhaps the use of the model to predict random categorical classifications might lead to similar proportions of correct classification. To further investigate this possibility, a simulation study was conducted with random categorical data. For all natural kind pairs and for three out of five artifact pairs (*toiletry–sewing gear*, *kitchen utensils–tableware*, and *cleaning utensils–gardening utensils*) 225 random category distributions were generated. The only constraint in generating random distributions was that the number of items in each concept within a concept pair was kept equal to the corresponding number in the stimulus set that was investigated in Experiments 1 and 2. A new series of logistic regression analyses, with exemplar coordinates in two to five dimensions (from the MDS solutions of Experiment 1) as predictors and the random categorizations as criteria, was conducted for the 11 concept pairs. This procedure resulted in 9,900 predicted category distributions (225 random category distributions for 11 pairs in four dimensionalities). All 225 random distributions for all concept pairs invariably yielded a maximum percentage correct lower than that of the real data, usually around 25% or more. In short, random categorical distributions cannot be predicted from the MDS coordinates as predictors.

A third consideration is that it may be possible to achieve perfect linear separability for artifact concepts if more than five dimensions are used. There is some evidence in Figures 1–4 that the separability was improving as dimensionality increased. We chose not to investigate higher levels of dimensionality mainly because of concerns about the reliability of the representations and the increased risk of overfitting the data. Our analysis in up to five dimensions has been sufficient to show both that natural kind categories have strongly separable representations, and also that artifact categories are much less easily separated.

Regardless of whether natural categories are in fact all perfectly linearly separable in some number of dimensions, or whether significant violations exist, the difference in linear separability between artifacts and natural kinds was very robust in our studies. The violations of linear separability in the artifact pairs documented in our experiments are in line with findings from Malt et al. (1999). Even though they considered similarity representations in only two dimensions, their finding fits nicely with our own results.

The observed difference between living kinds and artifacts is of particular importance given recent interest in the nature of category-specific dementias. A common pattern of neurological disorder is for patients to lose the ability to name artifacts while preserving the ability to name creatures and plants. Superordinate kind terms such as *fish* and *bird* are particularly likely to be preserved even in advanced cases of dementia. Our results therefore provide additional evidence of the well-separated nature of these categories in similarity space, which has been claimed to be one possible reason for this pattern (Humphreys & Forde, 2001; Tyler, Moss, Durrant-Peatfield, & Levy, 2000). They also provide support for more radical proposals concerning the different ontological status of semantic categories of living things as opposed to artifacts (e.g., Sloman & Malt, 2003; Wierzbicka, 1984).

One can argue that the clear linear separability of natural language concept pairs is good news for supporters of independent cue models. This result is particularly striking considering that it is just this domain of creatures and plants that is most often claimed to involve categorization principles not based on similarity (Rips, 1989). However, the considerable number of violations of linear separability in the artifact pairs yields evidence against the independent cue models, because it is not clear how these models, including the prototype model, can account for the classification decisions from individual participants in Experiment 2. Our data point to an important dissociation between similarity and categorization (Rips, 1989), and one that speaks very directly to the structure of the everyday concepts held in semantic memory.

Exemplar models such as the context model (Medin & Schaffer, 1978), the generalized context model (Nosofsky, 1984, 1986), or relational cue models in general, are flexible enough to account for linearly separable structures, as well as for nonlinear structures in natural concepts. Indeed, recent research by Storms, De Boeck, and Ruts (2001) and by Smits, Storms, Rosseel, and De Boeck (2002) on the concept pair *fruits–vegetables* showed that the best predictor of the classification of unfamiliar exotic plant foods was similarity to a nearest neighbor exemplar, rather than similarity to a category prototype.

REFERENCES

Ashby, F. G., & Maddox, W. T. (1992). Complex decision rules in categorization: Contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception & Performance*, **18**, 50-71.

Blair, M., & Homa, D. (2001). Expanding the search for linear separability constraint on category learning. *Memory & Cognition*, **29**, 1153-1164

Edgington, E. S. (1995). *Randomization tests* (3rd ed.). New York: Marcel Dekker.

Estes, W. K. (1994). *Classification and cognition.* New York: Oxford University Press.

Franks, J. J., & Bransford, J. D. (1971). Abstraction of visual patterns. *Journal of Experimental Psychology*, **90**, 65-74.

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought.* Cambridge MA: MIT Press.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning & Verbal Behavior*, **18**, 441-461.

Hampton, J. A. (1997). Psychological representations of concepts. In M. A. Conway (Ed.), *Cognitive models of memory* (pp. 81-110). Hove, U.K.: Psychology Press.

Hampton, J. A. (1998). Similarity-based categorization and fuzziness of natural categories. *Cognition*, **65**, 137-165.

Hayes-Roth, B., & Hayes-Roth, F. (1977). Concept learning and the recognition and classification of exemplars. *Journal of Verbal Learning & Verbal Behavior*, **16**, 321-338.

Humphreys, G. W., & Forde, E. M. E. (2001). Hierarchies, similarity, and interactivity in object recognition: "Category-specific" neuropsychological deficits. *Behavioral & Brain Sciences*, **24**, 453-509.

Komatsu, L. K. (1992). Recent views of conceptual structure. *Psychological Bulletin*, **112**, 500-526.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Beverly Hills, CA: Sage.

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory & Language*, **40**, 230-262.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, **6**, 462-472.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, **85**, 207-238.

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning & Memory*, **5**, 355-368.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 104-114.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39-57.

Onghena, P., & Van Damme, G. (1994). SCRT 1.1: Single-case randomization tests. *Behavior Research Methods, Instruments, & Computers*, **26**, 369.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, **3**, 382-407.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21-59). New York: Cambridge University Press.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, **7**, 573-605.

Sebestyen, G. S. (1962). *Decision-making processes in pattern recognition*. New York: Macmillan.

Sloman, S. A., & Malt, B. (2003). Artifacts are not ascribed essences, nor are they treated as belonging to kinds. *Language & Cognitive Processes*, **18**, 563-582.

Smith, E. E., & Medin, D. M. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Smith, J. D., Murray, M. J., & Minda, J. P. (1997). Straight talk about linear separability. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 659-680.

Smits, T., Storms, G., Rosseel, Y., & De Boeck, P. (2002). Fruits and vegetables categorized: An application of the generalized context model. *Psychonomic Bulletin & Review*, **9**, 836-844.

Sperber, D. (2000). Metarepresentations in an evolutionary perspective. In D. Sperber (Ed.), *Metarepresentations: A multidisciplinary perspective* (pp. 117-137). New York: Oxford University Press.

Storms, G., De Boeck, P., & Ruts, W. (2000). Prototype- and exemplar-based information in natural language categories. *Journal of Memory & Language*, **42**, 51-73.

Storms, G., De Boeck, P., & Ruts, W. (2001). Categorization of novel stimuli in well-known natural concepts: A case study. *Psychonomic Bulletin & Review*, **8**, 377-384.

Storms, G., De Boeck, P., Van Mechelen, I., & Ruts, W. (1996). The dominance effect in concept conjunctions: Generality and interaction aspects. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **22**, 1-15.

Storms, G., Ruts, W., & Vandenbroucke, A. (1998). Dominance, overextensions, and the conjunction effect in different syntactic phrasings of concept conjunctions. *European Journal of Cognitive Psychology*, **10**, 337-372.

Sutcliffe, J. P. (1993). Concepts, class, and category in the tradition of Aristotle. In I. Van Mechelen, J. A. Hampton, R. S. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis* (pp. 35-65). London: Academic Press.

Tyler, L. K., Moss, H. E., Durrant-Peatfield, M. R., & Levy, J. P. (2000). Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain & Language*, **75**, 195-231.

Verbeemen, T., Vanoverberghe, V., Storms, G., & Ruts, W. (2001). The role of contrast categories in natural language concepts. *Journal of Memory & Language*, **44**, 618-643.

Wattenmaker, W. D., Dewey, G. I., Murphy, T. D., & Medin, D. L. (1986). Linear separability and concept learning: Context, relational properties, and concept naturalness. *Cognitive Psychology*, **18**, 158-194.

Wierzbicka, A. (1984). Apples are not a "kind of fruit": The semantics of human categorization. *American Ethnologist*, **11**, 313-328.

NOTE

[1] Due to space limitations, lists of all 928 exemplars and 218 features are not included but are available with English translations on request.

(Manuscript received February 6, 2003;
revision accepted for publication August 5, 2003.)