Comparing the Utility of Pairwise and Feature-Derived Similarity Measures for Generating Spatial Representations of Semantic Concepts

Matthew J. Dry (Matt.Dry@psy.kuleuven.be)
University of Leuven, Department of Psychology, Tiensestraat 102
B-3000 Leuven, Belgium

Gert Storms (Gert.Storms@psy.kuleuven.be)
University of Leuven, Department of Psychology, Tiensestraat 102
B-3000 Leuven, Belgium

Abstract
This study compares the relative utility of similarity data gathered using either direct pairwise ratings or feature-derived methods in regards to generating spatial representational models of fifteen well-known semantic categories. We assess both the extent to which the different similarity data sets can be accurately represented by a spatial model (representational goodness of fit), and the ability of the resulting spatial models to predict an external empirical semantic variable (predictive validity). The results indicate that feature-derived similarities obtain a better representational goodness of fit than the pairwise similarities, and that the predictive validity of representations based on common features data is far superior to the predictive validity of representations based on any of the other similarity data sets.

Keywords: Semantic Concepts; Representational Models; Similarity; Typicality.

Introduction
Spatial models of semantic concepts represent category exemplars as points in a multidimensional psychological space. This class of representational model has a demonstrated utility in research dealing with the storage and retrieval of semantic concepts and has been applied in studies involving a wide range of empirical tasks including the categorization of novel stimuli (Smits, Storms, Rosseel, & De Boeck, 2002; Verbeemen, Vanpaemel, Pattyn, Storms, & Verguts, 2007), categorization response time in the semantic verification task (e.g., Rips, Shoben, & Smith, 1973) and the same-different task (e.g., Carramazza, Hersh, & Torgerson, 1976), analogical reasoning (e.g., Rumelhart & Abrahamson, 1973), and the prediction of empirical typicality ratings (e.g., Ameel & Storms, 2006; Verheyen, Ameel, & Storms, 2007).

Spatial representations are commonly generated using a multidimensional scaling (MDS) algorithm. These algorithms require as input a matrix of empirical data defining the similarity (or dissimilarity) between each of the \( n \) exemplars within a stimulus set and all of the other \( n \) exemplars within the set. Given the input data, the algorithm searches for the representation (with a given dimensionality) that best represents the data such that exemplars that have a high degree of similarity tend to be located close to each other within the space, and dissimilar exemplars located far apart. For example, within a spatial representation of the category mammals, the exemplars lion and tiger could be expected to be located closer to each other than the exemplars monkey and whale.

Numerous methods have been employed to obtain empirical similarity data including direct pairwise ratings, sorting tasks, triad ratings, conditional probability ratings, rank proximity in generation tasks, and a range of feature-applicability based measures (see Borg & Groenen, 2005 for a review). By far the most common method of obtaining similarity ratings in psychological research is the pairwise technique: a review of 120 published datasets of similarity ratings of semantic categories indicated that 65% of the ratings had been obtained using a pairwise comparison method. The next most common method of obtaining empirical similarities was to derive them from feature matrices (20.8%), with other methods each making up less than 10% of the remaining data sets.

Although these data collection methods differ considerably, it is implicitly assumed that the various measures all tap the same underlying psychological construct. However, it should be noted that very few studies have attempted to directly compare the predictions made by different similarity measures, and when comparisons have been made they have been restricted to the similarity structure of a single category (e.g., Bijnol & Wedel, 1995; Henley, 1969; Johnson & Tversky, 1984; Wish, 1976).

Current Study
In this study we aim to compare the relative utility of similarity data gathered using either direct pairwise ratings or feature-derived methods in regards to generating spatial representational models of fifteen well-known semantic categories. We do this in two ways. First, we are interested in the extent to which the different similarity data sets can be accurately represented by a spatial model. Second, we are interested in the ability of the resulting spatial models to predict an external empirical semantic variable.

The primary aim of a spatial model is to represent the similarity structure of a given category within a multidimensional space. Interestingly, past studies have indicated that MDS algorithms recover the inter-item structures of different forms of similarity data with varying success. Of particular interest to the current study is the finding in Johnson and Tversky (1984) indicating that...
feature-derived similarities were more successfully recovered by an MDS algorithm than pairwise similarities. However, the Johnson and Tversky study was limited to a single category (perceptions of risks), and only 2- and 3-dimensional representations were compared. In this study we employ 15 well-known semantic categories, and compare representational fits across 1 to 10 dimensions.

It is important to note that just because a given similarity data set can be accurately represented by a spatial model doesn’t mean that the resulting representation is of any psychological interest. Theoretically, it would be possible to perfectly represent the similarity structure of a randomly generated data set, but the resulting representation would be as devoid of meaning as the data upon which it was based.

A quantifiable approach to assessing the psychological validity of the representations is to compare them in regards to the prediction of independent psychological variables (Verheyen, Ameel, & Storms, 2007). For example, a widely replicated finding in the semantic category literature is the graded structure exhibited by many natural categories: although participants may agree that a set of exemplars all belong to the same category, they tend to differentiate between the exemplars in regards to their rated typicality as members of that category (e.g., Rosch & Mervis, 1975). Further, a number of studies have demonstrated that these graded structures can be predicted by distances within a multidimensional space (e.g., Ameel & Storms, 2006; Henley, 1969; Rips, 1975; Verheyen, Ameel, & Storms, 2007). Following this, in the current study we assess the validity of the representations derived from the different empirical typicality ratings.

Method

All of the data employed in this study were taken from the normed semantic category data set reported in De Deyne, Verheyen, Ameel, Vanpaemel, Dry, Voorspoels and Storms (submitted).

The present study focuses upon 15 semantic categories. Five of the categories are taken from the domain of animals (mammals, birds, fish, insects and reptiles), six from the domain of artifacts (kitchen utensils, musical instruments, clothing, weapons, vehicles and tools), two from the domain of foods (fruits and vegetables), and two from the domain of activities (sports and professions). Each of the categories contained between 20 to 33 exemplars.

Pairwise Similarity Data

The pairwise similarity data were obtained using a standard pairwise similarity rating task. For each of the categories, participants were presented with all of the unique pairings of the category exemplars, and they were asked to rate the similarity of the two presented exemplars using a scale ranging from 1 = totally dissimilar to 20 = totally similar. Between 15 and 25 participants rated each category.

Feature-Derived Similarity Data

The feature based similarities were derived from feature-by-exemplar matrices. Each matrix contained between 20 to 33 columns representing each of the exemplars within a single category, and between 156 to 382 rows representing descriptive features for the exemplars within the category. As is described in detail in De Deyne et al. (submitted), the features were obtained by asking 1003 participants to generate 10 descriptive features for a small, randomly selected subset (N = 6 to 10) of the category exemplars. An independent group of participants was then required to indicate for each feature/exemplar combination the applicability of the feature, using a 1 to indicate that the feature applied to the exemplar, or a 0 to indicate that it did not. The feature applicability task was completed by 8 participants in each category.

There are a number of methods available for obtaining similarities from feature-by-exemplar matrices. In this paper we focus upon five methods: a correlation-based measure, three feature matching measures, and a Euclidean distance measure.

Feature vector correlation. A common approach to deriving a measure of similarity between two exemplars is to calculate the correlation between the feature-vectors associated with each stimulus. Given the feature vectors v_i and v_j, the similarity between exemplars i and j is given by the correlation measure:

$$ s_{ij} = \frac{\sum_k (v_{ik} - \bar{v}_i)(v_{jk} - \bar{v}_j)}{\sqrt{\sum_k (v_{ik} - \bar{v}_i)^2}\sqrt{\sum_k (v_{jk} - \bar{v}_j)^2}}, $$

where k indicates the number of features in each vector.

Feature matching. An alternative approach to conceptualizing the similarity between two exemplars is in regards to their degree of featural overlap. Numerous methods have been proposed to account for inter-exemplar similarity based upon either common features, distinctive features, or a combination of the two (see Tversky, 1977 for a review). In this paper we employ three similarity measures based on Tversky’s (1977) Contrast model. The Contrast model generalizes a number of feature matching measures by employing a parameter which controls the relative contribution of common and distinctive features to the overall similarity measure. Following Navarro and Lee (2004), inter-item similarity under the Contrast model can be calculated as:

$$ s_{ij} = \lambda + \left[ \rho \sum_k v_{ik}v_{jk} - \left( 1 - \rho \right) \sum_k v_{ik} (1 - v_{jk}) \right] - \left[ (1 - \rho) \sum_k (1 - v_{ik})v_{jk} \right], $$

where \( \lambda \) is a constant and \( \rho \) is the weight given to common features, \( \lambda / \left( 1 - \rho \right) \) is the weight given to distinctive features, and \( \rho \) is a parameter which controls the relative contribution of common and distinctive features to the overall similarity measure.
where $0 \leq \rho \leq 1$, and $\lambda$ is an additive constant indicating a universal level of inter-exemplar similarity. From this it can be seen that setting the parameter $\rho$ to low values emphasizes distinctive features, whereas setting $\rho$ to high values emphasizes common features. In this study we employed three values of $\rho$ (0, 1, and 0.5) corresponding (respectively) to a purely distinctive features model, a purely common features model, and a balanced model that gives equal weight to common and distinctive features.

Euclidean distance. According to this approach each row in the feature-by-exemplar matrix can be conceptualized as a dimension in a multi-dimensional space, with the location of each exemplar along each dimension in this space indicated by the presence or absence of the corresponding feature. Following this, it is possible to calculate the similarity between any two exemplars in the feature space using:

$$s_{ij} = \lambda - \left[ \sum_k \left( v_{ik} - v_{jk} \right)^2 \right]^{\frac{1}{2}},$$

where $\lambda$ is a constant used to transform the distances into similarities.

MDS Representations
Spatial representations were generated from the six different similarity data sets (Pairwise, Correlation, Euclidean, Distinctive, Common and Balanced) using the metric multidimensional scaling algorithm described in Lee (2001). All of the similarity data sets were normalized to lie in the range 0-1, and were converted into proximities ($d_{ij}$) using a linear transformation ($d_{ij} = 1-s_{ij}$). Representations were generated with one to ten dimensions for each of the 15 categories.

Typicality Ratings
Participants in the typicality rating task were presented with a list containing exemplars of a single category and were asked to rate (on a 20 point scale) how good each exemplar was as a member of the category. Each category was rated by 28 participants. The reliability of the typicality ratings ranged from $r = .90$ to .98.

Results
Representational Goodness Of Fit
We measure representational goodness of fit as the proportion of empirical similarity ($s_{ij}$) variance accounted for (VAF) by the model similarity predictions ($\hat{s}_{ij}$):

$$VAF = 1 - \frac{\sum_{i<j} (s_{ij} - \hat{s}_{ij})^2}{\sum_{i<j} (s_{ij} - \bar{s})^2},$$

where $\bar{s}$ is the average of the empirical similarities. We compare the fit between each similarity data set and its resulting representation to determine if there are differences between the various similarity types in regards to the adequacy of their representational fit across different levels of dimensional complexity.

Figure 1 shows the representational fits for the directly rated Pairwise, and the feature-derived Correlation, Euclidean, Distinctive, Common, and Balanced similarity data sets using spatial representations with one through ten dimensions. Due to the difficulties of presenting the results for each category separately Figure 1 is a summary plot showing the VAF averaged across the fifteen categories.

As can be seen, regardless of the type of similarity data that the spatial models were generated from the representational fits all improve as the dimensionality of the representation increases from one to ten dimensions. However, there is a striking difference between the quality of the fits for the Pairwise and feature-derived similarity data sets, with the Pairwise similarity based representations providing consistently worse fits, regardless of the assumed underlying dimensionality.

Somewhat surprisingly, the ten-dimensional Pairwise based representation only manages to achieve the same level of fit as is achieved by the feature-derived representations using around half the number of dimensions. Given that in MDS the number of free parameters equals the number of exemplars multiplied by the number of dimensions this difference in parametric complexity can be considered non-trivial.

It is important to note that the superiority of the goodness of fit for the feature-based representations was also found...
when each category was considered in isolation. While the ordering of the fits for the feature-based representations showed some variability across categories, with the exception of a single category the fits of the Pairwise-based representations were consistently poorer than those of all of the feature-based representations.

**Predictive Validity**

Ameel et al. (2006) and Verheyen et al. (2007) demonstrated that empirical typicality ratings of category exemplars can be successfully predicted by the distance between each exemplar within a multidimensional category space and the category prototype or centroid. Specifically, highly typical exemplars tend to be located close to the category centroid, whereas atypical exemplars tend to be located closer to the edge of the category boundary. Following this we modeled typicality as the distance between each exemplar within a category space and the category centroid (where the centroid is the average of the exemplar locations along each dimension). Figure 2 shows the correlation between typicality and distance-from-centroid for the representations derived from the six different similarity data sets, averaged over the fifteen semantic categories.

As expected there is a negative correlation between typicality and distance-from-centroid, regardless of the form of the similarity data from which the representation was generated. Furthermore, as was found in Verheyen et al. (2007), the strength of the correlation shows a (mild) increase as the number of dimensions is increased from 1 to 10. Most importantly however, Figure 2 indicates that there is a large difference in the predictive validity of the different representations. The Euclidean and Distinctive similarity based representations only achieve average correlations of around -.2, whereas the best average correlations for representations based on the Common similarity data are as high as -.69.

Correlation coefficients are useful in that they provide an easily interpretable measure of predictive fit. However, it could be argued that they are not well suited to distinguishing between different models: a correlation of -.57 is stronger than a correlation of -.55, but it is difficult to determine if this difference is of any psychological interest. In light of this we employ the Bayesian information criterion (Schwarz, 1978) to compare the relative likelihoods of the typicality predictions made using the different spatial representations.

Following this approach we calculate the weighted sum-square error fit of the empirical and representation-based typicality ratings:

$$WSSE = \sum \frac{1}{\sigma^2_i} (t_i - \alpha + \beta \eta_i)^2$$

where $t$ indicates the empirical typicalities, $\sigma^2$ is the variance of the empirical typicalities, and $\alpha$ and $\beta$ are regression coefficients employed to linearly transform the distance ($\eta$) between each exemplar and the category centroid into a predicted typicality. Under the assumption of a Gaussian likelihood function the Bayesian information criterion (BIC) is calculated as:

$$BIC = WSSE + p \ln n,$$

where $p$ is the number of free parameters (which in this case is equal to the number of regression coefficients), and $n$ is the number of category exemplars. The relative likelihood of the data fits can then be compared using Bayes factors (Kass & Raftery, 1995).

Table 1 shows the Bayes factors for the typicality prediction fits calculated across all 15 categories. As can be seen, the Bayes factors indicate that the representations generated from the Common similarity data are at least 2.40x10^8 times more likely than the alternative representations, a result that can be interpreted as a ‘decisive’ difference (Jeffreys, 1961).

It should be noted that these findings do not apply to the averaged data alone. Even when each category is considered in isolation, the representations based upon the Common similarity data tended to consistently provide better typicality predictions than the other representations. For the categories Musical Instruments, Vehicles, Weapons, Tools, and Sports, the Common similarity representations obtain minimum BIC values regardless of the underlying dimensionality of the solution. For Fruits and Kitchen Utensils the Common representations provide the minimum BIC in all but a single dimension. Finally, for the categories...
Table 1. Bayes factors for the typicality predictions of the 1 to 10 dimensional representations generated from the six similarity data sets. The data are based on BIC values calculated across all 15 categories.

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pairwise</td>
<td>1.88x10^{10}</td>
<td>1.35x10^{9}</td>
<td>9.59x10^{8}</td>
<td>8.87x10^{8}</td>
<td>2.11x10^{8}</td>
<td>5.10x10^{7}</td>
<td>9.46x10^{7}</td>
<td>7.85x10^{7}</td>
<td>5.74x10^{7}</td>
<td>4.70x10^{7}</td>
</tr>
<tr>
<td>Correlation</td>
<td>3.80x10^{8}</td>
<td>2.02x10^{11}</td>
<td>6.10x10^{9}</td>
<td>1.18x10^{10}</td>
<td>1.64x10^{10}</td>
<td>3.94x10^{10}</td>
<td>6.22x10^{10}</td>
<td>3.96x10^{10}</td>
<td>2.38x10^{10}</td>
<td>1.49x10^{10}</td>
</tr>
<tr>
<td>Euclidean</td>
<td>7.50x10^{7}</td>
<td>6.91x10^{14}</td>
<td>7.67x10^{15}</td>
<td>6.37x10^{16}</td>
<td>1.73x10^{17}</td>
<td>7.05x10^{17}</td>
<td>1.49x10^{18}</td>
<td>1.39x10^{18}</td>
<td>1.16x10^{18}</td>
<td>9.38x10^{17}</td>
</tr>
<tr>
<td>Distinctive</td>
<td>6.26x10^{10}</td>
<td>4.36x10^{15}</td>
<td>3.16x10^{16}</td>
<td>2.56x10^{17}</td>
<td>8.31x10^{17}</td>
<td>3.04x10^{18}</td>
<td>6.03x10^{18}</td>
<td>5.21x10^{18}</td>
<td>3.96x10^{18}</td>
<td>3.08x10^{18}</td>
</tr>
<tr>
<td>Common</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Balanced</td>
<td>1.12x10^{10}</td>
<td>1.52x10^{11}</td>
<td>1.67x10^{9}</td>
<td>8.15x10^{8}</td>
<td>8.64x10^{8}</td>
<td>1.24x10^{9}</td>
<td>1.56x10^{9}</td>
<td>8.73x10^{8}</td>
<td>4.52x10^{8}</td>
<td>2.40x10^{8}</td>
</tr>
</tbody>
</table>

**Professions, Insects and Reptiles** the Common representations provide minimum BICs for 3-, 2-, and 1-dimension(s) respectively.

In total, across the 15 (categories) x 10 (dimensions) representations, the Common similarity data provide minimum BIC-values for 74 of the representations. Of the remaining representations the Pairwise similarity data obtains 26 best predictions, the Balanced data 23 best predictions, the Euclidean data 14 best predictions, the Correlation data 11 best predictions and the Distinctive data 2 best predictions. Importantly, however, even when a Common similarity representation failed to provide the best prediction, in all but three cases the Bayes factors indicate that the prediction of the Common representation was at most only 3.2 times less likely than that of the best prediction. According to Jeffreys (1961) guidelines for interpreting Bayes factors this difference is ‘not worth more than a bare mention.’

**Discussion**

The aim of this study was to assess the utility of different similarity data for generating spatial representations. We assessed utility in two ways: representational goodness of fit, and predictive validity.

In regards to representational goodness of fit the results suggest that the feature-derived similarities are more adequately represented by a spatial model than pairwise similarities. Furthermore this pattern of results holds regardless of assumptions made about the underlying dimensionality of the representations.

It is not immediately obvious why there should be such a large difference in the representational fits of the Pairwise and feature-derived similarities. One potential explanation can be found in the distributional properties of the similarity data. For example, simulation studies (Pruzansky, Tversky, & Carroll, 1982) have shown that spatial representations tend to produce positively skewed distributions of inter-exemplar distances. Analysis of the empirical data indicated that the mean skewness of the inter-exemplar distance distributions (averaged across the 15 categories) for the Pairwise, Correlation, Euclidean, Distinctive, Common, and Balanced similarity sets was -1.39, -0.52, -0.81, -0.36, -0.69, and -0.52 respectively.

Whilst none of these values indicates positive skew, the Pairwise data set has a much stronger negative skew than any of the feature-derived similarities. One interpretation of this finding is that the distributional structure of the Pairwise data makes it inherently less suited to spatial representation than the feature-derived data. It also suggests that the Pairwise data may be better suited to tree-like representations than spatial representations (Pruzansky, Tversky, & Carroll, 1982). In order to determine if the poor representational fit of the Pairwise data was due to the form of the representational model we generated Additive Tree representations for the 15 categories using 2 to 20 internal nodes. The results indicated that the VAF of the Pairwise data based tree representations was marginally better than the feature based tree representations (around 0.03% difference). In general, however, the quality of the fits was much poorer than those obtained using spatial representations, with the mean VAF asymptotating at around 0.76%.

The skewness data are also interesting because they tell us something about the tasks employed to generate the different similarity data sets. On the one hand, the pairwise task appears to emphasize the high latent similarity between the exemplars as members of the same category (relative to non-members), resulting in many high similarity ratings and few low similarity ratings. On the other hand, the feature applicability task forces participants to consider the similarity of the exemplars in light of a wide range of descriptive features (some of which may only apply to a small subset of the category exemplars). As a result the feature applicability task appears to be more successful than the pairwise task at differentiating between the exemplars in regards to within-category similarity structure.

In regards to the second assessment criterion, predictive validity, the results indicated that all of the spatial representations were able to provide reasonable predictions of empirical typicality ratings regardless of the type of data from which they were generated. However, the Bayesian analyses indicated that the predictions of the Common similarity data representations were far more likely than those of any of the other predictions. This results accords with the findings of previous research indicating that typicality is dependant upon the degree of commonality or overlap in the features of category exemplars: typical exemplars tend to have many features in common with other category members, and atypical items tend to have few features in common with other category members (e.g.,
Rosch & Mervis, 1975; Smith, Shoben, & Rips, 1974). The current study is unique in that the correspondence between featural overlap and typicality has not been measured using raw features, but with a spatial representation derived from Common features similarity data.

It should be remembered that predictive validity in the current study was based purely upon typicality, however it is highly likely that a similar result would be found with categorization response time given that this tends to correlate with typicality and has been shown to be modeled reasonably successfully using a distance-from-centroid approach (e.g., Carramazzza, Hersh, & Torgerson, 1976; Rips, Shoben, & Smith, 1973). It is less obvious whether similar results would be found using other semantic variables such as familiarity, age of concept acquisition and category association frequency.

**Conclusion**

The results of this study indicated that feature based similarity data was superior to Pairwise similarity data in regards to the goodness of fit for derived spatial representations of semantic concepts. Furthermore, Common features based representations were shown to have a much better predictive validity (in regards to empirical typicality ratings) than representations based upon any of the other similarity data sets.

These results are somewhat surprising given the predominance of the pairwise rating task for collecting similarity data. In psychology it is generally assumed that the pairwise rating task is the ‘gold standard’ for the collection of similarity data. This study suggests that in regards to the similarity structure of semantic concepts this standard needs to be revised. Whether or not the same conclusion applies to the similarity structure of perceptual and abstract stimuli is a question that remains to be tested empirically.

**References**


