Behav Res (2016) 48:1644–1652
DOI 10.3758/s13428-015-0680-2

Estimating affective word covariates using word association data

Bram Van Rensbergen[1] · **Simon De Deyne**[1,2] · **Gert Storms**[1]

Published online: 28 October 2015

Abstract Word ratings on affective dimensions are an important tool in psycholinguistic research. Traditionally, they are obtained by asking participants to rate words on each dimension, a time-consuming procedure. As such, there has been some interest in computationally generating norms, by extrapolating words' affective ratings using their semantic similarity to words for which these values are already known. So far, most attempts have derived similarity from word co-occurrence in text corpora. In the current paper, we obtain similarity from word association data. We use these similarity ratings to predict the valence, arousal, and dominance of 14,000 Dutch words with the help of two extrapolation methods: Orientation towards Paradigm Words and *k*-Nearest Neighbors. The resulting estimates show very high correlations with human ratings when using Orientation towards Paradigm Words, and even higher correlations when using *k*-Nearest Neighbors. We discuss possible theoretical accounts of our results and compare our findings with previous attempts at computationally generating affective norms.

Keywords Word meaning · Word associations · Concepts · Psycholinguistics · Semantics

✉ Bram Van Rensbergen
mail@bramvanrensbergen.com

[1] Laboratory of Experimental Psychology, University of Leuven, Tiensestraat 102 B3711, 3000 Leuven, Belgium

[2] Computational Cognitive Science Laboratory, University of Adelaide, Adelaide, SA, Australia

Introduction

Emotionally charged concepts are processed differently to emotionally neutral concepts. This intuitive idea is supported by research in multiple domains, including brain imaging (Lane, Chua, & Dolan, 1999; Lang et al., 1998; Maddock, Garrett, & Buonocore, 2003; Mourão-Miranda et al., 2003), semantic categorization (Moffat, Siakaluk, Sidhu, & Pexman, 2015; Newcombe, Campbell, Siakaluk, & Pexman, 2012; Niedenthal, Halberstadt, & Innes-Ker, 1999), affective priming (Fazio, 2001; Klauer, 1997), word associations (Cramer, 1968; Isen, Johnson, Mertz, & Robinson, 1985; Johnson & Lim, 1964; Matlin & Stang, 1978; Pollio, 1964), or word recognition reaction times (De Houwer, Crombez, Baeyens, & Hermans, 2001; Kuperman, Estes, Brysbaert, & Warriner, 2014).

Research on the emotional aspect of words traditionally makes use of three dimensions: (1) valence or evaluative attitude, generally rated on a good/bad or happy/unhappy scale, (2) arousal or activity, often represented on an active/passive scale, and (3) dominance or potency, usually expressed on a strong/weak or dominant/submissive scale. The importance of these dimensions was first described by Osgood, Suci, and Tannenbaum (1957). In an undertaking to quantify connotative meaning, they performed a factor analysis on a large number of verbal judgments of a wide variety of concepts and found that most of the variance in emotional assessments was accounted for by these three affective dimensions. Subsequent research has replicated these findings across dozens of cultures (see Heise, 2010, or Osgood, 1975, for an overview), indicating that the importance of these factors may be near universal.

Word values on these dimensions are commonly used both for investigating the influence of affective meaning on some other aspect, and to control for a possible confounding effect of the emotional charge of stimuli. As such, it is not surprising that there is a high demand for databases with affective norming data.

Traditionally, these norms are obtained by asking participants to rate a large number of words on each dimension. This procedure can be very expensive and time-consuming, as multiple persons have to rate each word in order to arrive at reliable measures (by means of average ratings). As a result, most norming databases are rather limited in the number of different words they contain, making generalizations towards the entire lexicon somewhat unfeasible. For example, the original Affective Norms for English Words (ANEW) dataset, likely the most frequently used norms, contains "just" 1,034 unique words (Bradley & Lang, 1999). Despite the cumbersome nature of gathering ratings word by word, some researchers have recently managed to construct a much more comprehensive English database, containing norms for 13,915 words (Warriner, Kuperman, & Brysbaert, 2013). Affective rating datasets in other languages are not nearly as extensive, such as in Dutch (4,300 words: Moors et al., 2013), Finnish (420 words: Söderholm, Häyry, Laine, & Karrasch, 2013), French (1,031 words: Monnier & Syssau, 2014), German (2,900 words: Võ et al., 2009), Italian (1,034 words: Montefinese, Ambrosini, Fairfield, & Mammarella, 2014), Spanish (1,034 words: Redondo, Fraga, Padrón, & Comesaña, 2007), Polish (1,586 words: Imbir, 2015), or Portuguese (1,034 words: Soares, Comesaña, Pinheiro, Simões, & Frade, 2012).

Estimating affective ratings using word co-occurrence data

As the procedure of having participants rate words manually is both expensive and time-consuming, there has been some interest in deriving affective norms from other sources of information. One approach that has been suggested starts by deriving similarity measures for large numbers of words using their position in text corpora. For any given word in the corpus, norm ratings are then estimated using that word's similarity to a number of words for which affective values are already known. This approach could lead to norming datasets significantly larger than those gathered using manual ratings, as large text corpora are available in many languages.

Two implementations of this technique have been put forward. A first approach makes use of latent semantic analysis (LSA; Landauer & Dumais, 1997), which quantifies the degree to which words are associated based on the assumption that similar words occur in similar pieces of text. LSA starts from a word by context matrix, where each cell contains how frequently that word occurs in that chunk of text (e.g., sentence, paragraph, or document). To diminish the influence of highly frequent words, a weighting function is applied to this matrix. Subsequently the most important dimensions (usually 300) are extracted from this matrix using singular value decomposition, yielding a relatively low-dimensional approximation of the original matrix. The similarity between any two words is then defined as the cosine of the angle between their corresponding row vectors in this new matrix. As a result, LSA can estimate the similarity between two words that never occur together, but do co-occur in similar contexts.

A second approach to predict similarity from text corpora makes use of pointwise mutual information (PMI: Church & Hanks, 1990; see also Bullinaria & Levy, 2007; Manning & Schütze, 1999), which derives relatedness from direct word co-occurrence rather than co-occurrence in contexts. Specifically, the PMI of two words x and y is defined as

$$\mathrm{PMI}(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)},$$

where $P(x,y)$ refers to the frequency of x and y co-occurring in some context divided by the total number of tokens in the corpus, and $P(x)$ and $P(y)$ refer to the frequency of x and y, respectively divided by the total number of tokens. Compared with LSA, the most prominent advantage of PMI is scalability, as it can be applied to corpora far larger than LSA can handle. Additionally, it has been suggested that PMI may be more plausible as a model of semantic organization (Recchia & Jones, 2009).

Once pairwise similarity estimates have been derived by applying either LSA or PMI to text corpora, one can estimate words' values on various dimensions using their similarity towards words for which the values on those dimensions are already known.

Turney and Littman (2003) predicted the valence of words using their similarity to a small number of paradigm words, words commonly used to describe very low or very high levels of valence (e.g., *good*, *bad*). They compared the predictions of this approach with binary manual ratings (words rated positive or negative) for 3,596 English words, and report a correlation of .65 when using similarity derived from LSA (on a corpus comprising 10 million tokens), and between .61 (corpus containing 10 million tokens) and .83 (corpus containing 100 billion tokens) when using similarity derived from PMI.

Bestgen and Vincze (2012) employed a somewhat different approach. Rather than examine a word's relation to a small number of seed words, they looked at its similarity to all words for which norming data exist: they define the estimated rating of words as the average of its k nearest neighbors included in the norming data, with k ranging from 1 to 50. Nearest neighbors were obtained from similarity indices between 17,350 English words, which were calculated by applying LSA to a corpus comprising 12 million tokens. The valence, arousal, and dominance of each of these words was then estimated as the average rating of its k nearest neighbors which were included in the ANEW norms. Note that a given word was never considered as one of its nearest neighbors, that is, predictions were based on a leave-one-out approach. Comparing obtained estimates with the ANEW norms they find the highest accuracy at $k = 30$, with a correlation of .71 for valence, .56 for arousal, and .60 for dominance.

Recchia and Louwerse (2015) used a comparable approach, with a number of differences. They obtained nearest neighbors through similarity measures derived with PMI rather than LSA, which allowed them to make use of a much larger corpus containing 1.6 billion English words. They also tested a wider array of values for neighborhood parameter *k*, with *k* ranging from 2 to 500. Additionally, instead of following a leave-one-out approach, predictions were based on the ratings of one dataset while accuracy was assessed through correspondence to ratings of a second dataset. This revealed correlations of up to .74 for valence (at $k = 15$), up to .57 for arousal (at $k = 40$), and up to .62 for dominance (at $k = 60$).

Finally, Mandera, Keuleers, and Brysbaert (2015) evaluated how the performance of these computational approaches is influenced by the size of available norming data. To that end, the 13,915 words in the Warriner et al. (2013) norms were split into a training set and a test set, using different splits (e.g., 90 %/10 % or 50 %/50 %). Similarity indices between all words were obtained through applying LSA or PMI to a corpus comprising 385 million tokens. These were then used to predict the valence, arousal, and dominance of words, with neighborhood parameter *k* set to 30 (the optimal value described by Bestgen & Vincze, 2012). They find that accuracy is somewhat reliant on the size of available norms. For example, when working with PMI-based similarity, increasing the training sample (i.e., the ratings that can contribute to the estimates) from 10 % of the Warriner norms to 90 % raises the correlation between the test sample and the norm ratings from .61 to .72 for valence, from .37 to .51 for arousal, and from .51 to .61 for dominance. (They also investigated a number of other extrapolation methods, all of which showed a similar or lower accuracy.)

Taken together, these studies indicate that ratings extrapolated from word co-occurrence data show medium to high correlations with human judgments, highlighting the usefulness of this computational approach. Moreover, the size of norming databases constructed using this method is likely to keep expanding in the coming years, as even more word corpora become available. This is especially useful for languages other than English, where existing norming datasets are often quite limited in size.

Word associations as a source of similarity

As we have seen, existing research on computationally estimating norms generally makes use of similarity values derived from word co-occurrences in text corpora. An alternate approach to obtaining similarity ratings is using word association data. In a word association task, participants respond with the first word(s) that come to mind after reading a certain cue word. A key assumption in using word associations to investigate meaning is that the probability of producing a certain response to a cue is a measure of the associative strength between cue and response in the mental lexicon (Cramer, 1968; De Deyne, Navarro, & Storms, 2013; Deese, 1966; Nelson, McEvoy, & Schreiber, 2004). This idea is supported by research on facilitation of word processing in associative priming (Hutchison, 2003), response times in lexical decision tasks (De Deyne et al., 2013), word recognition reaction times (De Deyne et al., 2013; Gallagher & Palermo, 1967; Nelson, McKinney, Gee, & Janczura, 1998), fluency task generation frequencies (Griffiths, Steyvers, & Tenenbaum, 2007), clustering in recall (Wicklund, Palermo, & Jenkins, 1965), and predicting cued recall (Nelson et al., 1998).

To obtain information about relatedness from word association data, one can make use of a cosine measure of similarity (Landauer & Dumais, 1997). While this measure is traditionally applied to spatial models such as LSA, it can also be used in the context of word association data (e.g., De Deyne et al., 2013; De Deyne, Verheyen, & Storms, 2015; Gravino, Servedio, Barrat, & Loreto, 2012). Here, the cosine similarity between two words reflects their overlap in associative links; two words that share no associations have a similarity of 0, while two words with the exact same associative responses have a similarity of 1. Similarity estimates obtained using this approach show a strong correspondence with relatedness judgments (De Deyne et al., 2013; De Deyne et al., 2015).

Research indicates that, compared with approaches based on text corpora, word association data can lead to a more valid measure of semantic relatedness. For example, (human) similarity judgments correlate more strongly with similarity estimates derived from association data than with predictions based on word co-occurrences (De Deyne, Peirsman, & Storms, 2009; De Deyne et al., 2015). Additionally, associative strength has been shown to predict priming effects on a word-level in both lexical decision tasks and naming tasks, while similarity derived from applying LSA to text corpora did not (Hutchison, Balota, Cortese, & Watson, 2008).

In the current study, we propose using word association data to obtain similarity estimates for a large number of words, and subsequently predict words' values on affective dimensions (e.g., valence) using their similarity towards words for which the values on those dimensions are already known (e.g., *pleasant*). Using this approach, we will estimate valence, arousal, and dominance ratings for a large number of words. To verify the validity of these estimates, we will compare them with existing norm ratings.

Method

Materials

To obtain the associative strength for a large set of words, we made use of the Dutch Small World of Words project,[1] which contains 3.7 million word associations collected in response to 14,000 cue words. Each cue was presented to roughly 100 participants, who gave up to three responses per cue (see De Deyne et al., 2013, for full details[2]).

Valence, arousal, and dominance ratings for 4,300 Dutch words were taken from Moors et al. (2013). In this study, words were rated on a Likert scale ranging from 1 (*very negative/unpleasant*, *very passive/calm*, and *very weak/submissive*, respectively) to 7 (*very positive/pleasant*, *very active/aroused*, and *very strong/dominant*). Ratings showed very high split-half reliabilities: .99 for valence, .97 for arousal, and .96 for dominance.

Procedure

We began by computing the cosine similarity (e.g., Landauer & Dumais, 1997) between each combination of the 14,000 cue words in the Dutch Small World of Words dataset. In this context, a cosine measure reflects the extent to which two words overlap in associative responses: two words that share no associations would have a value of 0, while two words with the exact same associative responses would have a value of 1. To obtain this measure, we first constructed a cue-by-cue count matrix, where cells reflected how often each cue was given as an association in response to each other cue. Rows of this matrix were normalized to sum to 1 and log-transformed. Finally, to obtain the cosines between the angles of these vectors, the matrix was multiplied by its transpose. At this point, cells of the matrix contained the cosine similarity between the cues corresponding to their rows and columns.

Subsequently, we used these similarity ratings to predict affective word covariates by applying two extrapolation methods, each of which estimates word's values on affective dimensions using that word's similarity to certain words for which affective ratings are already known.

The first extrapolation method we employed, *Orientation towards Paradigm Words*, predicted a word's valence, arousal, and dominance using that word's similarity towards certain paradigm words, words commonly used to describe extreme values on these dimensions (Kamps, Marx, Mokken, & de Rijke, 2004; Turney & Littman, 2003). Paradigm words were obtained from the instructions in the rating task described by Moors et al. (2013), which yielded two positive and two negative paradigm words for each dimension (Table 1).

At first, Orientation towards Paradigm Words predictions simply reflected the sum of a word's similarity towards both positive paradigm words minus the sum of its similarity towards both negative paradigm words. These estimates were consequentially refined by including the target word's similarities towards the k nearest neighbors of each of the paradigm words, that is, out of the 14,000 words, the k words with the highest similarity towards that paradigm word, where k ranged from 0 to 500. A target word's final score was computed as the sum of its similarity towards both positive paradigm words and the k nearest neighbors of each positive paradigm word, minus the sum of its similarity towards both negative paradigm words and the k nearest neighbors of each negative paradigm word.

The second extrapolation method we applied, k-*Nearest Neighbors*, was very similar to the approach described by Bestgen and Vincze (2012), with the notable difference that our similarity estimates were derived from word association data rather than from word co-occurrence in text corpora. Under this approach, the score of any target word on some dimension is calculated as the mean score of its k nearest neighbors (as assessed with cosine similarity) for which the value on that dimension is known (that is, the k closest words for which human judgments are included in the dataset of Moors et al., 2013), for k ranging from 1 to 500. Note that a target word is never considered as one of its own nearest neighbors; as such, the human judgment of some word does not contribute to that word's extrapolated rating.

It may be important to stress that with the k-Nearest Neighbors approach, k refers to the nearest neighbors of the target word (for which ratings were available), while under the Orientation towards Paradigm Words method, k refers to the nearest neighbors of the various paradigm words.

Results

We estimated the valence, arousal, and dominance of the 14,000 cue words in the Small World of Words dataset with the two extrapolation methods described above. Out of these 14,000 words, 3,872 are comprised in the norms of Moors et al. (2013) and can be used to assess the accuracy of the two methods. These 3,872 words represent 90 % of the 4,300 words in the norms, and 28 % of the cue words in the word association dataset.

The Orientation towards Paradigm Words method predicted the affective values of words using their similarity towards certain paradigm words (see Table 1), and the k nearest neighbors of each paradigm word. The left panel of Table 2 displays the correlations (Pearsons's r) between these estimates and the human judgments described by Moors and colleagues (2013)

[1] See www.smallworldofwords.com

[2] We use a more recent version of this dataset, which is somewhat larger (e.g., comprising 14,000 cue words rather than 12,000) but otherwise similar in all aspects.

Table 1 English translation of the paradigm words corresponding to valence, arousal, and dominance (Dutch source that was actually used)

Dimension	Positive paradigm words (Dutch source)	Negative paradigm words (Dutch source)
Valence	Positive (positief), pleasant (aangenaam)	Negative (negatief), unpleasant (slecht)
Arousal	Active (actief), busy (druk)	Passive (passief), calm (kalm)
Dominance	Strong (sterk), dominant (dominant)	Weak (zwak), submissive (onderdanig)

for valence, arousal, and dominance, for k values ranging from 0 (only the paradigm words themselves are used) to 500 (the paradigm words and the 500 nearest neighbors of each paradigm word contribute to the final estimate).[3] When estimates are based solely on similarity to the paradigm words themselves, we find correlations of .79, .53, and .59 to human judgments of valence, arousal, and dominance, respectively. As we increase the number of neighbors of each paradigm word that contribute to our predictions, these correlations increase to up to .86, .65, and .69 for valence, arousal, and dominance, respectively.

The k-Nearest Neighbors method estimated the valence, arousal, and dominance of the 14,000 words as the mean of the human ratings of its k nearest neighbors included in the Moors et al. (2013) dataset. The right panel of Table 2 displays the Pearson correlation between these estimates and human judgments of valence, arousal, and dominance, for k (the number of neighbors of a target word that contribute to its estimate) ranging from 1 to 500.[4] We find an optimal accuracy at $k = 10$, where the extrapolated ratings show a correlation of .91 for valence, .84 for arousal, and .85 for dominance.

We find that performance of both extrapolation methods shows a curvilinear function with respect to neighborhood parameter k: as k increases, accuracy improves up to a certain point and then starts to decline. This decreased performance at higher values of k is in line both with expectations, as "further" neighbors have a lower similarity to the target word, and with previous research (Recchia & Louwerse, 2015).

A downside of the k-Nearest Neighbors approach is that it relies on an existing set of human judgments. As a result, the number of words for which human ratings are available is certain to have an effect on the accuracy of this method. If only few norms are available, it is possible that some extrapolated values are based on ratings of words that are in fact not particularly close to the target word (if more similar words are not included in the norming dataset), which would certainly have consequences for the validity of those estimates. In Dutch, we have access to 3,872 words in the relatively large norms of Moors et al. (2013); in many languages, databases of this size are not available. To estimate how accurate this extrapolation method would be when only a limited set of norms is available, we followed an approach similar to that of Mandera et al. (2015) by running the k-Nearest Neighbors method restricted to random subsets of the available norming data (at $k = 10$, the optimal value in Table 2). We tested 12 different sample sizes, ranging from 100 words to 3,872 words (the entire dataset). To remove any sampling bias, this procedure was repeated 100 times for each sample size. Figure 1 indicates that even when only a small norming dataset is available, the k-Nearest Neighbors method manages to attain a high accuracy; for exampling, when norms for just 1,000 words are available, the extrapolated ratings show correlations with human judgments of up to .89 for valence, and up to .79 for arousal and dominance.

Finally, we wanted to have an idea of whether having access to a norming dataset larger than that of Moors et al. (2013) would lead to a significant improvement in accuracy. Although we cannot test this notion directly with the data currently at our disposal, we can estimate it by examining the slopes of the lines in Fig. 1. As all three lines keep increasing up to the largest sample size, it seems reasonable to assume that expanding the size of the used norming dataset would result in a small improvement in accuracy, especially for arousal and dominance.

Discussion

We have outlined two methods to computationally estimate subjective norms values. Both methods derive similarity from association data, and predict a word's norms using its similarity towards words for which affective values are already known. The two approaches were used to extrapolate the valence, arousal, and dominance for 14,000 Dutch words; these estimates are available at https://osf.io/pmbvc/.

In comparing the extrapolated norms to human judgments, we find high to very high correlations for all three dimensions. Correspondence is highest for valence, suggesting that compared with arousal and dominance, valence is represented

[3] We also investigated the effect of applying various monotonically decreasing weighting functions to the contribution of the various nearest neighbours of each paradigm word, so the similarity towards further neighbours contributed less to the final score. Somewhat contrary to our expectations, none of these functions led to a significant improvement in the overall accuracy of our approach; as such, these findings are not reported here.

[4] Here, too, we examined the effect of applying different weighting functions to these k values, with further neighbours contributing less to a target word's final score. As with the first extrapolation method, this did not lead to a considerable improvement in accuracy; as such, we will not report these findings here.

Table 2 Correlations between human judgments and estimates derived using the Orientation towards Paradigm Words extrapolation method (left panel) and estimates derived using the *k*-Nearest Neighbors extrapolation method (right panel)

k	Orientation towards Paradigm Words			*k*-Nearest Neighbors		
	Valence	Arousal	Dominance	Valence	Arousal	Dominance
0	.79	.53	.59	-	-	-
1	.80	.54	.60	.85	.76	.76
2	.80	.54	.62	.88	.80	.81
5	.79	.49	.63	.89	.83	.83
10	.81	.56	.67	.91	.84	.85
25	.83	.63	.69	.91	.84	.84
50	.84	.63	.69	.91	.83	.83
100	.84	.67	.68	.91	.83	.82
250	.85	.65	.68	.90	.81	.81
500	.86	.63	.68	.90	.78	.79

Note. $N = 3{,}872$

more strongly in the semantic similarity space. This finding is in line with the importance often attributed to this aspect, both in research on affective meaning (Osgood et al., 1957) and various other domains.

Of the two extrapolation methods we tested, accuracy is highest for the *k*-Nearest Neighbors technique, as would be expected because this method is based directly on the human ratings with which accuracy is assessed (although importantly, the human judgment of a given word does not contribute to that word's extrapolated value). Note, though, that this reliance on human ratings brings with it a huge drawback: the *k*-Nearest Neighbors method can only work when human judgments are already available for some amount of words. In contrast, the Orientation towards Paradigm Words approach does not depend on human judgments in any form, and aside from a selection of paradigm words, only requires similarity indices.

Considering the *k*-Nearest Neighbors method relies on human judgments, its accuracy is likely tied to the quality of available human ratings. As our research was performed in Dutch, we had access to the large norming dataset of Moors and colleagues (2013). In many languages, existing databases are considerably smaller. To assess how accurate our approach

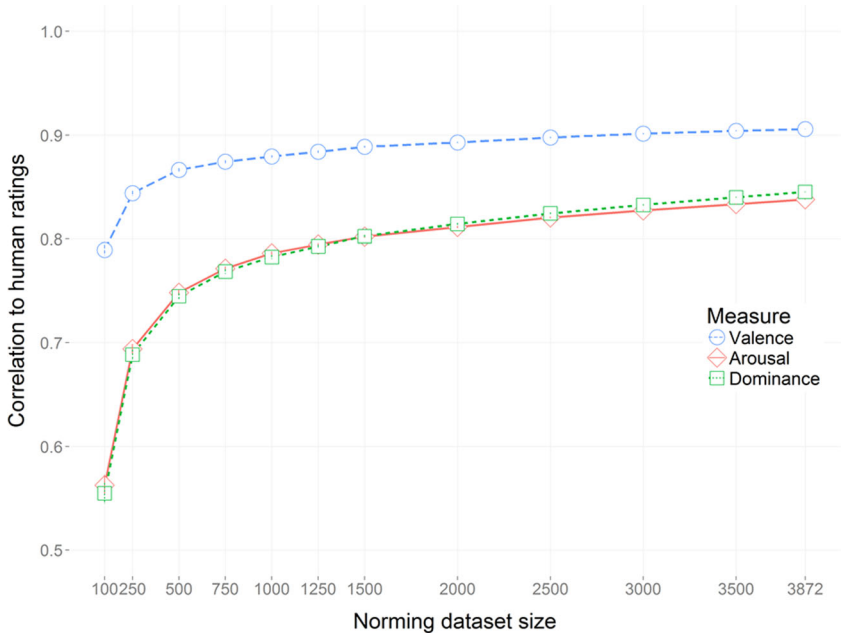

Fig. 1 Relation between accuracy of the *k*-Nearest Neighbors extrapolation method and the size of available norms. Correlations were obtained by averaging across 100 iterations of running the extrapolation method limited to a random subset of human judgments (out of the available 3,872 norm words). Neighborhood parameter *k* was set to 10, the optimal value reported for running this extrapolation method with all human judgments (Table 2). Error bars (very small, due to low error rates) indicate standard error in accuracy across the 100 iterations

is when limited to a smaller set of norms, we ran the *k*-Nearest Neighbors extrapolation method restricted to subsets of the available norming data. Correlations with human judgments were lower than when the method had access to all norming data, but still very high (between .78 and .88 when using a subset of 1,000 words). This suggests that even when only a small set of norms is available, the *k*-Nearest Neighbors method can be very effective at predicting affective word covariates.

In existing research on computationally predicting affective norms, similarity or semantic relatedness is generally derived from word co-occurrence data rather than from word associations. Using these similarity estimates, several studies have extrapolated affective ratings with the help of the same *k*-Nearest Neighbors technique we described. These studies report that their estimates display correlations with human judgments of up to .74, .57, and .62 (Bestgen & Vincze, 2012), up to .71, .56, and .60 (Recchia & Louwerse, 2015), and roughly up to .72, .51, and .61 (Mandera et al., 2015), for valence, arousal, and dominance, respectively.

In comparison, the predictions we present show a much higher accuracy, on all three dimensions. There are several potential explanations behind this improvement. It could be a result of a difference in language: we made use of Dutch associations and judgments, while the described corpus-based studies were performed in English. However, this seems an unlikely explanation, as similar corpus-based research has also been undertaken in French and Spanish, where estimates showed similar or lower correlations with human ratings (Bestgen, 2002, 2008; Vincze & Bestgen, 2011). Furthermore, as the importance of valence, arousal, and dominance is highly generalizable across cultures (Osgood, 1975), there is no a priori reason to expect these aspects to be represented differently in Dutch and English.

A more probable cause for the disparity between our findings and previous attempts at computationally estimating norms is the nature of the information from which similarity estimates were construed: existing research derived relatedness from word co-occurrence in text corpora, while we made use of word association data. Previous comparisons between corpus-based and association-based similarity estimates also report a higher accuracy for approaches reliant on word association data, in line with our findings (De Deyne et al., 2009; De Deyne et al., 2015; Hutchison et al., 2008). This is likely because word associations and text corpora represent information of a different nature. Written language is grounded in pragmatics; the goal is to communicate some discourse efficiently, and information that is known to both parties is often left out. Word associations, in contrast, are non-propositional, and generally free from pragmatics or intent (Deese, 1966; Szalay & Deese, 1978). As a result, mentally central concepts or properties (such as color or shape) are usually well represented in word associations, while they are somewhat uncommon in most written text. An additional asset of word association data is its very high signal to noise ratio, as almost every association reflects a meaningful relation; in contrast, text corpora are often characterized by a low signal-to-noise ratio, negating part of the advantage of scale that characterizes corpus-based approaches.

Taken together, we can conclude that word association data can be a very powerful source of information on semantic relatedness, and suggest that when computationally generating affective norms, an association-based approach may be a worthwhile addition to or substitute for procedures based on word co-occurrence in text corpora.

Of course, this approach does require access to word association data. While gathering word associations is a simple and straightforward procedure, it remains reliant on human participants. As a result, constructing a large dataset of this nature is far from effortless. Luckily, such databases already exist in many languages; for example, the Small World of Words project from which we obtained the Dutch associations also contains datasets in English, German, French, Spanish, Rioplatense Spanish, Vietnamese, Japanese, and Cantonese. Note that in terms of number of tokens, these databases are all much smaller than most text corpora. However, as we have seen, this quantitative shortcoming does not necessarily translate to deteriorated predictions; indeed, human judgments show a considerably higher correspondence to the estimates reported in the current paper, which are based on a dataset comprising 3.7 million tokens, than to the estimates based on word co-occurrence data described previously, which are based on much larger corpora (e.g., the predictions reported by Recchia & Louwerse, 2015, are based on a dataset containing 1.6 billion tokens).

An important caveat when working with computationally estimated word covariates is that even when they show a moderate to high correspondence with human judgments, they could lead to different conclusions than would be reached when using human ratings (Mandera et al., 2015). The data we present are likely somewhat less vulnerable to this issue, as our estimates show considerably higher correlations to human ratings; nevertheless, this is definitely a topic that should be investigated further in future research.

In the current paper, we estimated valence, arousal, and dominance ratings based on similarity values derived from word association data. The extrapolation methods we describe would conceivably work on other psychologically relevant dimensions as well, as long as these dimensions are captured by the associative technique, that is, as long as the associations people give to a certain word are in some way related to the cue's or association's value on that dimension. Existing research suggests that other examples of word covariates that could likely be predicted based on association data may include concreteness (the extent to which words refer to

something perceptible; see Mandera et al., 2015, or Van Rensbergen, Storms, & De Deyne, 2015), age of acquisition (the age at which a word was learned; see Mandera et al., 2015), or dimensions relevant to personality profiles (e.g., openness, conscientiousness, extraversion, agreeableness, or neuroticism; see Yarkoni, 2010, or Park et al., 2015).

Author note This research was supported by Research Grant OT/10/024 from the University of Leuven Research Council. Author contributions: BVR conceived and designed the study, analyzed the data, and wrote the manuscript. GS and SDD conceived and designed the study, and contributed to the writing of the manuscript.

References

Bestgen, Y. (2002). Détermination de la valence affective de termes dans de grands corpus de textes. *Actes de CIFT, 2002,* 1–14.

Bestgen, Y. (2008). Building affective lexicons from specific corpora for automatic sentiment analysis. *Proceedings of LREC, 2008,* 496–500.

Bestgen, Y., & Vincze, N. (2012). Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior Research Methods, 44*(4), 998–1006. doi:10.3758/s13428-012-0195-z

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings*.

Bullinaria, J., & Levy, J. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods, 39*(3), 510–526. doi:10.3758/BF03193020

Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics, 16*(1), 22–29.

Cramer, P. (1968). *Word association*. New York: Academic Press.

De Deyne, S., Navarro, D. J., & Storms, G. (2013). Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods, 45*(2), 480–498. doi:10.3758/s13428-012-0260-7

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

De Deyne, S., Verheyen, S., & Storms, G. (2015). The role of corpus size and syntax in deriving lexico-semantic representations for a wide range of concepts. *Quarterly Journal of Experimental Psychology*. doi:10.1080/17470218.2014.994098

De Houwer, J., Crombez, G., Baeyens, F., & Hermans, D. (2001). On the generality of the affective Simon effect. *Cognition & Emotion, 15*(2), 189–206. doi:10.1080/02699930125883

Deese, J. (1966). *The structure of associations in language and thought*. Baltimore, MD: Johns Hopkins University Press.

Fazio, R. H. (2001). On the automatic activation of associated evaluations: An overview. *Cognition & Emotion, 15*(2), 115–141. doi:10.1080/02699930125908

Gallagher, J., & Palermo, D. (1967). The effect of type associative relation on word recognition times. *Child Development, 38*(3), 849–855. doi:10.2307/1127262

Gravino, P., Servedio, V. D. P., Barrat, A., & Loreto, V. (2012). Complex structures and semantics in free word association. *Advances in Complex Systems*, *15*(03n04). doi:10.1142/S0219525912500543

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211–244. doi:10.1037/0033-295X.114.2.211

Heise, D. R. (2010). *Surveying cultures: Discovering shared conceptions and sentiments*. Hoboken, NJ: John Wiley & Sons.

Hutchison, K. (2003). Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic Bulletin & Review, 10*(4), 785–813. doi:10.3758/BF03196544

Hutchison, K. A., Balota, D. A., Cortese, M. J., & Watson, J. M. (2008). Predicting semantic priming at the item level. *Quarterly Journal of Experimental Psychology, 61*(7), 1036–1066. doi:10.1080/17470210701438111

Imbir, K. K. (2015). Affective norms for 1,586 polish words (ANPW): Duality-of-mind approach. *Behavior Research Methods*, *47*(3), 860–870. doi:10.3758/s13428-014-0509-4

Isen, A. M., Johnson, M. M., Mertz, E., & Robinson, G. F. (1985). The influence of positive affect on the unusualness of word associations. *Journal of Personality and Social Psychology, 48*(6), 1413–1426. doi:10.1037/0022-3514.48.6.1413

Johnson, R. C., & Lim, D. (1964). Personality variables in associative production. *Journal of General Psychology, 71*(2), 349–350.

Kamps, J., Marx, M., Mokken, R. J., & De Rijke, M. (2004). Using WordNet to measure semantic orientation of adjectives. *In Proceedings of LREC* (Vol. 4, pp. 1115–1118). doi:10.1.1.134.483 Klauer, K. C. (1997). Affective priming. *European Review of Social Psychology, 8*(1), 67–103. doi:10.1080/14792779643000083

Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. (2014). Emotion and language: Valence and arousal affect word recognition. *Journal of Experimental Psychology, General, 143*(3), 1065–1081. doi:10.1037/a0035669

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211–240. doi:10.1037/0033-295X.104.2.211

Lane, R. D., Chua, P. M. L., & Dolan, R. J. (1999). Common effects of emotional valence, arousal and attention on neural activation during visual processing of pictures. *Neuropsychologia, 37,* 989–997. doi:10.1016/S0028-3932(99)00017-2

Lang, P. J., Bradley, M. M., Fitzsimmons, J. R., Cuthbert, B. N., Scott, J. D., Moulder, B., & Nangia, V. (1998). Emotional arousal and activation of the visual cortex: An fMRI analysis. *Psychophysiology, 35,* 199–210. doi:10.1111/1469-8986.3520199

Maddock, R. J., Garrett, A. S., & Buonocore, M. H. (2003). Posterior cingulate cortex activation by emotional words: fMRI evidence from a valence decision task. *Human Brain Mapping, 18,* 30–41. doi:10.1002/hbm.10075

Mandera, P., Keuleers, E., & Brysbaert, M. (2015). How useful are corpus-based methods for extrapolating psycholinguistic variables? *The Quarterly Journal of Experimental Psychology, 68*(8), 1623–1642. doi:10.1080/17470218.2014.988735

Manning, C. D., & Schütze, H. (1999). In H. Schütze (Ed.), *Foundations of statistical natural language processing*. Cambridge, MA: MIT press.

Matlin, M. W., & Stang, D. J. (1978). *The Pollyanna principle: Selectivity in language, memory, and thought*. Cambridge, MA: Schenkman.

Moffat, M., Siakaluk, P. D., Sidhu, D. M., & Pexman, P. M. (2015). Situated conceptualization and semantic processing: Effects of emotional experience and context availability in semantic categorization and naming tasks. *Psychonomic Bulletin & Review, 22*(2), 408–419. doi:10.3758/s13423-014-0696-0

Monnier, C., & Syssau, A. (2014). Affective norms for french words (FAN). *Behavior Research Methods, 46*(4), 1128–1137. doi:10.3758/s13428-013-0431-1

Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2014). The adaptation of the Affective Norms for English Words (ANEW) for Italian. *Behavior Research Methods, 46*(3), 887–903. doi:10.3758/s13428-013-0405-3

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., van Schie, K., Van Harmelen, A.-L., … Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods*, *45*(1), 169–77. doi:10.3758/s13428-012-0243-8

Mourão-Miranda, J., Volchan, E., Moll, J., De Oliveira-Souza, R., Oliveira, L., Bramati, I., … Pessoa, L. (2003). Contributions of stimulus valence and arousal to visual activation during emotional perception. *NeuroImage*, *20*, 1955–1963. doi:10.1016/j.neuroimage.2003.08.011

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers, 36*(3), 402–407. doi:10.3758/BF03195588

Nelson, D. L., McKinney, V. M., Gee, N. R., & Janczura, G. A. (1998). Interpreting the influence of implicitly activated memories on recall and recognition. *Psychological Review, 105*(2), 299–324. doi:10.1037/0033-295X.105.2.299

Newcombe, P. I., Campbell, C., Siakaluk, P. D., & Pexman, P. M. (2012). Effects of emotional and sensorimotor knowledge in semantic processing of concrete and abstract nouns. *Frontiers in Human Neuroscience*, *6*. doi:10.3389/fnhum.2012.00275

Niedenthal, P., Halberstadt, J. B., & Innes-Ker, A. H. (1999). Emotional response categorization. *Psychological Review, 106*(2), 336–361. doi:10.1037/0033-295X.106.2.337

Osgood, C. E. (1975). *Cross-cultural universals of affective meaning*. Urbana, IL: University of Illinois Press.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., … Seligman, M. E. P. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology, 108*(6), 934–952. doi:10.1037/pspp0000020

Pollio, H. (1964). Some semantic relations among word-associates. *The American Journal of Psychology, 77*(2), 249–256. doi:10.2307/1420131

Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods, 41*(3), 647–656. doi:10.3758/BRM.41.3.647

Recchia, G., & Louwerse, M. M. (2015). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *Quarterly Journal of Experimental Psychology, 68*(8), 1584–1598. doi:10.1080/17470218.2014.941296

Redondo, J., Fraga, I., Padrón, I., & Comesaña, M. (2007). The Spanish adaptation of ANEW (affective norms for English words). *Behavior Research Methods, 39*(3), 600–605. doi:10.3758/BF03193031

Soares, A. P., Comesaña, M., Pinheiro, A. P., Simões, A., & Frade, C. S. (2012). The adaptation of the Affective Norms for English Words (ANEW) for European Portuguese. *Behavior Research Methods, 44*(1), 256–269. doi:10.3758/s13428-011-0131-7

Söderholm, C., Häyry, E., Laine, M., & Karrasch, M. (2013). Valence and arousal ratings for 420 Finnish nouns by age and gender. *PLoS One, 8*(8), e72859. doi:10.1371/journal.pone.0072859

Szalay, L. B., & Deese, J. (1978). *Subjective meaning and culture: An assessment through word associations*. Hillsdale, NJ: Lawrence Erlbaum.

Turney, P., & Littman, M. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS), 21*(4), 315–346. doi:10.1145/944012.944013

Van Rensbergen, B., Storms, G., & De Deyne, S. (2015). Examining assortativity in the mental lexicon: Evidence from word associations. *Psychonomic Bulletin & Review*. Advance online publication. doi:10.3758/s13423-015-0832-5

Vincze, N., & Bestgen, Y. (2011). Une procédure automatique pour étendre des normes lexicales par l'analyse des cooccurrences dans des textes. *Traitement Automatique Des Langues, 52*(3), 191–216.

Võ, M. L.-H., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., & Jacobs, A. M. (2009). The Berlin Affective Word List Reloaded (BAWL-R). *Behavior Research Methods, 41*(2), 534–538. doi:10.3758/BRM.41.2.534

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*(4), 1191–207. doi:10.3758/s13428-012-0314-x

Wicklund, D., Palermo, D., & Jenkins, J. (1965). Associative clustering in the recall of children as a function of verbal association strength. *Journal of Experimental Child Psychology, 2*(1), 58–66. doi:10.1016/0022-0965(65)90015-9

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality, 44*(3), 363–373. doi:10.1016/j.jrp.2010.04.001