Running head: GRADED STRUCTURE IN ADJECTIVE CATEGORIES

Accounting for graded structure in adjective categories with

valence-based opposition relationships

Simon De Deyne (simon.dedeyne@ppw.kuleuven.be)[1,2]

Wouter Voorspoels (wouter.voorspoels@ppw.kuleuven.be)[1]

Steven Verheyen (steven.verheyen@ppw.kuleuven.be)[1]

Daniel J. Navarro (daniel.navarro@adelaide.edu.au)[2]

Gert Storms (gert.storms@ppw.kuleuven.be)[1]

Corresponding Author: Simon De Deyne

Tel: +32 16 326294 or +32 16 325965

Fax: +32 16 326099

Address: Laboratory of Experimental Psychology,

Tiensestraat 102 box 3711, 3000 Leuven, Belgium

[1]University of Leuven, Faculty of Psychology and Educational Sciences,

Tiensestraat 102, 3000 Leuven, Belgium

[2]University of Adelaide, School of Psychology,

5000 Adelaide, Australia

Abstract

In contrast to noun categories, little is known about the graded structure of adjective categories. In this study we investigated whether adjective categories show a similar graded structure and what determines this structure. The results show that adjective categories like nouns exhibit a reliable graded structure. Similar to nouns, we investigated if similarity is the main determinant of the graded structure. We derived a low-dimensional similarity representation for adjective categories and found that valence differences in adjectives constitute an important organizing principle in this similarity space. Valence was not implicated in the categories' graded structure, however. A formal similarity-based model using exemplars accounted for the graded structure by effectively discarding the valence differences between adjectives in the similarity representation through dimensional weighting. Our results generalize similarity-based accounts of graded structure and highlight a closely-knit relationship between adjectives and nouns on a representational level.

Keywords: concepts; adjectives; typicality; valence; GCM; similarity.

One of the central questions in cognitive science is how people mentally represent categories of objects, events and relationships between entities in the world. Largely the research in this area has focused on artifact and natural kinds categories, that is to say, words - most often nouns - that refer to concrete objects such as cats, bats, apples and coconuts. While object categories are undoubtedly important in the human conceptual apparatus, the exclusive attention to this class of concepts leaves a considerable gap in our understanding, since many classes of concepts remain untouched (Medin, Lynch, & Solomon, 2000). Perhaps the most notable class of concepts that has remained out of focus are adjectives, that is to say, words that refer to features such as taste, smell, size and quality. Adjectives are often used in models of representation of object categories to denote their properties. Gärdenfors concisely summarizes the dominant attitude toward both word classes, stating that in natural language the distinction between properties and concepts shows up in the distinction between adjectives and nouns, with adjectives normally referring to properties and nouns referring to concepts (Gärdenfors, 2000). Adjectives are seldom given the "concept" status. Instead, there is a tendency to reserve this for nouns.

In the present study, we focus on adjectives, and more specifically, we investigate the extension of classes of adjectives that cover a semantic domain, for instance, adjectives that constitute *a description of someone's appearance.* Starting from the categories of adjectives, we focus on two key questions that have been frequently addressed in research on categories of nouns. First, we ask if members of the adjective categories reflect a graded structure, as do noun categories, and second, we ask whether a computational model that can account for the graded structure in nouns is apt at capturing a similar structure in the adjective categories. Before describing the empirics, we first give some background on the importance of graded structure and the challenges of examining graded structure in adjective categories.

Graded structure in categories

One of the most robust observations in a broad variety of categories is that membership is judged to be continuous rather than dichotomous. In the case of noun concepts, for instance, *cats* are considered to be more typical, or better examples, of the category mammals than are *bats*. This graded membership structure, or typicality gradient, has been reliably observed in a broad range of natural language categories, spanning animal and artifact categories (e.g., Rosch & Mervis, 1975), food categories (e.g., Ross & Murphy, 1999), activity categories (e.g., De Deyne et al., 2008), categories of verbs (Plant, Webster, & Whitworth, 2011), abstract noun categories (Hampton, 1981; Verheyen, Stukken, De Deyne, Dry, & Storms, 2011), goal-derived and ad hoc categories (Barsalou, 1983, 1985), conceptual combinations (e.g., Smith & Osherson, Rips & Keane, 1988), logically defined categories (Nosofsky, 1991) and artificial categories (e.g., Nosofsky, 1988). Importantly, graded membership has been shown a key influence in a wide range of tasks, generally in the form of a processing advantage for typical items as compared to atypical items. Such influence has been confirmed in a wide range of cognitive phenomena such as category verification, lexical decision, word production (Hampton & Gardiner, 1983), inductive reasoning (Rips, 1975), priming (Rosch, 1977), memory interference (Keller & Kellas, 1978) and naming (for a review see Hampton, 1993).

The empirical observation of the typicality gradient in itself does not imply a specific model of how a category is represented (see Rosch 1978, pp. 39-40 for a discussion). Moreover, different types of categories can have different principles underlying the gradient. For example, while for concrete concepts typical members tend to be those that are considered more similar to all other members of the category or have more features in common with the other exemplars

(e.g., Hampton, 1979), this does not hold for goal-derived categories. Instead, an item's ability to fulfill the implied goal provides a better account of the item's judged representativeness (Barsalou, 1985; Voorspoels, Storms & Vanpaemel, submitted). To understand the graded membership of a structure, and by extension, the processing advantage it grants to more typical items, it is imperative to identify the principles underlying the gradient. Surprisingly, for adjective categories little effort has been undertaken to establish a graded membership, let alone, examine its source. As such, an empirical test is needed to investigate whether such a graded structure is present and what the important determinants of this structure are.

From nouns to adjectives

While nouns and adjectives constitute different grammatical classes, there are salient similarities between them, both in structure and use. The two grammatical classes are neither clear-cut nor universal. For instance when asked to what extent words are good examples of the part-of-speech class of nouns, *teacher* and *table* are considered good examples of this grammatical category, while less concrete words such as *doorway* and *sky* are not, regardless of the fact that all these words are actually nouns (Taylor, 2003, p. 210). Similarly, adjectives vary in the degree to which they exhibit typical adjectival properties (e.g., whether they are used attributively or predicatively, whether they admit comparative and superlative forms, etc). Apart from typicality differences within grammatical classes of words, the distinction between the classes is blurry as well. In English, adjectives can function as nouns (ever since Robin Hood robbed the *rich* to give to the *poor*), nouns can be used as modifiers (e.g., *killer application* or *adjective noun*) and nouns and adjectives can be turned into verbs (i.e., nouns can be *verbed*, which *weirds* language altogether[1]).

Valence is an important determinant of the mental representation of both nouns and adjectives (e.g., Grühn & Smith, 2008; Kousta, Vigliocco, Del Campo, Vinson, & Andrews, 2011; Kousta, Vinson, & Vigliocco, 2009). Both abstract and concrete nouns such as *diamond*, *nurse* or *truth* are considered to have a positive valence, while words such as *dentist*, *mosquito* or *stress* have a negative valence (Bradley & Lang, 1999). Emotional adjectives in particular (*happy-sad*) have been found to contrast on the dimension of valence as well (Bernat, Bunce, & Shevrin, 2001; Herbert, Kissler, Junghofer, Peyk, & Rockstroh, 2006). Correspondences like these motivate the need for theories that take into account the points of convergence between nouns and adjectives.

Despite the strong analogies between adjectives and nouns, addressing adjectives in a way similar to nouns, that is, examining the graded structure of adjective categories, poses new challenges. The presence of various opposition relationships for many adjective pairs such as *fast-slow*, *hot-cold*, or *young-old* is particularly challenging. While certain nouns such as *husband-wife*, *winner-loser* come in oppositional pairs, classical theories of concept representation such as the prototype or exemplar view have often been able to ignore these relationships, presumably because they are not very prominent among the exemplars of the noun categories that are generally studied (*fruit*, *furniture*, *tools*). They may prove impossible to ignore in adjective categories, however, as opposition has been argued to be the basic relation in adjective pairs (Murphy & Andrew, 1993)[2], affecting even the order and difficulty of adjective acquisition (Landau & Gleitman, 1985). It is not immediately clear how such opposition can be reconciled with theories of category representation that derive category representations by simultaneously considering all exemplars in a set. For instance, if we assume that *hot* and *cold* are dissimilar, but both typical adjectives to describe a temperature, it is not clear how similarity-

based approaches -- such as in exemplar theory -- can be used to derive that a different word to describe temperature is highly typical without being similar to both *hot* and *cold* at the same time.

A number of practical issues need to be considered as well. One of them is the fact that many adjectives are highly context dependent compared to nouns. *Cool*, for example, has a different meaning when it modifies *breeze* than when it modifies *music*. Another issue is the fact that it is not clear how to derive a similarity structure for adjectives, since one of the most successful approaches for nouns based on a feature generation task has no analog for adjectives(what, e.g., are the characteristics of *cool*?). If our approach to uncover the structure of adjective categories succeeds in face of these challenges, we will have unlocked a number of models that have been proven successful in the noun domain. The representational assumptions and psychological processes involved in these models may be generalized to the new domain, allowing the integration of findings within a broader view on semantic structure.

Outline

In the following sections, we first describe how categories for adjectives reflecting natural domains such as colors, shapes, weather descriptions or quality judgments, can be used to study graded structure. Based on these categories, we obtain a direct measure of graded structure by asking participants to judge the goodness of the exemplars for a category. These data allow us to answer the question whether a reliable graded structure can be identified. We then examine whether the graded structure in the adjective categories can be captured by an approach based on the general similarity-based approach in noun categories. In particular, we implement an exemplar model to account for the obtained structure. Next, we investigate how it handles opposition relationships, by interpreting the underlying dimensions that determine the adjectives'

similarity representation in terms of independently obtained attributes and investigating the weights the model assigns to them when accounting for the adjectives' graded structure. Previewing our results, this analysis suggests that opposition in terms of valence is an important part of the similarity structure of adjective categories, but not of their graded structure. The exemplar model overcomes this discrepancy by assigning a low weight to the valence dimension. We end with a brief discussion on rivalry, non-similarity based approaches.

Study 1: Categories of Adjectives

One difference between nouns and adjectives is that ~~ate fact that~~, unlike nouns, adjectives seem to lack a hyponymic relationship. According to some researchers, the IS-A relationship is undefined for adjectives without specifying a noun domain they might modify (Gross & Miller, 1990; Murphy & Andrew, 1993). For instance, one can reflect on different adjectives to describe a feeling, and some of these adjectives might be better examples than others, but no adjective can be considered a superordinate. The apparent absence of an IS-A relationship does not necessarily preclude some kind of hyponomic organization for certain adjectives. For example, *smooth* could be further differentiated to distinguish different ways of smoothness ranging from *velvety* or *silky* to *rough*, *bumpy* or *jagged*. This is somewhat similar to hierarchical relations between verbs such as *whispering* which is a way of *talking* (Plant, Webster & Withworth, 2011). In addition, linguists have proposed various ways in which adjectives can be organized in a small number of classes (Dixon, 1982; Raskin & Nirenburg, 1996). The work by Dixon (1982), for example, used semantic, syntactic and morphological properties of adjectives to distinguish seven classes of adjectives in English. These classes relate to dimension (*big*, *long*, etc.), physical properties

(*hard*, *heavy*, etc.), colors (*orange*, *yellow*, etc.), human properties (*happy*, *generous*, etc.), age (*young*, *old*, etc.), value (*good*, *pure*, etc) and speed (*fast*, *quick*, *slow*, etc.). Even in languages with only few adjectives, age, dimension, value and color are generally present.

Recent psychological theories on embodied and grounded cognition indicate that the meaning of adjectives can be organized along modality specific simulations (Lynott & Connell, 2009; van Dantzig, Cowell, Zeelenberg, & Pecher, 2011). Lynott and Connell, for instance, had 423 prenominal adjectives that are used to describe various properties rated on five perceptual modalities: vision, audition, touch, smell and taste, providing an organization of adjectives in terms of the senses. In an attempt to combine these approaches, we cast our net widely, aiming to capture a number of adjective categories that cover organizing principles reflecting natural domains related to abstract entities, objects, person and emotion properties, and the senses. Consistent with previous work using nouns, we ask participants to generate exemplars for each of the suggested categories, as a method for approximating their extensions. These exemplar generations provide a first, tentative answer to the questions how adjectives are psychologically organized in categories. A more elaborate answer will be provided in the subsequent studies, where within- and between-exemplar variability of typicality judgments will be addressed.

Method

We adapted the exemplar generation procedure for nouns described in Ruts et al. (2004) for use with adjectives and asked participants to generate exemplars for adjective categories. The adjective categories were described in terms of organizing principles that included specific modifiers. Because adjective meaning often depends on context, an indication of the noun domain they modified was provided where applicable. In other words, instead of asking persons

to generate adjectives for objects or persons, they were asked to generate adjectives to describe the *shape* of an object or the *appearance* of a person.

Participants. Thirty-nine native Dutch-speaking volunteers (32 of them female and 7 male) were paid €8/hour. Their ages ranged from 19 to 57 years ($M = 23$).

Stimuli and materials. The generation cues were 22 adjective category descriptions, covering a broad range of adjectives. They can be divided in four overarching classes: descriptions of abstract properties (*a quality judgment*, *description of a quantity*, *degree to which something is difficult or hard*, *degree of certainty*, *description of weather conditions*, *departure from a norm*), object properties (*description of a landscape*, *appreciation of a work of art*, *description of a work of art*, *the shape of an object*, *the value of an object*, *the position of objects*), sensory properties (*description of music*, *description of the taste of food*, *the color of objects, temperature*, *the feel of an object*) and person and emotion properties (*description of someone's character*, *description of a person's appearance*, *description of the sound of someone's voice*, *description of intelligence*, *description of a mood*).

As it would be infeasible to ask any one participant to generate exemplars for all categories, four different subsets of these descriptions were constructed, each consisting of 11 categories. In this way, exemplars were generated for each category by half of the participants. To prevent related adjective categories co-occurring in the same permutation, two restrictions were imposed: These restrictions were applied to two category pairs (1) *appreciation of a work of art* and *description of a work of art* and (2) the *description of music* and the *description of the sound of someone's voice*.

Procedure. Each participant generated exemplars for one of the subsets of adjective categories. The task was presented in an Excel-file containing 11 sheets, one sheet for each

category in the subset containing the category description followed by 24 blank lines. The following instructions were given:

In this experiment, we investigate what adjectives people come up with spontaneously when trying to express or describe certain things. On each sheet, you will find the name of a category of concepts that can be described by various adjectives. Try to write down as many as possible in the allotted space.

Thirteen examples were given for the category of adjectives used to *describe buildings* and five examples were given for *phases in life*. The categories for each subset corresponded to a different sheet in the Excel file and participants were asked to complete these categories in the fixed order. Finally, the participants were instructed that there were no right or wrong answers but they should avoid verb forms corresponding to adjectives. The participants were allowed to work on this task by their own pace but were instructed to complete it within the hour.

Results and discussion

Responses for each category are summarized by tabulation. Table 1 shows the total number of tokens, the number of types and idiosyncratic types, the mean, standard deviation and the skew of the frequency distributions.

INSERT TABLE 1 ABOUT HERE

The counts show that for most categories participants found it easy to generate exemplars. For example, for *description of someone's character* 386 responses were recorded, which means that the participants who completed this category generated 20 adjectives on average, which is near the maximum space of 24 lines we provided. The counts also indicate considerable variation among the number of adjectives generated for a specific category, ranging from 183 for *degree of certainty* to 417 for *color of objects*, and a tendency for idiosyncratic responses (on average, 65% of the types that were generated, occurred only once). Moreover, out of 1,918 adjective types, 594 adjectives were generated for more than one category, reflecting the polysemous nature of these words. Despite the instructions asking for descriptive adjectives, qualitative judgments such as *good* and *bad* occurred in many categories such as *description of weather conditions* or *description of the taste of food*. Furthermore, as indicated by the positive skew, the adjectives that were generated most frequently were much more frequent than the subsequently generated adjectives, similar to Zipfian frequency distributions, where the frequency of words is inversely proportional to their rank.

Study 2: Typicality of Adjectives

In Study 1, we established the extension of 22 adjective categories. In Study 2 we examine whether - similar to noun categories - they demonstrate a graded structure. To this end we had participants judge exemplars' typicality towards their respective adjective categories.

Method

Participants. Twenty-seven female and ten male volunteers were paid €8/hour. Their ages

ranged from 19 to 29 years ($M = 23$). All participants were native Dutch speakers.

Stimuli and materials. Twelve categories from the initial set of 22 categories described in Study 1 were retained for the typicality judgment study. These categories were selected so as to cover the proposed adjective domains, while avoiding the inclusion of similar categories. In addition categories with only a small number of non-idiosyncratic exemplars (e.g., *departure from a norm*, where as indicated in Table 1 only 30 types were non-idiosyncratic) or categories that contain an adjective in the name (e.g., *degree to which something is difficult or hard*) were deemed less appropriate and were also not included. For each category, 30 adjective exemplars were sampled to cover the entire range of the production frequencies[3]. Note that some adjectives like *good* and *bad* were included in multiple categories.

Procedure. Participants completed a web-administered questionnaire. They were presented with the exemplars of an adjective category, and were asked to indicate how good an example each adjective was of the category on a seven-point Likert-scale ranging from 1 (a very bad example) to 7 (an excellent example). Every participant rated the typicality of all the exemplars of every category. The order of the categories and the items within the category were completely randomized for each participant and the task was completed in less than an hour.

Results and discussion

The reliability of the typicality judgments for each of the categories was estimated using the split-half correlation with Spearman-Brown correction. Nine categories were found to be very reliable ($r_{splithalf} > .90$). The categories *description of quantity*, *description of a work of art* and *description of a person's character* were only slightly less reliable. For these categories the values were, respectively, .83, .85 and .89. The median of the ratings varied between 4.43

(*description of a quantity*) and 5.45 (*color of an object*) indicating that most of the category members were considered to be more typical than atypical to the category[4]. The selection of category members across the range of the generation frequency distribution resulted in a clear graded structure for all categories, (mean $SD = .77$). For each category, we confirmed that the value of the standard deviation was different from what can be expected from categories with no typicality gradient by a permutation test. This test was performed by a Monte Carlo test consisting of permuting the subject ratings for each category 1,000 times, after which we calculated the mean of the standard deviation of each stimulus. We then compared the distribution of standard deviations with the original standard deviations. None of the distributions included the original standard deviation values (for the permuted values the maximum SD over all categories $= .39$, while the minimum observed among all categories was higher than the one for the permuted values, $SD = .55$). We also confirmed the validity of the exemplar selection procedure based on the generation frequency by calculating the correlations between the log-transformed generation frequencies and the mean typicality judgments. The correlation with generation frequency was significant for all categories (average $r = .59$, one-sided t), except for description *of a mood, r(30) = .29, p = .06*, one-sided t). In sum, the general pattern generalizes findings in noun categories to adjective categories. We find a stable and reliable typicality gradient in the adjective categories. Moreover, the positive correlation between typicality and the number of times an adjective is generated as a category exemplar confirms and generalizes previous findings for category membership and typicality of nouns (Barsalou, 1985; Mervis, Catlin & Rosch, 1976, Verheyen, Stukken, De Deyne, Dry, & Storms, 2011).

Study 3: Similarity of Adjectives

The data from Study 2 clearly show that a reliable, graded structure exists for the 12 adjective categories. The next step is to account for the established structure. In noun categories, the typicality gradient is often related to the underlying similarity structure that exists within a category (e.g., Heit & Barsalou, 1996; Storms, De Boeck, & Ruts, 2000). In the present Study, we will explore whether a similarity-based explanation also determines the graded structure present in adjective categories. In what follows, we will first derive a multidimensional spatial category representation, based on a measure of pairwise similarity between the adjectives within each category. The obtained spatial representation will be used to examine whether a similarity-based exemplar model can predict the typicality data that were obtained in Study 2. Finally, we will investigate what kind of structure the dimensions of the spatial representation measure.

Similarity Measure and Similarity Scaling

Feature-based similarity measures have been very successful in predicting conceptual data including typicality judgments (e.g., Dry & Storms, 2009). Since adjectives often correspond to concept features, standard feature listing tasks used for concrete nouns cannot be applied here. However, previous studies have shown that word association data capture the semantic representation well among a wide range of concepts (De Deyne, Peirsman, & Storms, 2009). For example, using a similarity judgments task, De Deyne, Peirsman and Storms found that similarity measures derived from word associations provided a good account for the judgment of animal $r(300) = .85$, and artifact concepts, $r(435) = .76$. While these values are slightly lower than the golden standard obtained using a similarity measure derived from judged semantic features (.89 for both domains), word association norms are more versatile since they are not restricted to defining information but capture thematic information as well.

To derive a similarity space from word associations, we relied on existing norms (De Deyne & Storms, 2008; De Deyne, Navarro, & Storms, 2012). The adjective exemplars were part of a dataset containing more than 12,500 cue words. For each word, 300 association responses were collected and the association responses were tabulated. The meaning of each adjective is represented by the association response distribution, which encodes the number of times a certain association was generated to the adjective cue. Using these distributions, similarity indices were derived in a manner identical to that in De Deyne et al. (2009). First, the counts were transformed using a *t*-score measure of concordance following a proposal by Church, Gale, Hanks, and Hindle (1991). Next, the similarity between two adjectives was calculated using the cosine measure. This was done for all adjective combinations in a category. These were then subjected to multidimensional scaling (MDS; Borg & Groenen, 1997), which converts the similarities between a category's exemplars into distances between points in a multidimensional space.

For each of the twelve categories solutions with dimensions varying from 2 to 6 were obtained. Kruskal (1964) suggests that solutions with a stress-value (i.e., a measure of the discrepancy between the input similarities and the output distances) exceeding .10 should not be considered for further analyses. Using this criterion to select the lowest dimensionality, results in dimensionalities with a mode of 4. A full description of the dimensionality and the stress values for each category is shown in the second and third column of Table 2.

Model Description

Similar to nouns, we expect that the underlying similarity space of the adjectives can be used to account for typicality. To examine this hypothesis, we consider an exemplar model of typicality that is grounded in a multidimensional similarity representation. Such a model-based

analysis can be informative in two ways: While it aims to provide a psychological account of

category representation and gradedness, it can equally fulfill the role of data-analytic tool. By

identifying the meaning and contribution of the dimensions that constitute the similarity data, the

model allows us to understand the underlying structure of the data.

The model used in this study corresponds to the Generalized Context Model (GCM;

Nosofsky, 1986). While this model was originally developed to account for categorization, it can

also be adapted for typicality (Nosofsky, 1991; Voorspoels, Vanpaemel, & Storms, 2008).

According to the GCM, the typicality of an exemplar is assessed by summing the exemplar's

similarity to all other exemplars. The typicality T_{iA} of an exemplar i for category A is thus given

by:

(1)

where η_{ij} represents the similarity between category exemplars i and j. The similarity between

two exemplars is derived from their psychological distance in a multidimensional spatial category

representation of the kind we obtained in the previous section using MDS. It is defined as:

(2)

$$\eta_{ij} = exp(-cd_{ij})$$

where d_{ij} represents the distance between exemplars i and j in the representational space and c is a

scaling parameter that shrinks or magnifies the space.

The psychological distance d_{ij} between exemplars i and j is given by

(3)

where x_{ik} and x_{jk} are the coordinates of exemplars i and j on dimension k of the space, w_k is the

weight granted to dimension k and K is the number of dimensions constitute the space. The

dimension weights w_k are constraint to sum to one and provide the model with a mechanism to

take into account that depending on the task requirements, different dimensions receive more or

less emphasis. In this study, the parameter r was fixed at 2 to correspond to Euclidean distances,

which are more appropriate for integral dimensions (Shepard, 1964, 1987).

Model Fit

The GCM was fitted by optimizing the correlation between the predicted typicality and

the observed typicality for each category consisting of 30 members separately. The results are

shown in the last column of Table 2. All correlations were significant at the .01 level (one-tailed

t). The strength of the correlations varied depending on the categories and ranged from moderate

to high for all categories.

INSERT TABLE 2 ABOUT HERE

To ensure that these results were not due to flexibility in (over)fitting the free parameters

to the data, we performed a permutation test. This test consisted of permuting the observed

typicality values a 1,000 times and finding the optimal prediction of the model for each of the

permutations. If the free parameters in the model ~~are~~were able to capture every pattern to the

same extent, we would expect optimal correlations for the permuted data sets that are within the same range as the correlations in Table 2. This was not the case. Averaged over categories, the correlation was $r = .31$ ($r_{min} = .20$, $r_{max} = .40$) which is considerably lower than the observed correlations in Table 2. To see if the correlations were statistically significant different by transforming them to a t value (see Meng, Rosenthal, & Rubin,1992) and performing a one-sided test. All permuted and original correlations were significantly different from each other, except for the *color of objects*, which was borderline significant $t = -1.49$, $n = 30$, $p = .07$. The results of the model fitting therefore indicate that the GCM exploits the structure that is present in the multidimensional representation of each category to account for the gradedness of its exemplars[5].

The application of a well-known model such as the GCM to adjectives allows a full comparison with previous results for nouns. To further validate our results for adjectives, we compared the performance of the GCM with that for the well-studied domain of noun concepts. Voorspoels, Vanpaemel and Storms (2011) applied the GCM to noun categories delineating animals and artifacts. The results in this study showed that the GCM obtained correlations with rated typicality of .65 averaged across 5 animal categories and .76 averaged across 6 artifact categories. The average correlation of .73 across all adjective categories indicates that the present results are similar in terms of magnitude. Because the similarity space in Voorspoels, et al. (2011) was derived from participant-generated semantic features restricted to a particular domain, rather than word associations we replicated these findings, using the association data that were presented earlier as input. Using similarity spaces derived from associations, the average optimized correlations for typicality were .61 for the animal categories and .77 for the artifact categories. This shows that the prediction for the adjectives is on par with that of concrete nouns, regardless of whether semantic features or word associates are used.

By optimally weighting the dimensions of the representation, the model succeeds in ordering the category exemplars in terms of judged typicality. As is the case for categories of nouns, similarity appears to be an important determinant of the graded structure participants perceive among sets of adjectives. The more similar an adjective is to an adjective category's exemplars, the more typical of the category it is deemed. To obtain a measure of similarity between adjective pairs, this study cast differences among word classes aside and employed association responses (regardless of word class) to both adjectives in a pair. The final Study in this paper is concerned with one of the dimensions that spans these similarities and its role in judgments of typicality: valence.

Study 4: Valence Opposition in Adjectives

Ideally, the multidimensional scaling solution should not only fit the similarity data well; it should also teach us something new about the data. In this case, it is important to consider the underlying structure in the MDS solutions in combination with the values of the dimensional weights derived in the GCM model to understand how the psychological distances affect typicality. The GCM dimension weights determine the contribution of each MDS dimension to the prediction of a category's graded structure. In the following section, we first interpret the dimensions that capture most of the structure in the similarity data and validate them with additional ratings that have previously been proposed to account for the semantic variation in adjectives. The last section investigates the role of the dimension weights in the account of the categories' graded structure, with particular emphasis on those dimensions that could be interpreted and validated by means of the additional ratings.

Interpreting the Spatial Representation

First, we investigated the underlying structure in the similarity space of adjectives. Previous research has shown that many concepts - especially abstract ones - carry an evaluative force , which we refer to as valence (Kousta et al., 2011). Related adjectives often come in pairs that contrast on this dimension of valence. In terms of similarity of meaning, adjectives such as *valuable* and *worthless* are clear opposites in terms of their valence and should be distal in a MDS space. At the same time, these concepts are closely related on a semantic level. They differ in terms of valence but might be similar on all other dimensions. By visualizing the first two dimensions of the MDS spaces, it becomes clear that valence distinguishes the adjectives. The *description of a mood* and *description of a character trait* categories provides a clear example of valence polarization. Indeed, as shown in Figure 1a and 1b, the first dimension distinguishes the adjectives in terms of valence.In Figure 1a this dimension spans negative moods on the left and positive moods on the right. A similar interpretation can be made for *description of a character trait* (Figure 1b).

INSERT FIGURE 1 ABOUT HERE

To verify if our interpretation of this structure can be quantified, we collected a subjective measure of valence for all adjectives that does not require the explicit comparison between a pair of two adjectives. In this way, problems related to specific contexts in which evaluative judgments are presented are minimized and adjectives can be differentiated even if a specific word does not have an obvious valence opposite. The valence judgments were collected as part of a norming study (Verheyen, De Deyne, Linsen, & Storms, 2012). ~~In this~~In this study, participants

were asked to indicate on a seven-point scale whether a word evoked a negative or positive feeling.

To evaluate the extent to which valence is indeed an organizing principle in the spatial stimulus representations that were derived from the association similarity data, we performed a property fitting procedure (Kruskal & Wish, 1978). In this analysis valence is the dependent variable in a multiple regression analysis, with all MDS dimensions as predictors. A high R^2 for the multiple regression analysis reveals that an optimal linear combination of the MDS-dimensions can produce a dimension that aligns nicely with the attribute. A graphical example of the property fitting procedure is shown in Figure 1 for the category *description of a mood and description of a character trait*. Fitting the regression line for valence (dashed line) approximately aligns with the first dimension and thus confirms our interpretation based on visual inspection.

The results of the property fitting procedure in which the exemplars' coordinates along each of the dimensions were added as predictors are presented in Table 3. To aid the interpretation, both the dependent variable (valence) and the ~~independent variables~~independent variables (the dimension coordinates) were centered. This allows us to interpret the standardized regression weights as correlation coefficients. A large regression weight indicates that the attribute dimension aligns nicely with the MDS-dimension that produces the regression weight. Only the models that were statistically reliable are displayed. This excluded the category *shape of an object*.

INSERT TABLE 3 ABOUT HERE

Except for *feel of an object* ($R^2 = .35$), we find moderate ($R^2 = .60$) to high ($R^2 = .94$) fits for valence as indicated by the multiple correlation coefficients. While the prediction of valence benefits from the contribution of multiple dimensions, the results indicate that a single dimension is often correlated strongly with valence, in line with the results of Table 3. This dimension is generally among the first dimensions that are extracted by the multidimensional scaling algorithm. As these dimensions capture the largest variability among the similarity data, this result indicates that valence is the most important organizing principle of most adjective categories.

Contribution to Graded Structure

The previous section showed that valence is strongly involved in the similarity structure of adjective classes. Despite being an important organizing principle in semantic space, valence itself does not correlate with any of the typicality ratings except for *description of an artwork* ($r(30) = .46$, $p < .05$, two-tailed t) and *color of an object* category ($r(30) = .52$, $p < .05$, two-tailed t). If valence does not provide any useful information about the graded structure of adjectives while dimensions corresponding to valence accounted for most of the similarity structure at the same time, how did the GCM manage to provide such a good account of the graded structure in these categories? To answer this question, we now focus on how the GCM accounts for typicality in the presence of valence polarization.

INSERT TABLE 4 ABOUT HERE

Crucial to the analysis is the interpretation of the dimension weights w_k. These allow us to infer which dimensions in the category representation are of importance in judging the typicality of the adjectives for their respective categories. While a particular attribute might correspond to an important axis in the psychological similarity space, this does not necessarily imply that this dimension contributes to the perceived typicality of the category exemplars. The GCM account of graded structure might yield a relatively low weight for the dimension. To investigate the proportional weight attributed to valence, the dimension with the highest weight (k_{max}) was compared with the dimension that was most closely related to valence (cfr. Table 3). As can be seen from Table 4, for all but three categories (*description of a quantity*, *color of an object* and *feel of an object*), the estimated weight for the valence dimension was relatively low, and often close to zero compared to the highest weight (w_{max}) assigned to a certain dimension k_{max} in the MDS solution. This indicates that the importance of the valence dimensions is down-weighted by the model to account for the typicality judgments. This holds for most categories, with the effect most prominently present in the person related adjective categories. In other words, the GCM attributes most of its weight to other dimensions than those that encode the valence of adjectives. This confirms our earlier intuitions that the valence dimension, which has a crucial role in the similarity structure, is far less important in typicality judgments, and therefore this attribute will have a low weight. These results indicate an important property of the model when applied to adjectives: its flexibility offered by dimensional weighting can account for valence-free typicalities using valence opposition-rich similarity representations, by selecting the most appropriate dimensions in the similarity structure.

General Discussion

Adjectives constitute a crucial element of meaning. Various theories of concept representation reserve an important role for adjectives. On the one hand, the meaning of a concept derives from experience with the world, that is to say, knowledge about the characteristics, verbalized through adjectives, of the extension of a concept (e.g., Hampton, 1979; Rosch & Mervis, 1975). For example, we know what the concept *dog* means because we have experience with the class of dogs, and have knowledge about the features and characteristics that are common for the concept (e.g., *furry*, *loyal*, *playful*). On the other hand, meaning can depend on the semantic network in which a concept is embedded. In such a network, all classes of concepts – including nouns and adjectives – are represented uniformly by nodes (e.g., Collins & Quillian, 1969).

In the present study, we examined the graded membership structure of adjective categories, an often-studied aspect of noun categories. Adjective categories, just like noun categories, show a reliable graded structure. This finding generalizes previous results for nouns (e.g., Rosch & Mervis, 1975) and verbs (e.g., Pulman, 1983, Plant, Webster, & Whitworth, 2011). Moreover, we found that a similarity-based model was able to account for the graded structure. Similar models have been shown successful at explaining graded structure in different types of noun categories (e.g. Hampton, 1979; Rosch & Mervis, 1975; Voorspoels et al., 2008). In particular, the model relates typicality of a certain adjective for its semantic domain (that is, its category) to the extent it is similar to other members of the domain. The more similar an adjective is to other adjectives of the category, the more it is judged typical of the category. Furthermore, the model implements a flexible notion of similarity by allowing differential weighting of underlying dimensions of the similarity structure.

Interestingly, while valence proved a crucial organizing principle in the similarity structure of the majority of adjective categories, we found that it contributed little to the model-based account of the observed graded structure. Our findings thus only partly converge with an earlier an claim that, in contrast to nouns and verbs, the representation of adjectives is primarily organized around relative polar pairs such as *big–small* or *clean–dirty* (Bierwisch, 1967; Landau & Gleitman, 1985). However, regardless of valence, discounting a major source of similarity structure represents a strong difference with nouns, where there is no evidence so far that the primary dimension of variability in terms of similarity (such as the size or ferociousness of animals) is systematically discounted. In fact, valence even masks a similarity-based explanation of graded structure in adjectives, due to its dominance in the similarity structure. Without considering dimensional weighting, a similarity based approach predicts that adjectives half way in between the relative polar pair (for example '*moderately sized*' for the *big-small* pair) are more typical for the corresponding category (*description of a quantity*), since they are presumably most similar to all members in a domain (both the members referring to *big* and the members referring to *small*). However, it is obvious that both *big* and *small* are far more typical of the corresponding category than is *moderately sized*. The GCM escapes this conundrum by collapsing the valence dimension.

Comparison to Other Accounts

Some category exemplars are judged more typical than others, regardless of whether the categories comprise nouns, verbs, adjectives or even well-defined entities such as odd numbers (Armstrong, Gleitman, & Gleitman, 1983). As such, graded structure appears to be a universal property of categories (Barsalou, 1983). However, graded structure needs not be tied to an underlying similarity-based structure in all these categories.

Availability. According to the concept accessibility view (e.g., Hampton & Gardiner, 1983, Janczura & Nelson, 1999), typicality judgments do not depend on similarity between category members but reflect how easy it is to retrieve category exemplars due to factors such as the pre-existing associations with the category, word frequency, or familiarity with the exemplar. When the exemplars that are studied correspond to adjectives, similar factors might explain a graded structure. First, some researchers have proposed that category and instance dominance might explain typicality effects (e.g. Larochelle & Pineau, 1994, Larochelle, Richard, & Soulières, 2000). Category dominance refers to the frequency with which a superordinate is produced in response to a category member, while instance dominance refers to the frequency with which an instance or category member is produced in response to a category label. However, both these variables have little explanatory power, because they are often seen as a variant of the typicality effect itself to be explained (Storms, De Boeck & Ruts, 2000). Other measures of availability of availability do not suffer from this circularity.

Previous studies with noun categories have shown that typicality increases as exemplars become more familiar (e.g., Hampton & Gardiner, 1983; Malt & Smith, 1982). To investigate if this explains the typicality structure of the adjective categories, we correlated word frequency and familiarity with the typicality judgments. Word lemma frequencies were obtained from the CELEX lexical database (Baayen, Piepenbrock, & van Rijn, 1993), while familiarity ratings were taken from Verheyen et al. (2012). Neither log-transformed lemma frequency, nor subjective frequency measured by familiarity correlated significantly with the typicality ratings of any of the categories.

Polysemy. A second alternative explanation is based on the observation that adjectives have multiple related senses. van Dantzig et al. (2011), for instance, showed that an adjective like

plain might be interpreted as a predominantly visual or a predominantly gustatory property, depending on the item it modifies (*fabric*, respectively, *food*). Since their meaning differs depending on the nominal context, one could assume that highly polysemous adjectives would be perceived to be less typical. To investigate if polysemy affects the judgment of typicality, we collected a direct measure of polysemy. This was done by counting the number of senses for each adjective by checking them against the senses listed in the 14th edition of the Dutch Van Dale dictionary (den Boon, Geeraerts, 2005). The number of senses hypothesis was confirmed for a single category only, *description of someone's character, r*(30) = -.44, *p* <.05. No other correlations were significant. These findings suggest that neither the number of contexts in which adjectives occur nor the number of senses they have systematically affects the perceived typicality.

Altogether, none of the factors we investigated provided a full account of the graded structure, neither in terms of the strength of the correlation nor in the scope of categories under investigation. These results corroborate the earlier findings by Larochelle, Richard and Pineau (2000). They used a categorization task and found that well-defined categories such as *seasons, numbers,* or *part of the human anatomy* fail to show a typicality effect after other factors are taken into account. Importantly, in their studies a dissociation was found with natural categories where a similarity-based category structure determined typicality, even when a number of other factors are controlled. The current findings for adjectives are therefore in line with a similarity-based explanation of graded structure common to natural categories.

Semantic Structure

Apart from demonstrating a graded structure in adjectives and providing a model that captures such a structure, our results also confirm that a specific form of antonymy based on valence is arguably the most important structural factor that determines the semantic representation of adjectives (Gross & Miller, 1990; Murphy & Andrew, 1993). Inspection of the similarity spaces shows that valence differences often align with antonomy. However, in contrast to antonomy, valence seems to indicate a more stable property of adjectives that does not depend on the organization of a specific category class. To illustrate this, we obtained a MDS solution using a similar procedure as presented earlier but now using the entire set of adjectives. A satisfactory MDS solution with stress lower than 0.10 was found in 10 dimensions. After centering these dimensions and the predictor variable (valence), the regression resulted in a beta-weight or correlation of $r(294) = -.88$, $p < .001$ for the first dimension. Moreover, this model as a whole captured valence extremely well, $R^2 = .83$. The MDS solutions point towards valence as a strong semantic factor involved in the representation of adjectives, both for specific adjective categories as well as the adjective domain as a whole, and provide empirical support for previous theoretical claims.

Since we extracted between 4 and 6 dimensions for each category, and valence corresponded to only one or two of these dimensions, other factors than valence affect the organization of adjectives as well. To investigate this possibility we employed additional ratings for the semantic attributes of arousal and intensity (described in Verheyen et al., 2012). Arousal was defined as the degree to which a word evokes tension, while intensity was defined as the degree to which certain words such as *strong* or *heavy,* or *small* and *soft* are experienced as powerful or intense. This choice of attributes was influenced by the work of Osgood on the attitudes present in human judgments of words and phrases (Osgood, Suci, & Tannenbaum, 1957) and studies that differentiate between valence and arousal using emotion adjectives (Grühn &

Smith, 2008). While we did not use bipolar scales as Osgood did and focused on different concepts, the scale of valence could be considered analogous to Osgood's evaluation factor, arousal with the activity factor and intensity with potency.

In contrast to earlier findings for emotion adjectives (e.g., Kensinger & Corkin, 2004; Lewis, Critchley, & Dolan, 2007), the arousal ratings did not differentiate strongly from the valence ratings in our categories. A high correlation was established between both variables, ranging from *r(30) = −.79, p < .001* to *r(30) = . -.94 p < .001* for all categories except *description of a quantity, r(30) = .15, ns, shape of an object, r(30) = -.58, p < .001* and *description of a landscape, r(30) = -.65, p < .001.* Moreover, using a similar property fitting procedure as described previously, we found both attributes to correspond to the same dimensions[6] but with smaller multiple correlation coefficients for arousal for nearly all categories (*with feel of an object being the exception* (arousal $R^2 = .44$ compared to $R^2 = .35$ for valence).

Intensity was present in only a few categories (notably the personality related ones) and resulted in multiple correlation coefficients much lower than valence or arousal. For these categories, intensity correlated relatively higher to different dimensions than the other attributes. To illustrate the case for *description of a mood* and *description of a character trait,* we fitted the intensity attribute alongside the valence attribute and show the result in Figure 1a and 1b. Figure 1a shows a second dimension spanning moods going from *rude, aggressive, furious* at the positive side of the second dimension to *melancholic, gloomy* and *dreamy* at the negative side. For this dimension, intensity fits rather well (dotted line) and confirms our intuition of its meaning. A similar interpretation can be made for the *description of a character trait.* Figure 1b shows the two dimensional solution for this category with again a first dimension corresponding to valence and a second one to intensity.

Since the results for arousal were similar to those for valence but accounted for less variance and intensity distinguished itself only for personality related categories, a conclusive interpretation of the underlying similarity space in terms of these two factors seems preliminary. Valence is the only attribute that consistently determines the similarity representations of the adjective categories under consideration.

Similarity-based graded structure in lexico-semantic models.

Since Rosch and Mervis' (1975) seminal paper on family resemblances, it has been shown for a wide range of nouns that refer to natural categories of objects that the similarity structure in the environment is a strong determinant of the membership structure of these categories. In short, typicality is strongly related to the co-occurrence of certain physical features in objects of the same category. This isy isa key aspect inof our ability to classify the environment, and by extension, judge the representative of members of a category: A raven can fly and builds nests in trees, two features that are highly correlated in our environment and characterize the category of birds. A penguin is a rather atypical bird because of the lack of these two properties (a rare thing among birds).

This finding has been crucial for the development of similarity-based models of categorization such as the GCM, which essentially assume the correspondence between feature correlation in the world and the mental representation of categories.

Interestingly, the GCM also seems to stand its ground in the present context, in which the input - ; that -is, the similarity structure of the adjectives -; is -not derived from the perceptual features in the physical environment, but usesis based on word co-occurrences derived from the linguistic environment. The word association modelapproach is an example of a class of lexico-semantic

models in which the notion of similarity hinges on the lexical context in which a word occurs. One of the core assumptions of these models is the idea that humans are sensible to ~~weak~~ correlations in the lexical environment. In this sense, the representations of verbs, nouns and adjectives have a common origin, i.e., their co-occurrence with other words. Moreover, one can expect some degree of correspondence between the lexical system and physical world in terms of co-occurrence of words in language use and co-occurrence of features in the physical world: If language ~~was~~were completely different from the world, it ~~may~~would most not be ~~as~~ useful. Obviously, lexical co-occurrence is not the only factor that determines the language context ~~representation~~, and non-lexical~~, perhaps embodied~~ information comes into play as well, but the usefulness of ~~this~~the lexico-semantic system in activities that involve language might be more pervasive than recently assumed (cf. Louwerse, 2010). Inputting lexical co-occurrences, that is, different contexts in which the adjectives are used, into the similarity-based cognitive model – in much the same way as co-occurrences of features are traditionally fed into the models in noun categories – clearly leads to promising results.

A number of studies by Louwerse and colleagues support the notion that the lexico-semantic system is pervasive in language use. This system can predict the modality of words (Louwerse & Connell, 2011), infer the social relationships between people (Hutchison, Datla & Louwerse, 2012), and predict the location of places on maps (Louwerse & Benesh, 2012). In this sense, it is not surprising that the word associations encode valence of words as well, since this dimension is plausibly of great importance in differentiating between different lexical contexts. As suggested by Glaser (1992), it is likely that in many daily activities such as reading, access of one system might be bypassed in favor of the other, in the same way one does not resort to apple counting when doing more complex summation. Importantly, both a language-based similarity and world-derived similarity both compatible with the ideas of Rosch (1977, 1978) on how we structure the

world around us in a non-arbitrary way. As such, the membership structure in nouns and adjectives can be attributed to the same underlying mechanism.

Conclusion

The results of the present study strengthen the view of a closely-knit relationship between nouns and adjectives by showing that a graded structure is not only present in noun categories, but in a wide variety of adjectives categories as well. Furthermore, a similarity-based approach can account for the graded structure in adjective categories just as it does for nouns. While valence clearly is an important organizing principle of the similarity structure of adjectives, it is for a large part disregarded in the membership structure of categories of adjectives.

References

Armstrong, S. L., Gleitman, L. R. & Gleitman, H., (1983). What some concepts might not be. *Cognition*, ~~13,~~ *13*, 263-308

Author Note

Footnotes

1.   Example from Calvin and Hobbes

2.   Antonymy has a more restricted meaning than opposition (e.g., Cruse, 2004) and we shall argue later on that a specific type of opposition based on valence is an important determinant of adjective representation.

3.   All data are available as a downloadable file from http://ppw.kuleuven.be/concat/.

4.   The range of these ratings was similar to the range in typicality values found for the 12 natural kind noun categories, which varied between 4.64 for *vehicles* and 5.66 for *musical instruments* (De Deyne et al., 2008).

5.   We also considered related models. A first model that we considered was a central prototype model (e.g., Minda & Smith, 2010, Voorspoels et al., 2008). According to this model, the typicality of an adjective is derived by comparing it to an average representation of the category exemplars. The pattern of results of this model was very similar to the GCM results and only differed in terms of the absolute correlations that were achieved, which were slightly lower, on average .71 for the prototype model compared to .73 for the exemplar-based model.

6.   We can think of at least two factors that might explain the lack of independence between arousal and valence. A first explanation is based on the differences between instructions in the rating task. For reasons of consistency with other rating procedures in Verheyen et al. (2012) 7-point rating scales were used, in contrast with studies that used a Manikin rating scale (e.g., Bradley & Lang, 1994). Second, our adjective categories cover a wide variety of properties instead of focusing on emotion words only.

Figure Captions

Figure 1. *Property fitting for Valence and Intensity (cfr. General Discussion) for (a) the categories description of a mood (showing the first two dimension with K = 4 and stress = 0.069) and (b) description of a character trait (showing the first two dimensions, with K = 5 and stress = 0.087). Larger crosses indicate more typical adjectives.*