Title
Graded structure in adjective categories

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 33(33)

ISSN
1069-7977

Authors
De Deyne, Simon
Voorspoels, Wouter
Verheyen, Steven
et al.

Publication Date
2011

Peer reviewed

Graded structure in adjective categories

Simon De Deyne (simon.dedeyne@psy.kuleuven.be)[a]
Wouter Voorspoels (wouter.voorspoels@psy.kuleuven.be)[a]
Steven Verheyen (steven.verheyen@psy.kuleuven.be)[a]
Dan Navarro (daniel.navarro@adelaide.edu.au)[b]
Gert Storms (gert.storms@psy.kuleuven.be)[a]

[a] University of Leuven, Department of Psychology, Tiensestraat 102, 3000 Leuven, Belgium
[b] University of Adelaide, School of Psychology, 5000 Adelaide, Australia

Abstract

In contrast to nouns, little is known about the graded structure of adjective categories. In this study we investigate if adjectives categories show a similar graded structure and explain it using a similarity-based account. The results show a reliable graded structure which is adequately explained by two formal models based on prototypes and exemplars, therefore generalizing the model performance from concrete nouns to adjectives. Finally, we show the attention weights of these models deal with the additional challenge of antonymy in adjectives and discuss the findings in the light of alternative accounts that do not rely on item-similarity.

Keywords: concepts; adjectives, typicality, antonymy, GCM, prototypes.

Research on natural language concepts generally focuses on object categories; that is, words, most often nouns, that refer to concrete objects such as *cats*, *bat*, *apples* and *coconuts*. Category members vary in typicality, with *cats* being more typical mammals than *bats* for example, and one of the most robust effects across a wide range of tasks is a processing advantage for typical items. This effect is thought to reflect a graded category structure, with the gradations corresponding to the degree that concept features are shared between category members (Rosch & Mervis, 1975).

Surprisingly, no attempt has been undertaken to establish whether a similar structure exists for adjectives. Filling this void is crucial, since understanding adjective meanings is central to understanding natural language representations as a whole. It is natural to expect some similarities to exist: adjective and noun concepts are intimately related, since adjectives convey information about the shape, color, taste, etc. of the nouns they modify and nouns in their turn restrict the intended sense of the adjectives. Moreover, it is often adjectives that are used in models of representation of object categories to denote features. With this in mind, we aim to tackle the adjective domain in a way that is similar to the way object categories have been approached.

Despite the similarities, the shift from nouns to adjectives poses new challenges. First, adjectives differ from nouns because there is no hyponymic relationship. In other words, 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 to be a superordinate. Second, most adjectives are highly polysemous. Their meaning often depends on the context in which they are used. *Small*, for example, has a different meaning when it modifies *elephant* than when it modifies *fly*. Third, 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. Finally, and perhaps most challenging for traditional models of representation, is the presence of antonomy: the basic relation of many adjective pairs appears to be one of opposition (e.g., *fast-slow*, *hot-cold*)[1]. It is not immediately clear how antonomy can be reconciled with theories of category representation that rely on a summary representation such as prototypes. For instance, if we assume that *hot* and *cold* are a-similar, but both typical adjectives to describe a temperature, it is not clear how to derive an average representation of this category representation in which a different word to describe temperature can be highly typical and similar to both *hot* and *cold* at the same time.

The aim of this study is to investigate the extent to which the graded structure that exists with nouns is replicated in adjectives. To this end, we apply well-studied formal models of typicality that have been previously used for noun categories. In the following sections, we first describe categories for adjectives that can be used to study graded structure. The findings of this task are used to obtain a direct measure of graded structure by asking participants to judge the typicality or goodness of the exemplar for a category. These data allow us to answer the question whether a reliable graded structure can be identified. Given the presence of such a structure, we look for possible accounts. According to the structural view, the category gradient is a consequence of the similarity of the items in the set. We test this view by implementing both an exemplar and prototype model to account for the obtained structure and investigate how they handle antonymy relationships. We end with a brief discussion on other, non-similarity based approaches and possible lines of further improvements to the present approach.

[1] Antonymy has a more restricted meaning than opposition (e.g., Cruse, 2004), but for the current purposes, we will make no further distinction.

Experiment 1: Categories of Adjectives

Since it is not immediately obvious how adjectives are organized in categories, a number of categories will be used, aiming to cover psychologically interesting domains related to abstract entities, objects, the senses, and person and emotion properties. These organizing principles were chosen to capture a large variety of categories in terms of the nouns they can modify and their abstractness.

Consistent with previous work using nouns (e.g., Ruts et al., 2004), we propose a method where participants generate exemplars for each of the categories, as a method for approximating a natural language category. To reduce the intrusion of qualitative judgments (e.g., *good* and *bad*), specific modifiers were used. For example, adjectives for *description of a work of art* were used instead of simply asking subjects to generate adjectives used for artworks. Since adjectives are often polysemous, 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 generated adjectives to describe the *shape* of an object or the *appearance* of a person.

Method

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

Stimuli and materials The stimuli consisted of twenty-two adjective category descriptions that can be used to describe 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*), sensorial properties (*description of music*, *description of the taste of food*, *the color of objects*, *temperature*, *the feel of an object*) or 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*).

Four different subsets of these descriptions were constructed, each consisting of 11 categories. Two restrictions were imposed, to prevent related categories co-occurring in the same permutation. 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 Every participant filled in an Excel-file containing 11 sheets. Each sheet consisted of the name of a category and 24 blank lines for filling in words. Participants were instructed to generate as many adjectives as they could think of in the allotted space. Two examples were given for the category of adjectives used to describe buildings and phases in life. The examples included both adjectives with positive and negative connotations (e.g., *young*, *old*).

Results and discussion

The responses for the different participants were summarized by tabulation for each category. The counts indicated considerable variation among the number of adjectives generated for a specific category, ranging from 183 for *degree of certainty* to 415 for *color of an object*. The count distributions of the types were positively skewed with on average 65% of the types being generated more than once. The results also showed that out of 1918 adjective types, 594 adjectives were generated for more than one category. This indicates the polysemous nature of these words.

Despite the instructions asking for descriptive adjectives for certain categories, qualitative judgments such as *good* and *bad* occurred in many categories such as *description of weather conditions* or *description of the taste of food*.

Experiment 2: Typicality of Adjectives

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 the previous task 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 exemplars (e.g., *departure from a norm*) or categories containing an adjective in its name (e.g., *degree to which something is difficult or hard*) were also not included. For each category, 30 adjectives were sampled to cover the entire range of the production frequencies[2]. Note that some adjectives like *good* and *bad* were included in multiple categories.

Procedure The participants were asked to fill in an online questionnaire in which they were asked to indicate how good an example each adjective was of the category mentioned on the screen. This was done by providing ratings 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.

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.

[2] All data are available upon request.

The selection of category members across the range of the generation frequency distribution resulted in categories with variability in the degree in which a graded structure was present (mean $SD = .77$, minimum $SD = .56$ for *description of someone's character* and maximum $SD = 1.30$ for *color of an object*). 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 average correlation was $r = .57$ and ranged between $r_{min} = .29$ (*description of a mood*) and $r_{max} = .80$ (*shape of an object*). Although somewhat lower in strength, this confirms previous findings that show a positive correlation between the number of times an item is generated as a category member and typicality (Barsalou, 1985).

Modeling Adjective Categories

The experimental data clearly show that a reliable, graded structure exists for these concepts. To model this structure, we consider two prominent theories that have been applied successfully to object categories. The exemplar view states that a category is represented as a collection of all its previously encountered members (e.g., Medin & Schaffer, 1978). According to the prototype view on the other hand, categories are represented by an abstract summary or a prototype (e.g., Rosch, Simpson, & Miller, 1976). We will briefly describe the models we used to implement these fundamental views on representation and their account of typicality.

Exemplar and prototype models

In this study, we make use of the Generalized Context Model (GCM; Nosofsky, 1986) and the Central Prototype Model (CPM; Minda & Smith, 2010). Both are similarity based models in which the psychological distance between two items i and j is given by

$$d_{ij} = \left(\sum_{k=1}^{K} w_k |x_{ik} - x_{jk}|^r \right)^{\frac{1}{r}} \quad (1)$$

where x_{ik} and x_{jk} are the coordinates of exemplars i and j on dimension k, w_k is the attention weight granted to dimension k and K is the number of dimensions. In this study, the parameter r was fixed at 2 to correspond to Euclidean distances which are more appropriate for integral dimensions (Shepard, 1964). The similarity is then defined:

$$\eta_{ij} = \exp(-cd_{ij}) \quad (2)$$

where c is the scaling parameter which determines the contribution of similarity as points become more distant in space. In the exemplar model, the typicality of an item is assessed by summing the items similarity to all other exemplars. In the prototype model, typicality is assessed by measuring the similarity to a "prototypical" item located at the centroid which is calculated by averaging across the K coordinates of all category members. For both models, the free parameters consist of $K - 1$ dimension weights and the parameter c.

Constructing a psychological space

Feature-based similarity measures have been very successful in predicting conceptual data including typicality judgments (Dry & Storms, 2009). Since adjectives often correspond to concept-features, standard feature listing tasks used for concrete nouns cannot be applied. However, previous studies have shown that word association data captures the semantic representation among a wide range of concepts well (De Deyne, Peirsman, & Storms, 2009). To derive a similarity space from word associations, updated norms of the data described in De Deyne and Storms (2008) were used. The meaning of each adjective is represented by the association response distribution which encodes the number of time 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. The multidimensional representations underlying the GCM and the CPM were constructed for each of the twelve categories by varying the number of dimensions from 2 to 6. Following Kruskal (1964), solutions with stress-values exceeding .10 are not considered for further analyses. Using this criterion to select the lowest dimensionality resulted in dimensionalities with a mode of 4 (see Table 1). Two examples of these spaces using 2 dimensions are shown in Figure 1.

Performance of the models

The GCM and CPM models were fit 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 Table 1. For both the GCM as the CPM, 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. The correlations indicate a slight advantage for the GCM but this was only found to be statistically significant for *a quality judgment*, $Z = 1.90$, $p < .05$ (Meng, Rosenthal, & Rubin, 1992). The general pattern of results was consistent across the tested dimensionality range.

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 both models. If the free parameters in the model are 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 1. 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 ones.

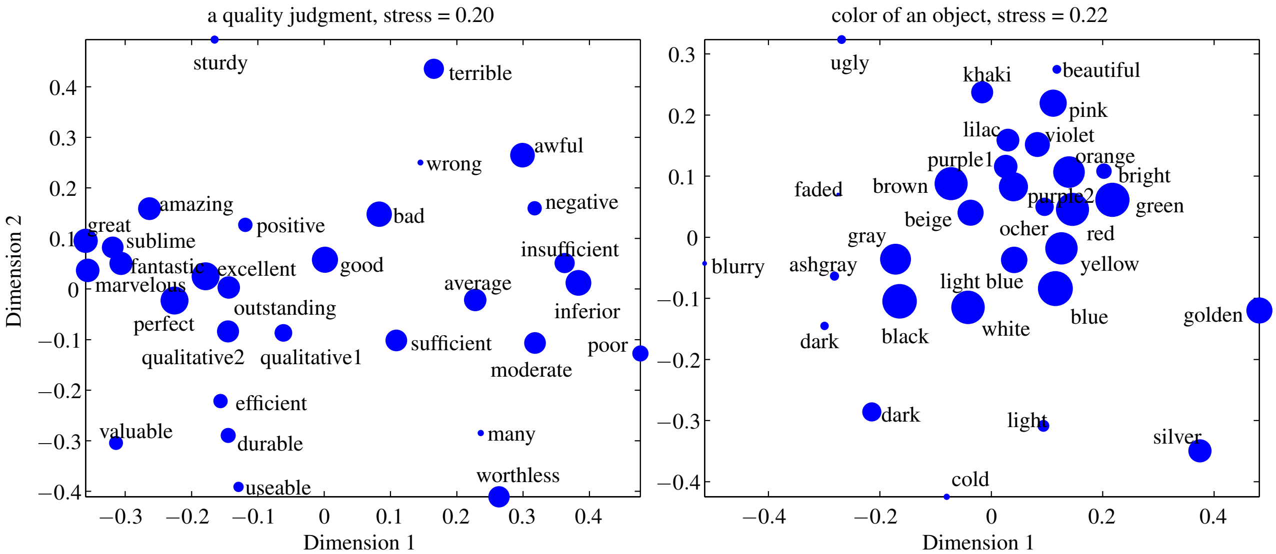

Figure 1: Multidimensional scaling solutions for two adjective categories showing the two most important dimensions. The marker size for each adjective reflects its judged typicality.

Table 1: GCM and CPM correlations based on K dimensional MDS solutions.

$n = 30$	K	stress	GCM	CPM
a quality judgment	4	0.095	.85	.82
a quantity	4	0.096	.78	.73
weather circumstances	4	0.097	.76	.76
landscape	5	0.097	.62	.49
artwork	6	0.086	.78	.76
shape of an object	4	0.099	.77	.64
taste of food	6	0.081	.68	.63
color of an object	5	0.098	.74	.72
feeling of an object	6	0.086	.57	.55
someone's character	5	0.087	.49	.51
person's appearance	5	0.089	.79	.78
mood	4	0.069	.72	.76

Novel insights from the models

For a model to truly be useful, it should not only fit data well; it should also teach us something new about the data. In this case, it is important to consider the values of the parameters to understand how the psychological distances affect the predicted typicalities. Most interesting in this respect are the dimensional weights. Unlike nouns, adjectives are often characterized by the presence of antonymy relationships. In terms of similarity of meaning, words like *valuable* and *worthless* are clear opposites and should be distal in an MDS space. At the same time, these concepts are semantically closely related. They differ in terms of valence but might be similar on all other dimensions. The *quality judgment* category provides the clearest example of antonymy effects. Indeed, as shown in Figure 1, the first dimension distinguishes the adjectives in terms of connotation or valence[3].

The issue of antonymy suggests an interesting application of the CPM and the GCM, since the adjective valence should not affect how typical an adjective is for the category. If the hypothesis is correct, we should expect the attentional weights for a stimulus dimension that corresponds to valence to be *low* in categories characterized by antonymy relations.

To test this hypothesis, we used valence judgments collected in a previous study, in which participants were asked to indicate on a seven-point scale whether a word evoked a negative or a positive feeling (Verheyen, De Deyne, Linsen, & Storms, 2011). As illustrated in the left plot in Figure 1, high typicality ratings can be found at both extremes of the first dimension, which largely reflects valence. Moreover, this valence itself did not correlate with any of the typicality ratings except for *color of an object category* ($r = .52$, $p < 0.1$, two-tailed t). Despite the lack of correlation with typicality ratings, Table 2 shows that the valence ratings tended to correlate most strongly with dimension 1 in the MDS solutions[4], and was significant for 10 of the 12 categories. The comparison between the weight w for this dimension and the dimension with the highest weight identified by the model (see k_{max} and w_{max}, for the dimension number and the weight) shows that the importance of the valence dimensions were down-weighted by the model for all categories except *descriptions of a quantity*. While not necessarily intuitively obvious, the flexibility offered by dimensional attention weights explains why the models can account for antonymy-free typicalities using antonymy-rich similarity representations.

Comparison to other accounts

Typicality can also be due to factors other than those tied to an underlying similarity-based structure. According to the concept accessibility view (e.g., Janczura & Nelson, 1999),

[3]Interestingly, *good* and *bad* and *black* and *white* are not at the extremes of this dimension but much more similar. This is most likely an artifact of the association data, since opposites tend to be strong associates for focal concepts like *good* and *bad*.

[4]Although the statistical model in MDS does not require it, the nature of MDS algorithms tends to ensure that dimension 1 explains the largest share of the variance in the similarity data: this occurred for 11 out of 12 categories.

Table 2: Maximum correlations r between valence ratings and dimensions scores with corresponding dimensional weights w and maximal weights w_{max} found for k_{max} dimensions.

category	n	r	k	w	k_{max}	w_{max}
a quality judgment	12	.83	1	0.01	4	0.75
a quantity	15	.86	1	0.72	1	0.72
weather circumst.	18	.65	1	0.18	3	0.28
landscape	14	.80	1	0.38	2	0.62
artwork	20	.66	2	0.08	1	0.37
shape of an object	24	–	–	–	4	0.47
taste of food	19	.55	2	0.28	1	0.58
color of an object	21	.78	1	0.02	3	0.97
feeling of an object	21	–	–	–	5	0.59
someone's character	16	.95	1	0.07	2	0.54
person's appearance	22	.58	2	0.03	3	0.56
mood	18	.96	1	0.39	3	0.55

typicality judgments do not depend on shared features or similarity between category members but reflect how easy it is to retrieve category members due to factors such as the pre-existing associations with the category, word frequency, or familiarity with the concept. When the concepts that are studied correspond to adjectives, similar factors could explain a graded structure.

One of these alternative explanations is based on the observation that adjectives like *cold* or *bitter* have multiple related senses. Since their meaning differs depending on the nominal context, one could assume that highly polysemous adjectives would be perceived to be less typical. As part of a large scale norming study on adjectives, Verheyen et al. (2011) asked participants to indicate in how many different contexts an adjective could be used. These context variability ratings were correlated with rated typicality, in order to investigate the role of polysemy. Context variability did not correlate significantly with the typicality ratings of any of the categories (all $p > 0.5$, one-tailed t) except for *shape of an object*, $r = -.34$, $p < 0.5$.

Previous studies show that typicality increases as exemplars become more familiar (e.g., Hampton & Gardiner, 1983; Malt & Smith, 1982). To investigate if this is the case, we also 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. (2011). Both the log-transformed lemma frequencies and subjective frequency measured by familiarity did not correlate significantly with the typicality ratings of any of the categories (all $p > .05$, one-tailed t).

A final possibility is that typicality simply reflects how often a certain adjective is used as a modifier of the category concept. There are at least two problems with this view. First, since the category concepts are defined at a more abstract superordinate level, it is not clear how to retrieve pre-existing associations between the category description and the adjectives for the majority of the categories used in this study. For instance, in describing a person's appearance, one could count how many times certain adjectives are used to modify the noun *person* or consider nouns at a more basic level such as *girl* or *cowboy*. Such an approach is still ill-defined since no distinction can be made between other person characteristics such as a person's status, character,etc. Second, generation frequencies are often considered a dependent measure themselves (e.g., Dry & Storms, 2009) since they could depend on the underlying similarity structure of the concepts involved. Although beyond the scope of this paper this topic clearly merits additional attention in future research.

Discussion

Adjective categories, just like noun categories, show a reliable graded structure. This structure was successfully explained using two formal similarity-based models based on exemplars (GCM) and prototypes (CPM). The application of the exemplar and prototype model was not only successful, it also gave us additional insight in the role of antonymy in adjectives.

To further validate our results for adjectives, we compared the performance of the GCM and CPM with the well-studied domain of noun concepts. Voorspoels, Vanpaemel, and Storms (in press) applied the GCM and CPM models to nouns referring to *animals* and *artifacts*. The results in this study showed that the GCM model obtained correlations with rated typicality of .65 for *animals* and .76 for *artifacts*. Moreover, the average correlations of .71 for the GCM and .68 for the CPM model indicate that the present results are both similar in terms of magnitude and relative performance of the models. This is remarkable given the domain differences and the fact that the similarity space in Voorspoels et al. (in press) was derived from participant-generated semantic features restricted to a particular domain, rather than word associations which do not incorporate domain restrictions. Given these encouraging results, we also considered related models. From a theoretical point of view, formal models based on ideals such as recently proposed by Voorspoels et al. (in press) could be of interest given that the adjective classes described here might share some similarity to the ad hoc goal-derived concepts for which an ideal-based account is preferred (Barsalou, 1985). However, preliminary results showed that an ideal-based approach performed worse than the basic GCM or CPM model. Within the same formal framework, the notion of contrasting categories might also prove to be valuable, since it would allow to penalize polysemous adjectives occurring within different domains. Additional preliminary results did not fit the data as well as the basic GCM and CPM models. Future research is needed to uncover why these models fail when applied to adjectives.

Apart from experimenting with models, other improvements might involve deriving the structural representations for adjectives differently. For instance, one could also opt to implement a more text-based model to derive a similarity space. A promising approach is the use of models that derive specific syntactic dependencies from text (Padó & Lapata, 2007). We explored this possibility but found that antonymy relations are not encoded as clearly as in the association data and cannot be discounted as easily by the models. This can be attributed to the fact that adjectives often modify nouns to incorporate new information, in a sense that they indicate a departure from the implicit default values. Thus, people will not often combine the adjective *yellow* with *banana*, while the collocation *red bananas* occurs more frequently than what can be expected from their occurrence in the environment. Moreover, there are other lexical reasons why people prefer some adjective-noun combinations over others. As noted by Cruse (2004) the fact that a combination such as *spotless kitchen* is more acceptable than *spotless taste* does not seem to depend on the meaning of the individual words. These observations support the idea that the success of our approach hinges on the use of association data since these data reflect more of the underlying semantic properties and are less biased by syntactic or lexical constraints.

Ultimately, nouns and adjectives will need to be integrated in a single semantic framework. Our study showed that this framework will have to incorporate flexibility towards antonymy in explaining graded structure of adjectives. While this endeavour provides a challenge for many years to come, it will broaden our understanding of natural concepts including nouns as well.

Acknowledgments

This work was supported by a research grant funded by the Research Foundation - Flanders (FWO) to the first author and by the interdisciplinary research project IDO/07/002 awarded to Dirk Speelman, Dirk Geeraerts, and Gert Storms. Dan Navarro was supported by an Australian Research Fellowship (ARC grant DP0773794).

References

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX lexical database [CD-ROM]*. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 629–654.

Church, K., Gale, W., Hanks, P., & Hindle, D. (1991). Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. In U. Zernik (Ed.), (p. 115-164). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cruse, D. A. (2004). *Meaning in language: An introduction to semantics and pragmatics.* (2 ed.). Oxford: Oxford University Press: Oxford: Oxford University Press.

De Deyne, S., Peirsman, Y., & Storms, G. (2009). Sources of semantic similarity. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th Annual Conference of the Cognitive Science Society* (p. 1834-1839). Austin, TX: Cognitive Science Society.

De Deyne, S., & Storms, G. (2008). Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods*, *40*, 198-205.

Dry, M., & Storms, G. (2009). Similar but not the same: A comparison of the utility of directly rated and feature-based similarity measures for generating spatial models of conceptual data. *Behavior Research Methods*, *41*, 889-900.

Gross, D., & Miller, K. (1990). Adjectives in WordNet. *International Journal of Lexicography*, *3*, 265-277.

Hampton, J. A., & Gardiner, M. M. (1983). Measures of internal category structure: A correlational analysis of normative data. *British Journal of Psychology*, *74*, 491-516.

Janczura, G. A., & Nelson, D. L. (1999). Concept accessibility as the determinant of typicality judgments. *The American Journal of Psychology*, *112*, 1-19.

Kruskal, J. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, *29*, 115-129.

Malt, B. C., & Smith, E. E. (1982). The role of familiarity in determining typicality. *Memory & Cognition*, *10*, 69-75.

Medin, D., & Schaffer, M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207-237.

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, *111*, 172-175.

Minda, J., & Smith, J. (2010). Formal approaches in categorization. In E. Pothos & A. Wills (Eds.), (chap. Prototype models of categorization: basic formulation, predictions, and limitations.). Cambridge University Press: Cambridge, UK.

Murphy, G. L., & Andrew, J. M. (1993). The conceptual basis of antonymy and synonymy in adjectives. *Journal of Memory and Language*, *32*, 301-319.

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

Padó, S., & Lapata, M. (2007, June). Dependency-Based Construction of Semantic Space Models. *Computational Linguistics*, *33*, 161–199.

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

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural Bases of Typicality Effects. *Journal of Experimental Psychology: Human Perception and Performance*, *4*, 491-502.

Ruts, W., De Deyne, S., Ameel, E., Vanpaemel, W., Verbeemen, T., & Storms, G. (2004). Dutch norm data for 13 semantic categories and 338 exemplars. *Behaviour Research Methods, Instruments, and Computers*, *36*, 506–515.

Shepard, R. N. (1964). Attention and the metric structure of the stimullus space. *Journal of Mathematical Psychology*, *1*, 54-87.

Verheyen, S., De Deyne, S., Linsen, S., & Storms, G. (2011). *Lexical and semantic norms for 1,000 Dutch adjectives.* (Unpublished manuscript)

Voorspoels, W., Vanpaemel, W., & Storms, G. (in press). A formal Ideal-based Account of Typicality. *Psychonomic Bulletin & Review*.