THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 2014
http://dx.doi.org/10.1080/17470218.2014.970204

Routledge
Taylor & Francis Group

Avoid violence, rioting, and outrage; approach celebration, delight, and strength: Using large text corpora to compute valence, arousal, and the basic emotions

Chris Westbury[1], Jeff Keith[1], Benny B. Briesemeister[2,3], Markus J. Hofmann[2,3,4,5], and Arthur M. Jacobs[2,3]

[1]Department of Psychology, University of Alberta, Edmonton, AB, Canada
[2]Department of Psychology, Free University Berlin, Berlin, Germany
[3]Experimental and Neurocognitive Psychology, Dahlem Institute for Neuroimaging of Emotion (DINE), Berlin, Germany
[4]Department of Psychology, University of Wuppertal, Wuppertal, Germany
[5]General and Biological Psychology, University of Wuppertal, Wuppertal, Germany

Ever since Aristotle discussed the issue in Book II of his *Rhetoric*, humans have attempted to identify a set of "basic emotion labels". In this paper we propose an algorithmic method for evaluating sets of basic emotion labels that relies upon computed co-occurrence distances between words in a 12.7-billion-word corpus of unselected text from USENET discussion groups. Our method uses the relationship between human arousal and valence ratings collected for a large list of words, and the co-occurrence similarity between each word and emotion labels. We assess how well the words in each of 12 emotion label sets—proposed by various researchers over the past 118 years—predict the arousal and valence ratings on a test and validation dataset, each consisting of over 5970 items. We also assess how well these emotion labels predict lexical decision residuals (LDRTs), after co-varying out the effects attributable to basic lexical predictors. We then demonstrate a generalization of our method to determine the most predictive "basic" emotion labels from among all of the putative models of basic emotion that we considered. As well as contributing empirical data towards the development of a more rigorous definition of basic emotions, our method makes it possible to derive principled computational estimates of emotionality—specifically, of arousal and valence—for all words in the language.

Keywords: Affect; Valence; Arousal; Lexical access; Co-occurrence; Semantics; Emotion.

Aristotle discussed the issue of basic emotions in Book II of his *Rhetoric* (Aristotle, 200 BCE/1941). Since that time, humans have attempted to identify a set of basic emotion labels—labels describing something akin to universal, irreducibly basic affect states, like *good, bad*, *happy*, or *sad*. This task is complicated by the inherent complexity and ambiguity in defining what it means for an emotion to be "basic" (as reviewed in a later section), but also by the lack of methods that might make it

Correspondence should be addressed to Chris Westbury, Department of Psychology, University of Alberta, P220 Biological Sciences Building, Edmonton, AB, Canada T6G 2E9. E-mail: chrisw@ualberta.ca

We certify that none of the authors has any conflict of interest with any organization or person regarding the material discussed in this paper.

This work was supported by the Natural Sciences and Engineering Research Council of Canada.

possible to approach the question in an empirically grounded, quantifiable way. In this paper, we present an algorithmic method for evaluating proposed sets of basic emotion labels—a method we further employ to evaluate a number of putative models of basic emotion. We do so by using the relationship between a large set of words for which human emotionality judgements have been collected and co-occurrence distances of those words to "basic" emotion labels. These co-occurrence distances serve as proxies for measuring semantic relatedness between words/labels and do not depend on human judgement. We then demonstrate a generalization of our method that uses backwards regression to determine the most predictive "basic" emotion labels from among all tested models of basic emotion. As well as contributing empirical data towards the development of a more rigorous definition of basic emotions, our method makes it possible to derive principled computational estimates of emotionality—specifically, arousal and valence—for all words in the language.

In this introductory section, we first elaborate on basic emotions, as they are conceptualized in this study. We then introduce lexical co-occurrence models and present an overview of how they are used to model human judgements of meaning in general and of how we use them here to model semantic differentiation via emotionality judgements.

What is a basic emotion?

In their discussion (and defence) of the idea of "basic emotions", Scarantino and Griffith (2011) note that there are two broad approaches: a *folk emotion approach*, which focuses on human intuitions and conceptions of emotion, and a *scientific emotion approach*, which tries to use empirical methods to identify basic emotions. The method we use in this paper is a hybrid of these two methods, since it uses quantitative empirical methods to look at normative use of emotion labels in written language. Scarantino and Griffith also noted that across the two approaches there are at least three different ways that a set of emotions might be basic: *conceptually, biologically*, and *psychologically*. A *conceptually basic emotion*, such as "anger" or "joy", is a basic-level kind (in the sense outlined by Rosch, 1973, 1978) in a conceptual taxonomy. Such emotions would be identified by evidence that the emotion labels were psychologically privileged in tasks, by being, for example, more likely to be produced as exemplars of the category of emotion, producing quicker reactions times in decision tasks, being acquired earlier in the lifespan, and so on (e.g., Fehr & Russell, 1984; Shaver, Schwartz, Kirson, & O'Connor, 1987). *Biologically basic emotions*, like "fear" and "lust", are emotions for which a common evolutionary origin, unambiguous evolutionary adaptations, and/or cross-species and universal human behavioural markers may be identified (Darwin, 1872; Ekman, 1980, 1999; Izard 1977, 1992; Panksepp, 1982, 1998, 2007, 2008a, 2008b). Finally, *psychologically basic emotions*, like "pleasure", are emotions that are neither blends nor proportions of any other emotions.

However, across both approaches, and within each of the three different ways of being basic, there is no consensus about what should constitute a canonical list of basic emotions. In fact, as is detailed in Experiment 1, there have been many theories of—and many different sets of labels for—basic emotions proposed by researchers over the last century and beyond. One possibility that is orthogonal to all three forms of basic emotions is that the linguistic emotion labels that we use for labelling basic emotions are not in fact the proper unit for labelling whatever it is that is "basic" about emotion. The concept of "*a* basic emotion" may be what is known in philosophy as a *category mistake*, a term introduced and demonstrated by Ryle (1949) through an analogy in which a visitor to a university, after having been shown its various buildings and grounds, asks, "But where is the *University*?" Rather than representing discrete, basic units, emotion labels may be terms we attach—perhaps rather loosely and using folk psychology—to various combinations of definable lower level neurobiological activities that are themselves the proper basic units of emotion. For example, in their study of

the bases of emotion as gleaned from neural activation measured with functional magnetic resonance imaging (fMRI), Kassam, Markey, Cherkassky, Loewenstein, and Just (2013) concluded from a factor analysis that brain activity underlying emotional experience consisted mainly of four relevant separable dimensions (listed here in terms of decreasing importance in explaining variance): one encoding *valence* (whether a stimulus was is positively or negatively valenced); one encoding *arousal* (the strength of the associated affect); a third encoding whether the emotion has a *social aspect* (as jealousy necessarily must, but disgust need not); and a fourth that uniquely applied to emotionality associated with *sexual desire*.

Viewing emotion as a combination of multiple underlying dimensions cleaves the issue of basic emotions into two questions that may have only a loose association with each other. One question is addressed by Kassam et al. (2013): What is the *structure* of the biological variation that underlies emotion labels? The second question, which is relevant to the present study, is: How are emotion labels *related to each other*? More precisely: To what extent do emotion labels share variance—that is, overlap in underlying "meaning"—with each other? If emotion labels are analogous to regions in a high dimensional space defined by multiple underlying factors at the neurobiological level—an "emotion space" for simplicity—then it may be the case that no lexically labelled emotion is basic, even if some emotion labels might be more or less distinct than others. Slightly different lexical terms, which are strongly associated with the same alleged basic emotion (say, *mad* versus *furious*), may nonetheless behave differently in such an emotion space. For example, the region of emotion space that we label as *furious* may be much closer, on average, to the region of emotion space that we label as *disgust* than it is to the region we label as *mad*, while *mad*, in turn, might be closer to *furious* than it is to *disgust*. As another example: Perhaps the strong term *joy* labels a clearer, smaller region in emotion space than the label *happiness* does, in much the way that the term *Mont Royal* (a hill near the centre of Montreal) labels a more specific region of the province of Quebec than *Montreal* does, even though both label well-defined regions. A question like "How far is Montreal from Quebec City?" is a question that admits of a much less precise answer than "How far is Mont Royal from Quebec City?", because the region of Montreal sprawls much more widely in all directions than the region of Mont Royal. The question "How similar is *disgust* to *mad*?" may be a question that has a different answer than "How similar is *disgust* to *furious*?", perhaps because (let us assume for the sake of example) the region labelled as *mad* sprawls much more widely than the region labelled as *furious*. To the extent that this is true, looking for a set of "basic" emotions, for whatever purpose, means first of all understanding how *emotion labels* are related to each other.

To approach this question of how emotion labels are interrelated, we employ a method that allows us to measure the shared variation among emotion labels with respect to the identified underlying dimensions of emotion. In this paper we introduce a method that allows us to estimate the independent and combined contributions of any word—in particular, emotion labels—to human ratings of *valence* and *arousal*, the two main dimensions underlying blood-oxygen-level-dependent (BOLD) signal variance in the Kassam et al. (2013) study (and widely agreed to be relevant by others). The method relies upon the fact that we can obtain information about the similarity in "meaning" (semantic association) between any two words from a co-occurrence model of semantics. In the next subsection we briefly outline how these models work.

Co-occurrence models and semantic judgements

Lexical co-occurrence models are a class of computational models that derive word meaning by capturing how words co-occur in human-generated text. These models infer word meaning through considering latent relationships between words in a large corpus of text by creating a high-dimensional

semantic space, where words are distributed heterogeneously in this space with their relative "positions" determined by how often words occur in similar textual contexts—meaning the more often words co-occur in similar textual contexts, the closer they will be in this semantic space, and, by inference, the greater the semantic association there will be between words. A number of co-occurrence models have been developed, and each have been shown to be effective in modelling human semantic judgement behaviour in a number of experimental paradigms and applied settings (Burgess & Lund, 2000; Durda & Buchanan, 2008; Hofmann, Kuchinke, Biemann, Tamm, & Jacobs, 2011; Jones & Mewhort, 2007; Landauer & Dumais, 1997; Lund & Burgess, 1996; Rhode, Gonnerman, & Plaut, 2007; Shaoul & Westbury, 2006a, 2008, 2010, 2011).

In this study, we use the open-source co-occurrence model HiDEx (Shaoul & Westbury, 2006a, 2008, 2010, 2011). Although the various co-occurrence models differ in a number of ways, HiDEx shares the basic defining feature of these models, which is to count how often each word occurs within a small window of text in front of and behind every other word in a large corpus of human-created text. This allows us to define *global co-occurrence vectors* for each word, which are a record of how often each word appeared close to every other word. If we have, for example, 60,000 words in our dictionary, we can define a 60,000 × 120,000 matrix, in which each row vector corresponds to a word (which we call the *target word*), each word in the dictionary is represented by two columns—one to record co-occurrences in front of, and another for behind, the target word—and each entry in a row vector corresponds to how often the target word co-occurred within the specified window before or after the column word (see Figure 1). Concretely, if the first word in our dictionary is *aardvark*, and we define a co-occurrence window of size five words in either direction, the first cell in each row of our matrix will be a count of how often the target word defining that row occurred no more than five words before the word *aardvark*, within some large corpus of text.

In practice, co-occurrence models deal with some complications to this simple overview, including weighting co-occurring words by their distance from the target word, adjusting the raw counts for the frequency of the target word, and compressing the co-occurrence matrix. More technical details about HiDEx can be found in Shaoul and Westbury (2010).

In all analyses reported in this paper, we used HiDEx's default settings (Table 1). Notably, we use the cosine between word vectors as a semantic similarity measure (i.e., words close to each other in semantic space will have a larger cosine value for the angle between their vector representations). Varying the settings for the free parameters listed in Table 1 can have a large impact on the model's performance, and a number of studies have explored optimal parameterization across a wide range of semantic tasks (Bullinaria & Levy, 2007, 2012; Lifchitz, Jhean-Larose, & Denhière, 2009; Shaoul & Westbury, 2010). The parameters used in this study have not been selected due to their being optimal for any given semantic task; rather, they reflect an aggregation of those general parameter settings that have been shown to be optimal across the majority of semantic tasks/paradigms within which they have been studied.

Our approach of using co-occurrence similarity measures to model emotionality judgements is similar to Osgood, Suci, and Tannenbaum's (1957) *semantic differential* approach to the measurement of meaning. Briefly, this theoretical framework accounts for learned associations between *signs* (e.g., words) and *significants* (e.g., a word's referent qua "meaning"). This was an extension to Morris's (1946) dispositional view of meaning, in which a sign becomes a sign of a significant because, through conditioning/association, the presentation of the sign comes to produce a disposition in the organism to make responses previously evoked by the significant. Osgood et al. (1957) posited that a sign evokes a mediating process that is some fractional part of the total behaviour elicited by the significant (see also, Osgood, 1952). Although the question of what the underlying nature of these mediating processes might be was left explicitly open, meaning is implicitly

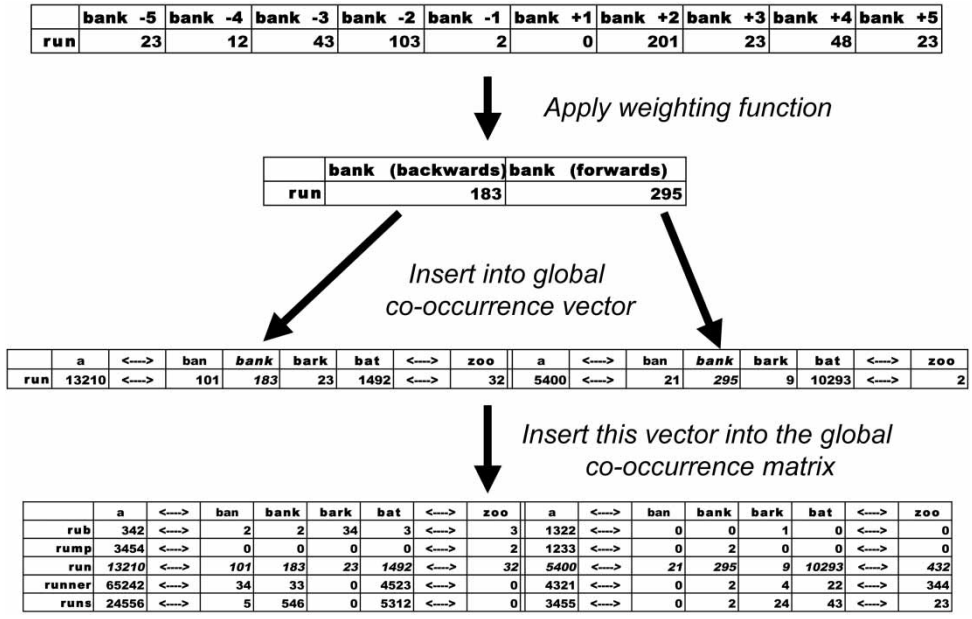

Figure 1. *Schematic illustration of how HAL (hyperspace analogue to language) word vectors are aggregated (by summing weighted backwards and forwards co-occurrence schemes into a single value each) and are inserted in the global matrix. In this example, the target word is "run", and the co-occurring word is "bank". Values are normalized for frequency after all vectors have been inserted in the global co-occurrence matrix (this normalization step is not shown here). From C. Shaoul and C. Westbury, HiDEx: The high dimensional explorer. In* Applied natural language processing and content analysis: Identification, investigation, and resolution *(p. 233), by P. McCarthy and C. Boonthum, 2011, Place of Publication: IGI Global. Copyright 2011 by IGI Global. Adapted with permission.*

defined in this view by the context of statistical redundancies in multitudinous multisensory elements perceived in environment—that is, meaning is composed of more factors than what can be found in analysing text or scale ratings alone. In this theoretical framework, significants evoke a behavioural disposition that becomes paired with a sign, and the sign evokes some of the significants' dispositions, which in turn influences both dispositional associations created/modulated with other signs and the selection of any subsequent signs in a given context. This is one (simplified) account for why are able to find emotionality in language through latent co-occurrence relationships.

Table 1. *HiDEx settings used for computing co-occurrence distances reported in this paper*

Corpus	12,714,502,395 words of USENET postings (Shaoul & Westbury, 2010)
Normalization	Positive pointwise mutual information
Vector similarity metric	Cosine
Weighting scheme	Inverse ramp
Window size	5 words in either direction
Dimensions used ("context size")	10,000

In their programme of research, Osgood et al. (1957) used bipolar scales to approximate semantic differentials and, through a series of studies and factor analyses, consistently found that using affective anchors (i.e., emotion labels) provided the best semantic differentiation in a number of contexts (in fact, we use the emotion labels from their optimal model in Experiment 1). Collecting human ratings of concepts (mostly individual words) on these bipolar emotionality scales allowed the construction of a semantic space—using the emotionality scales as dimensions—in which each concept/word was represented by a point in the space defined by its average rating by humans on each

of the scales. Similar to what was earlier described regarding co-occurrence models, semantic similarity is represented by relative distance—Euclidean distance in Osgood et al.'s work—between concepts/words in such a space.

Though differing in implementation, the present study uses a very similar approach to modelling semantic differentiation. Instead of using bipolar scales, we use co-occurrence similarity to emotion labels, which is analogous to using unipolar scales, and, instead of using small samples of human judgements on very limited numbers of concepts/words to develop our semantic space, we use latent relationships between tens of thousands of words found in analysing massive quantities of text, generated by hundreds of thousands (if not millions) of humans.

In Westbury, Briesemeister, Hofmann, and Jacobs (2014), we analysed a set of five emotion labels (*happiness, sadness, fear, disgust*, and *anger*) that have been suggested as basic emotion terms by many researchers (Ekman, 1999; Johnson-Laird & Oatley, 1989; Levenson, 2003). Similar to the present study, we used co-occurrence distances (between these emotion labels and thousands of target words) to model human judgements of arousal, valence, and the effects on behavioural measures of lexical access (lexical decision reaction times) of the associated judged emotionality of each word. We found support for co-occurrence similarities serving as objective measures of human semantic judgements, using emotionality as a semantic differential. Here we extend that work by considering 12 other proposed models of basic emotions, to see which model's set of emotion labels might be the best predictor of human judgements of arousal and valence. As much work has focused on the effect of affective variables in lexical access (i.e., Kousta, Vigliocco, Vinson, Andrews, & Del Campo, 2011; Larsen, Mercer, Balota, & Strube, 2008; Robinson, Storbeck, Meier, & Kirkeby, 2004), we also examine how well co-occurrence distance from the labels in each model predicts lexical decision reaction times (LDRTs).

EXPERIMENT 1: COMPARING 12 PUBLISHED SETS OF BASIC EMOTIONS

The 12 published models of the basic emotions that we considered are described here in order of their publication date, from oldest to newest.

The first model was Wundt's (1896), which proposed that emotion consisted of three basic axes: one for evaluation ("*pleasant/unpleasant*"), one for arousal ("*excitement/depression*"), and one for attention ("*tension/relaxation*").

The second model we used, which is closely related to Wundt's but derived using entirely different methods, as earlier discussed, was Osgood et al.'s (1957) model derived from a factor analysis of a large set of affective ratings. They elucidated three bipolar dimensions that were labelled as "*good/bad*", "*active/passive*", and "*strong/weak*". The similarity of the first two dimensions to Wundt's first two is obvious. The third dimension is more similar to Wundt's third dimension than it might appear to be, since Osgood and colleagues intended it to be broadly construed as a "potency" dimension (or what we would today call "arousal"), which is probably related to attentional force.

Tomkins (1962, 1963) proposed eight basic emotions.[1] Tomkins defined the dimensions representing these basic emotions with two labels each, the first for the milder version of the emotion and the second for a stronger related emotion. His milder/strong labels are *"interest/excitement"*, *"enjoyment/joy"*, *"surprise/startle"*, *"distress/anguish"*, *"fear/terror"*, *"shame/humiliation"*, *"contempt/disgust"*, and "*anger/rage*". Because most of these labels are close synonyms, we assessed this model as two separate models, consisting of the first and second labels, respectively.

Ekman, Sorenson, and Friesen (1969) proposed a list of six basic emotions: *happiness, surprise, fear, sadness, anger*, and *disgust*. Ekman (1999) extended this list with an additional 11 terms: *amusement, contempt, contentment, embarrassment, excitement, guilt, pride, relief, satisfaction, pleasure*, and *shame*.

[1]We ignore Tomkins's nonlexicalized ninth emotion, *dismell* or reaction to a bad smell.

We considered labels from both the shorter 1969 model and the longer 1999 model.

The next set of labels were taken from Plutchik's (1980) model derived from consideration of evolutionarily adaptive emotions relevant across species. Plutchik's model proposed seven primary emotions: *anger, fear, sadness, disgust, surprise, anticipation*, and *joy*.

In the context of presenting his computational belief desire theory of emotion, Reisenzein (2009) proposed that most or all emotions could be expressed as variants of just a few: *happiness, unhappiness, hope, fear, surprise, disappointment*, and *relief*. Because of the analysis of these emotions as basic emotions depends in Reisenzein's theoretical framework on *desire* and *aversion*, we also included these two labels.

Panksepp's (2005) model was driven by similar considerations of cross-mammalian universality. Building on the four basic emotions (*expectancy, rage, fear*, and *panic*) he had originally argued for in Panksepp (1982), Panksepp (2005) proposed that neural hard-wiring in the mammalian brain underlay seven primary emotions: *seeking, fear, rage, lust, care, panic*, and *play*.

Rather than focusing on particular emotions, Robinson et al. (2004) focused on the ubiquity in emotion theories of the general emotional characteristics of arousal and valence. In a series of behavioural studies, they showed that RTs on a number of different tasks were quickest when a negative stimulus was high in arousal (signalling a potentially dangerous stimulus, to be *avoided*) or a positive stimulus was low in arousal (signalling a safe stimulus, to be *approached*). Since Robinson et al. did not suggest a specific set of labels for basic emotions, we took the general focus of their analysis and created a set of labels related to danger or the lack thereof, to arousal, and to approach/avoidance behaviour: *approach, avoid, towards, away, to, from, evaluate, arouse, danger*, and *safe*.

Stevenson, Mikels, and James (2007) extended the affective norms (valence and arousal judgements) collected by Bradley and Lang (1999). Stevenson et al. had subjects rate the relatedness of 1034 words on five discrete emotions that they considered to be cross-culturally universal basic emotions (following Ekman et al., 1969; Ekman 1980; Levenson, 2003): *happiness, sadness, fear, disgust*, and *anger*.

Kassam et al. (2013) looked at the neural correlates of eight basic emotions: *anger, disgust, envy, fear, happiness, lust, sadness*, and *shame*. They also listed 18 labels related to these basic emotions (*angry, enraged, disgusted, revulsed, envious, jealous, afraid, frightened, happy, joyous, lustful, horny, proud, admirable, sad, gloomy, ashamed*, and *embarrassed*). These labels were included in the analyses that we report in Experiment 2, with the exception of the term *revulsed*, which did not appear in the HiDEx dictionary.

Although these models do not exhaust the space of possible models of basic emotion labels (cf. Johnson-Laird & Oatley (1989), who list 590 distinct emotion labels), they do cover a wide and representative range of possible emotion labels. Together they propose 78 distinct terms as possible basic emotion labels, which are reproduced in the Supplemental Material, Appendix A.

Method

In order to assess each of these models, we used the arousal and valence ratings collected for 13,915 words by Warriner, Kuperman, and Brysbaert (2013). In total, 10,931 of these rated words appeared both in the dictionary of our co-occurrence model and in the English Lexicon Project database (Balota et al., 2007) of visual lexical decision reaction times (LDRTs). We randomly split this subset of the rated words in half, to define a model development set of 5465 terms and a model validation set of 5466 terms.

Before adding emotion label predictors for LDRTs, we covaried out the effects of a number of lexical variables that are well known to account for RT variance in the lexical decision task: the logarithm of orthographic frequency (Grainger & Jacobs, 1996; Jacobs & Grainger, 1994; Shaoul & Westbury, 2006b), word length, number of syllables, orthographic neighbourhood size, and the logarithm of place-controlled summed bigram frequency (the last three measures came from the English Lexicon Project database). All of these

predictors entered reliably into the regression model, together accounting for 39.9% of the variance in LDRTs in the set of 10,931 measures ($p < 2E\text{-}16$). We saved the residuals from this regression (RT-RESIDUALS) as the target variable for predicting reaction times from the co-occurrence similarity of the target word to emotion labels. We discuss some potential pitfalls of this approach later in this paper.

We tested the ability of each of our 12 models to predict three target measures: the arousal ratings, the valence ratings, and RT-RESIDUALS. We used the co-occurrence similarities (measured as the cosine between word vectors) between each emotion label and each word in the test sets as predictors—for example, for Wundt's model, the following regression formula would be used to predict our three target measures:

$$\begin{aligned} <\text{targetMeasure}> \sim\ & \beta_1 \times {}^*\text{COS}(\text{target, pleasant}) \\ & + \beta_2 \times {}^*\text{COS}(\text{target, unpleasant}) \\ & + \beta_3 \times {}^*\text{COS}(\text{target, excitement}) \\ & + \beta_4 \times {}^*\mathit{COS}(\text{target, depression}) \\ & + \beta_5 \times {}^*\text{COS}(\text{target, tension}) \\ & + \beta_6 \times {}^*\text{COS}(\text{target, relaxation}). \end{aligned}$$

Using linear regression with backwards elimination, we eliminated each term that did not contribute to each model with $p < .05$. The resultant models were then validated by running the final regression equation from the test set on the 5466 terms in the validation set. Models were evaluated (within and between test and validation sets) by both comparing target~model correlations, and by comparing each model's Akaike information criterion (AIC) value, an information-theory-motivated measure of the relative quality of statistical models for given data, considering both goodness of fit and number of parameters.

Results

As shown in the summary of the results in Table 2, all 12 models were highly reliable predictors on both the test and validation set of our three target measures.

All models were also well validated, with an average (standard deviation, *SD*) *r*-squared difference between the test and validation sets over all 12 sets of basic emotion terms of just .012 (.005) for the arousal ratings, −.003 (.005) for the validation ratings, and −.002 (.003) for the predictions of RT-RESIDUALS.

The best model of valence was Ekman's (1999), having a validated correlation of .58. This model also had the highest number of predictors (18), giving it an advantage over other models. The best model for predicting arousal was Kassam and colleagues' (2013) eight-term model, which validated with a correlation of .23. The best model for predicting RT-RESIDUALS (by a very small margin) was Wundt's (1896) model, with a validated correlation of .22. Only four emotion label predictors entered into this model: DEPRESSION, EXCITEMENT, and PLEASANT, all of which had negative weights (indicating faster RTs with closer similarity to these terms), and UNPLEASANT, which had a positive weight (indicating slower RTs with closer similarity to this term).

Co-occurrence similarities to emotion labels were much better predictors of valence ratings [average (*SD*) validation-set correlation: .49 (.06)] than of arousal ratings [average (*SD*) validation-set correlation: .17 (.03); by Fisher's *r*-to-*z* test on the averages, these are reliably different: $z = 19.04$, $p < 2E\text{-}16$]. This may be in part because most emotion labels were reliable predictors of valence, with terms dropping out primarily in models that had a large number of predictors. Many fewer labels were predictive of arousal.

Discussion

In the context of using co-occurrence similarities as predictors, valence seems to reflect average (signed) semantic similarity to a wide range of emotion labels. We refine this high-level summary in Experiment 2.

It is difficult to draw any further conclusions from this comparison, because different models

Table 2. *Comparison of 12 published sets of "basic emotion terms", by prediction of human ratings of arousal and valence and RT residuals, on both a test and a validation set*

		TEST SET (N = 5465)			*VALIDATON SET (N = 5466)*		
AUTHOR/YEAR	*AFFECT TERMS*	*AROUSAL*	*VALENCE*	*RT RESIDUALS*	*AROUSAL*	*VALENCE*	*RT RESIDUALS*
Ekman et al. (1969)	*ANGER*, DISGUST, FEAR, HAPPINESS, *SADNESS*, *SURPRISE*	.16	.49	.15	.13	.50	.17
Ekman (1999)	*AMUSEMENT*, *ANGER*, *CONTEMPT*, *CONTENTMENT*, *DISGUST*, *EMBARRASSMENT*, *EXCITEMENT*, *FEAR*, GUILT, *HAPPINESS*, *INTEREST*, *PLEASURE*, PRIDE, *RELIEF*, *SADNESS*, *SATISFACTION*, SHAME, SURPRISE	.23	.57	.21	.20	.58	.21
Kassam et al. (2013)	*ANGER*, *DISGUST*, *ENVY*, FEAR, *HAPPINESS*, *LUST*, *SADNESS*, SHAME	.25	.50	.16	.23	.51	.18
Osgood, Suci, & Tannenbaum (1957)	ACTIVE, *BAD*, *GOOD*, *PASSIVE*, *STRONG*, *WEAK*	.17	.43	.09	.15	.45	.10
Panksepp (2005)	*CARE*, *FEAR*, *LUST*, PANIC, PLAY, RAGE, SEEKING	.19	.38	.19	.15	.38	.19
Plutchik (1980)	*ANGER*, *ANTICIPATION*, DISGUST, *FEAR*, *JOY*, *SADNESS*, SURPRISE	.20	.47	.21	.15	.48	.21
Reisenzein (2009)	*AVERSION*, *DESIRE*, DISAPPOINTMENT, *FEAR*, *HAPPINESS*, *HOPE*, *RELIEF*, *SURPRISE*, *UNHAPPINESS*	.21	.52	.17	.20	.52	.20
Robinson et al. (2004)	*APPROACH*, *AROUSE*, *AWAY*, *DANGER*, *EVALUATE*, FROM, *SAFE*, TO, TOWARDS, *WITHDRAW*	.23	.37	.17	.20	.38	.16
Stevenson et al. (2007)	*ANGER*, DISGUST, FEAR, HAPPINESS, *SADNESS*	.16	.47	.15	.12	.47	.17
Tomkins (1962, 1963): Mild terms	*ANGER*, CONTEMPT, *DISTRESS*, *ENJOYMENT*, *FEAR*, *INTEREST*, SHAME, SURPRISE	.19	.52	.20	.14	.51	.20
Tomkins (1962, 1963): Strong terms	ANGUISH, *DISGUST*, EXCITEMENT, *HUMILIATION*, *JOY*, *RAGE*, STARTLE, *TERROR*	.22	.52	.20	.17	.52	.20
Wundt (1896)	DEPRESSION, *EXCITEMENT*, *PLEASANT*, *RELAXATION*, TENSION, *UNPLEASANT*	.22	.53	.21	.21	.53	.22

Note: Terms that reliably predicted arousal are in italics. Terms that reliably predicted valence are bolded. Terms that reliably predicted RT residuals are underlined. RT = reaction time.

were better for different purposes and contained a different number of predictors. We know that linear models are sensitive to both the intercorrelation between their predictors and the number of their predictors. We suggest that both distantly related and closely related emotion labels (for example, *anger* and *rage*) may co-occur differently with the words for which we have valence and arousal judgements, making them unequal as predictors even if they seem to be close synonyms. We must therefore confront the fact that comparing these 12 models is in many ways comparing apples and oranges. However, this analysis does establish a relationship between co-occurrence distances to affect terms, on the one hand, and arousal judgements, valence judgements, and lexical access times, on the other. In the second study we look at these relationships in a more general way.

EXPERIMENT 2: MAXIMIZING PREDICTION OF AROUSAL, VALENCE, AND LDRT

A useful aspect of the approach we have taken is that it is easily extensible. From our methodological perspective, a model is just a set of lexical emotion labels. Our method is therefore well defined not only for all the 12 proposed models of basic emotions considered so far, but for all possible models of basic emotions. This allows us to consider the possibility of using large-scale simulations of "basic emotion models" to "titrate" the contribution of individual emotion labels. In Experiment 2, we looked for the best possible models definable using all 78 emotion labels.

Method

Starting with a linear regression model containing predictors for all 78 emotion labels considered in Experiment 1, we conducted a backwards regression for each target measure (i.e., valence, arousal, and RT-RESIDUALS), eliminating terms until we had an optimally performing eight-predictor (i.e., using eight emotion labels) model. We chose eight predictors because the average number of terms in the models we considered was 8.1 (or 7.9 if we eliminate the terms we ourselves selected to represent the ideas of Robinson et al., 2004, who did not offer an explicit model themselves). This selection resulted in a very strict criterion for entry into the model, with a maximum p-value of .0005, and usually much less, in every case.

One problem with this approach is that the co-occurrence distances tend to be very highly intercorrelated. Across the 10,931 words, our 78 co-occurrence predictors had an average (*SD*) correlation of .48 (.23), $p < 2E\text{-}16$. In one way this is not a problem, since this multicollinearity does not cause any problem in fitting new data, assuming that the predictor variables have the same relationship in the validation set as the test set (an assumption we have no reason to doubt, since our test and validation sets were randomly defined). In other words, the reported r-squares are accurate in the face of collinearity, in the sense that the models do account for the validated variance that is reported. However, collinearity can lead to inaccuracies in computation of the proper weights on the models —for example, weights for covarying predictors will be heavily influenced by the order in which they are entered into the regression model. The result is that we know how well our model performs (we can trust that the obtained r-squared is true for *some* combinations of the predictors in the model), but we have uncertainty about what exactly that model is.

One possible approach to this problem (albeit not a solution, since it radically changes the predictors) is to regress out the shared variance and examine the regression table produced by using only the residuals of the original predictors. Although this approach is guaranteed to misestimate the problem if there is any true shared variance at all (because it simply eliminates that shared variance) it can give an "alternative view" of the problem. Comparing and contrasting the regression tables with shared and unshared variance can present an overall picture of the nature of the relationships being studied. Accordingly, we undertook 78 regressions, predicting each individual emotion label's distance measures as a linear function of the other 77 predictors, plus

LNFREQ—the log-transformed orthographic frequency of target words in our data set. We used a dictionary of 35,654 words with orthographic frequencies between 0.1 and 600 per million words (Shaoul & Westbury, 2006b) for this regression. Each regression threw out any predictors that contributed to the model with $p < .05$. We took the residuals from each one of these models and used them as predictors in three alternative regressions to predict the arousal and valence judgements and the RT-RESIDUALS. Across the 35,654 words, our 78 residualized co-occurrence predictors had an average (*SD*) correlation of −.01 (.08), $p > .2$.

Table 3. *Best regression model for predicting arousal ratings using raw co-occurrence distances, developed on 5465 ratings and validated on 5466 ratings*

Affect term	*Estimate*	SE	t	p
(Intercept)	4.32	0.08	52.56	<2E-16
LUST	10.59	0.91	11.57	<2E-16
PANIC	9.31	0.88	10.53	<2E-16
HUMILIATION	7.7	0.98	7.86	4.71E-15
SADNESS	−7.27	0.78	−9.32	<2E-16
AWAY	−5.43	0.72	−7.59	3.75E-14
DEPRESSION	−4.65	0.57	−8.21	2.79E-16
ASHAMED	−4.46	0.98	−4.54	5.68E-06
PLEASANT	−4.4	0.57	−7.71	1.50E-14

Note: Test-set $r = .31$; validation-set $r = .27$. $F(8, 5456) = 73.6$, $p < 2.2E-16$.

Results

Arousal

The best model for predicting arousal ratings with the full co-occurrence distance predictors is reproduced in Table 3. It had a correlation of .31 with the ratings in the test set and .27 with the ratings in the validation set. This is 3.0*Z* better on the validation set than the average of the 12 models considered in Experiment 1 [average (*SD*) = 0.17 (0.03)] and 1.3*Z* better than the best of those models, Kassam and colleagues' (2013) six-term model. The model suggests that higher arousal ratings are associated most strongly with similarity to the emotion labels HUMILIATION, LUST, and PANIC, which are all words associated with autonomic nervous system emotions. This is modulated by (generally smaller magnitude) negative weights for similarity to the labels ASHAMED, AWAY, DEPRESSION, PLEASANT, and SADNESS, which are therefore all associated with lower arousal ratings. The model is very similar to the Kassam et al. (2013) six-label model. Two predictors (SADNESS and LUST) appeared in both models, and a third was closely semantically related (Kassam et al.'s (2013) HAPPINESS matched with this model's PLEASANT).

The best model for predicting arousal using the residualized co-occurrence distances is shown in Table 4. It had a correlation of .18 with the test ratings and a correlation of .17 with the validation ratings, which was close to the average value of the best of the 12 models considered in Experiment 1, and much worse than the model with the full distances that was considered in the previous paragraph. Four of the predictors overlapped exactly with those in the full distance model and had the same sign (REZLUST, REZHUMILIATION, REZPLEASANT, and REZPANIC), and a fifth (REZLUSTFUL) was a close synonym, also signed the same way. The remaining three predictors were all associated with autonomic nervous system emotions (REZEXCITEMENT, REZTERROR, and REZANGUISH). The top predictor in both models was distance to the emotion label "lust".

Together these two models support the idea that arousal ratings are fairly weakly predicted by co-occurrence distances and almost entirely from words that are associated with autonomic system arousal, most notably "lust". Given the much worse performance of the second model using residualized co-occurrence distances, it appears that much of the variance in arousal ratings that can be accounted for by co-occurrence distances is due to variance that is shared between words, possibly due to shared co-occurrence relationships to other words in the model.

Valence

The best model for predicting valence ratings using the full co-occurrence distances is reproduced in Table 5. It had a correlation of .60 on both the

Table 4. *Best regression model for predicting arousal ratings using the residuals of each emotion label's co-occurrence distance after removing the effect of the other 77 emotion labels and LNFREQ, developed on 5465 ratings and validated on 5466 ratings*

Affect term	*Estimate*	SE	t	p
(Intercept)	4.38	0.01	303.01	<2E − 16
REZLUST	15.06	2.35	6.42	1.49E − 10
REZLUSTFUL	11.60	2.46	4.72	2.45E − 06
REZHUMILIATION	11.34	1.86	6.09	1.19E − 09
REZPLEASANT	−9.83	1.30	−7.59	3.88E − 14
REZEXCITEMENT	9.28	1.93	4.80	1.63E − 06
REZPANIC	8.42	1.77	4.75	2.14E − 06
REZTERROR	6.77	0.91	7.42	1.35E − 13
REZANGUISH	6.26	1.46	4.28	1.92E − 05

Note: Test-set $r = .18$; validation-set $r = .17$, $F(8, 5456) = 23.58$, $p < 2.2E - 16$. LNFREQ = the natural log of the orthographic frequency of the word.

test set and the validation set, which is 1.9Z better on the validation set than the average of the 12 models considered in Experiment 1 [average (SD) = .49 (.06)], and 0.3Z better than the best performing of those 12 models (Ekman, 1999), which included 12 terms. This top-performing eight-term model is of particular interest because it is has a very clear structure, consisting of three dimensions with anchoring emotion labels carrying positive/negative beta weights: STRONG/WEAK (*potency*), JOY/SADNESS (*happiness*), and PLEASURE/BAD (*approachability*). The last two labels in the model are the closely synonymous ANGRY (beta weight of −4.18) and RAGE (beta weight of −14.15), without a corresponding opposing anchor. Valence is therefore predicted mainly by similarity to positive emotion labels and dissimilarity to negative labels on the three named dimensions, with a "bonus" decrement in valence rating for labels associated with anger.

Table 5. *Best regression model for predicting valence ratings using raw emotion label co-occurrence distances, developed on 5465 ratings and validated on 5466 ratings*

Affect term	*Estimate*	SE	t	p
(Intercept)	4.34	0.07	61.85	<2E − 16
STRONG	17.58	0.8	22.1	<2E − 16
JOY	17.04	1.01	16.87	<2E − 16
RAGE	−14.15	1.11	−12.8	<2E − 16
PLEASURE	13.19	0.96	13.79	<2E − 16
SADNESS	−10.2	0.75	−13.57	<2E − 16
BAD	−8.21	0.64	−12.75	<2E − 16
WEAK	−7.03	0.69	−10.19	<2E − 16
ANGRY	−4.18	0.68	−6.19	6.64E − 10

Note: Test-set $r = .60$; validation-set $r = .60$, $F(8, 5456) = 377.7$, $p < 2.2E - 16$.

The best model for predicting valence using the residualized co-occurrence distances contained 57 predictors (74% of all predictors), each one of which entered into the model with $p < 2E\text{-}16$. Removing any one of the 57 predictors lowered the amount of variance the model could account for. The eight predictors with the largest absolute beta weights are reproduced in Table 6. Four of those predictors had negative beta weights: Three (leading to lower valence predictions) had to do with anger (REZANGER, REZANGRY, and REZENRAGED), and one had to do with fear (REZFEAR). The remaining four predictors among the top eight had positive beta weights (leading to higher valence predictions), and all had to do with happiness (REZJOY, REZHAPPY, REZHAPPINESS, and REZPLEASURE). The full 57-predictor model correlated with the test set at $r = .59$ ($p < .000001$) and with the validation set at $r = .55$ ($p < .000001$).

With so many predictors in the residualized co-occurrence distance model, fine-grained comparisons between the two models make little sense. However, we note two points. One is that the highly structured eight-predictor model

Table 6. *Most strongly weighted eight predictors for predicting valence ratings using the residuals of each emotion label's co-occurrence distance after removing the effect of the other 77 emotion label distances and LNFREQ, developed on 5465 ratings and validated on 5466 ratings*

Affect term	*Estimate*	SE	T	p
(Intercept)	5.15	0.01	357.63	<2E − 16
REZANGER	−222.37	7.00	−31.77	<2E − 16
REZFEAR	−179.09	4.95	−36.10	<2E − 16
REZJOY	165.91	4.39	37.82	<2E − 16
REZANGRY	−139.27	6.01	−23.17	<2E − 16
REZENRAGED	−139.17	6.66	−20.91	<2E − 16
REZHAPPY	126.06	5.08	24.80	<2E − 16
REZHAPPINESS	125.84	3.48	36.19	<2E − 16
REZPLEASURE	120.65	3.62	33.33	<2E − 16

Note: Eight predictors of 57 that entered into the model with $p < 2E - 16$. LNFREQ = the natural log of the orthographic frequency of the word.

that allowed shared variance performed just as well as the second model with 57 residualized predictors. The second is that the entry of so many residualized predictors into the model is consistent with our tentative conclusion in Experiment 1 that "valence seems to reflect average (signed) semantic similarity to a wide range of emotion labels".

RT-RESIDUALS

The best regression model found for predicting RT-RESIDUALS using the raw co-occurrence distances is reproduced in Table 7. It had a correlation of .24 on the test set and .21 on the validation set, which is 1.0z better on the validation set than the average of the 12 models evaluated in Experiment 1 [average (*SD*) = .18 (.03)]. However, it performed marginally worse on the validation set than the best of those 12 models, Wundt's (1896) model, which predicted RT-RESIDUALS with a correlation of .21 on the test set and a correlation of .22 on the validation set, using just four predictors: DEPRESSION, EXCITEMENT, PLEASANT, and (the only label with a positive weight) UNPLEASANT. Wundt's model can be slightly improved by deleting the relatively weak predictor EXCITEMENT, which has a negligible effect [validation r (AIC) with the term = .22 (63419); validation r (AIC) without the term = .22 (63421)]. Because this simple model has fewer predictors and performs slightly better than the eight-item model at predicting validation set RT-RESIDUALS, it should be considered the best model for predicting RT-RESIDUALS and is therefore reproduced in Table 8.

Further consideration of this model with the full (unresidualized) LDRTs revealed that it was amenable to additional refinements. As previously mentioned, the variance in the validation set of unresidualized LDRTs accounted for by the base model consisting only of the lexical predictors is 39.8% (AIC: 63461). When the three emotion label predictors from Wundt's model are added to this base model, DEPRESSION does not enter reliably into the model, so the model can be

Table 7. *Best regression model for predicting LDRT residuals, developed on 5465 ratings and validated on 5466 ratings*

Affect term	*Estimate*	SE	t	p
(Intercept)	54.16	8.39	6.46	1.17E − 10
SATISFACTION	−598.4	85.06	−7.04	2.23E − 12
JOY	−406.85	91.01	−4.47	7.96E − 06
HORNY	−402.2	110.49	−3.64	0.00028
GOOD	331.36	54.97	6.03	1.77E − 09
ADMIRABLE	316.17	62.38	5.07	4.14E − 07
JOYOUS	308.87	93.64	3.3	0.00098
PLEASANT	−269.73	43.58	−6.19	6.47E − 10
CONTEMPT	251.29	47.12	5.33	1.00E − 07

Note: After covarying out the effect of several lexical predictors. Test-set $r = .24$; validation-set $r = .21$, $F(8, 5456) = 43.53$, $p < 2.2E - 16$. LDRT = lexical decision reaction time.

Table 8. *Wundt's (1869) model for predicting LDRT residuals, developed on 5465 ratings and validated on 5466 ratings*

Affect term	*Estimate*	SE	t	p
(Intercept)	57.05	4.79	11.91	<2E − 16
DEPRESSION	−201.58	41.26	−4.885	1.06E − 06
PLEASANT	−433.69	34.38	−12.616	<2E − 16
UNPLEASANT	245.42	69.05	3.554	0.000382

Note: After covarying out the effect of several lexical predictors. Test-set $r = .21$; validation-set $r = .22$, $F(4, 5460) = 64.89$, $p < 2.2E - 16$.

further simplified by deleting this term. Entering just PLEASANT and UNPLEASANT increases variance accounted for to 40.9%, a very substantial improvement by AIC comparison (AIC: 63399). Allowing for an interaction between PLEASANT and UNPLEASANT (which interact reliably, $p <$ 2E-15) increases the variance accounted for to 41.6%, again a substantial improvement by AIC values (AIC: 63337). The model performed similarly well on validation, accounting for 42.6% of the variance in that set. The final model for predicting LDRT, with just these two interacting emotion label predictors, is shown in Table 9.

We note that both emotion term predictors in the final model using the full distances (Table 9) have (roughly equal magnitude) negative beta weights, indicating that they are associated with faster RTs. This may appear to be a contradiction to their role in predicting the residuals above (i.e., compare to beta weight estimates shown in Table 8). However, the dependent measures in the two models are not the same. In the first case, we developed a model using only emotion label predictors to estimate the LDRT residuals remaining after first covarying out lexical predictors; in the second case our model included both the lexical and emotion label predictors in the same regression equation to estimate the full, unresidualized LDRTs. As shown in Figure 2, the correlations of the individual lexical predictors and the co-occurrences distances to the emotion labels PLEASANT and UNPLEASANT are quite different. Further, across all of the Warriner et al. (2013) words, a regression of all the predictors onto the distances from PLEASANT accounts for 8.7% of the variance, while a regression of the same variables onto the distances from the emotion label UNPLEASANT accounts for 16.0% of the variance—a significant difference by Fisher's r-to-z test ($Z = 8.84$, $p < .00001$). In other words, regressing out the same lexical predictors does not have the same effect on each of the PLEASANT and UNPLEASANT emotion label predictors, making it difficult to adjudicate whether it is right or wrong to regress them out before looking for effects. We discuss some general implications of this problem in the final discussion.

Table 9. *Best model for predicting LDRTs, developed on 5465 ratings and validated on 5466 ratings*

Affect term	*Estimate*	SE	t	p
(Intercept)	928.29	24.68	37.62	<2E − 16
PLEASANT	−1042.95	100.50	−10.38	<2E − 16
UNPLEASANT	−1008.45	114.17	−8.83	<2E − 16
LENGTH	7.08	0.82	8.63	<2E − 16
SYLLABLES	26.80	1.85	14.46	<2E − 16
LNCONBG	−5.21	1.18	−4.40	1.09E − 05
LNFREQ	−28.57	0.94	−30.37	<2E − 16
PLEASANT:UNPLEASANT	5447.52	677.27	8.04	1.06E − 15

Note: Test-set $r = .42$; validation-set $r = .43$. $F(7, 5457) = 557$, $p = <2.2E - 16$. LDRT = lexical decision reaction time.

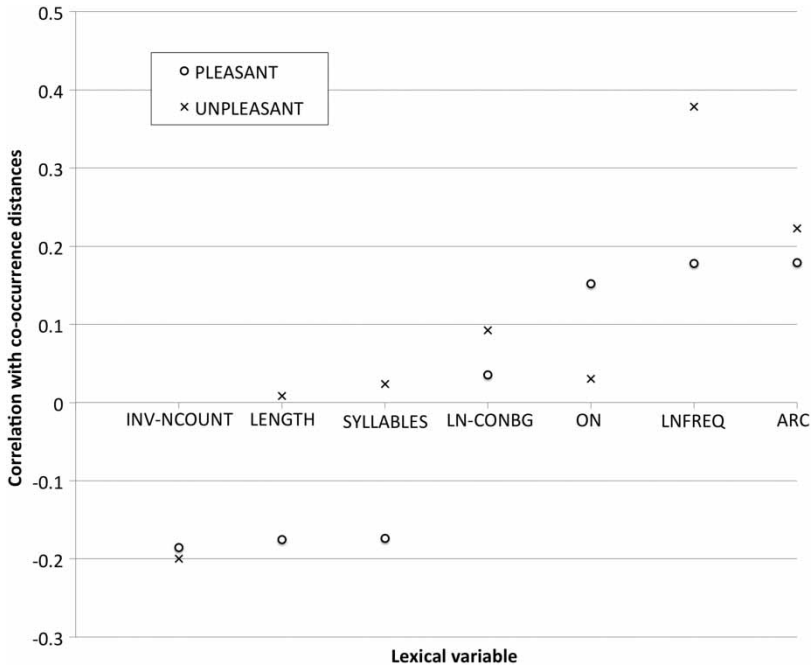

Figure 2. *Correlations between co-occurrence distance to the emotion labels PLEASANT and UNPLEASANT and the six lexical predictors used in this study. INV-NCOUNT = inverse N-count, a measure of co-occurrence neighbourhood size. LENGTH = word length in letters. SYLLABLES = number of syllables in word. LN-CONBG = the natural log of the place-controlled summed bigram frequency. ON = orthographic neighbourhood size. LNFREQ = the natural log of the orthographic frequency of the word. ARC = average radius of co-occurrence, a measure of co-occurrence neighbourhood density. LENGTH, SYLLABLES, and ON are better predictors of distances from the emotion label PLEASANT than UNPLEASANT. LNFREQ is a notably better predictor of distances from the emotion label UNPLEASANT than PLEASANT.*

Discussion

The models for predicting arousal and valence found by searching the space of emotion labels using backwards regression were substantially better (by at least 1.9*Z*) than the average model composed of labels suggested by other researchers as describing basic emotions. The models support the hypothesis of Westbury et al. (2014) that arousal is associated more strongly with autonomic reactivity than valence, predicted by co-occurrence similarity to emotion labels naming automatic emotional reactions (HUMILIATION, LUST, and PANIC).

It is interesting to see such a simple model in estimating LDRTs, contrasted with the more complex, multidimensional model of valence ratings, which includes eight predictors characterizing four dimensions—what we have labelled *potency*, *happiness*, *approach,* and *anger*. If we use the best emotion label predictors for valence to instead predict LDRTs, we end up with a reliably worse model, accounting for 41.1% of the variance with an AIC value of 63392. This dissociation of valence and the best LDRT emotion label predictors suggests that it may not be valence per se that affects LDRT. One possible interpretation of this is that valence is an estimate of some function of *pleasantness*, which appears to be the actual driver of the LDRT effect that is normally attributed to valence itself. This is consistent with evidence from a recent event-related potential (ERP) study (Briesemeister, Kuchinke, & Jacobs, 2014) suggesting that discrete nameable emotion states (i.e., *happiness*) are processed prior to more general positive valence.

Across the 10,931 words from the Warriner et al. (2013) ratings, there is a strong correlation between valence ratings and the co-occurrence distance to the emotion label PLEASANT ($r = .49$, $p < 2E\text{-}16$) but no relationship between valence ratings and the emotion label UNPLEASANT ($r = .006$). This explains why valence is not as good a predictor of LDRT as similarity to PLEASANT and UNPLEASANT. Co-occurrence similarities to both these terms are reliable contributors to predicting LDRT and are combined additively (plus an interaction term) in the regression equation for predicting LDRT.

There are many reasons why co-occurrence distances from the emotion label UNPLEASANT might not be a good predictor of valence. We speculate that it is mainly because a weak negative valence is associated with the colloquial use of the word. Being rained upon, feeling exhausted, and going to the dentist are unpleasant. It would be odd (and, we would suggest, incorrect) to use the word "unpleasant" to describe many of the most negatively valenced words in our data sets (e.g., *AIDS, homicide, castration*, *rapist*, and *genocide*, all of which appear on the Warriner et al., 2013 list). These refer or allude to things that are unquestionably very much worse than just unpleasant. The appropriate response when asked to rate such words from *pleasant* to *unpleasant* would be to say "But how can I possibly do that? Your scale does not go low enough! *Homicide* is not just *unpleasant*!". Faced with the impossible task they have been given, we suspect that what subjects do (quite sensibly) is to recalibrate the scale they were given so that "Unpleasant" is taken to mean (what it does not mean) "Absolutely terrible" and "pleasant" is take to mean (what it does not mean) "Absolutely wonderful". After that necessary recalibration, the alleged anchors "Pleasant" and "Unpleasant" (and similar terms that are used to anchor the valence ratings scales) will be middling words clustered close to the centre of a scale that actually goes from "Really terrible" to "Really wonderful".

Although no emotion label appears in all three models, all models do contain either PLEASANT or its near-synonym PLEASURE. High co-occurrence similarity to these terms is associated with faster LDRTs, higher valence, and lower arousal ratings.

By assessing these models in terms of correlation with human judgements, there is an implicit notion that the proper "gold standard" should be a correlation of 1.0. Of course this is not true, since human judges themselves cannot agree about 100% of the variance in arousal and valence judgements. The presence of noise in the ratings being modelled makes it impossible to model those ratings perfectly. In order to get a more accurate understanding of how much of the variance in human ratings is explained by our computational estimates, we looked at the correlation of human ratings for a subset of 2132 words for which we have human ratings from two sources (Adelman & Estes, 2013; Warriner et al., 2013) and which do not appear as predictors in any of the computational models. The results are shown in Table 10. The two human rating sources correlate reliably both for ratings of arousal ($r = .51$, $p < 2E\text{-}16$) and for valence ($r = .81$, $p = 2E\text{-}16$). Correlations between humans and the computer estimates are reliable, but also lower than the correlation between human ratings both for arousal (with the Warriner et al., 2013 ratings: $r = .28$; with the Adelman & Estes, 2013 ratings, $r = .42$; by Fisher r-to-z comparison: $z > 3.5$, $p < .0005$) and for valence (with the Warriner et al., 2013 ratings: $r = .56$; with the Adelman & Estes, 2013 ratings, $r = .64$; by Fisher r-to-z comparison: $z > 12.0$, $p < 2E\text{-}16$). Taking ratios of r-squared—that is, the squared correlation of the computer estimate with one rating set divided by the squared correlation of the two rating sets—the

Table 10. *Correlations of computer estimates of human ratings to two independent human ratings of arousal and valence?*

Estimate source	*Warriner et al. (2013)*	*Adelman & Estes (2013)*	*Computed*
Warriner et al. (2013)		.81	.56
Adelman & Estes (2013)	.51		.64
Computed	.28	.42	

Note: Arousal: lower diagonal; valence: upper diagonal.

computer estimates account for 30%/68% (Warriner et al., 2013/Adelman & Estes, 2013 respectively) of the variance accounted for by one human arousal rating against another, and 48%/62% (Warriner et al., 2013/Adelman & Estes, 2013 respectively) of the variance accounted for by one human valence rating against another.

These are very large differences, which underscore the problematic nature of thinking of human ratings as a "gold standard". Since our computed estimates are fixed, the differences in the estimates of variance accounted for are due entirely to variation in the human ratings themselves. When a "fixed measure" such as our computational estimates accounts for 30% of the variance accounted for by one set of human ratings and 68% of the variance accounted for by another set of human ratings of the same measure, we must ask of the human ratings: *Which human rating set is the most accurate measure of the construct of interest?* This question is impossible to answer *except by reference to a fixed standard*. It is perhaps not surprising that we believe that one can conclude that the Adelman and Estes (2013) ratings are the better ratings, since they correlate much better with our fixed estimates (recall that our estimates were derived from the Warriner et al., 2013 ratings). One might say it is tautological to make this claim, since we are justifying our own measure by a post hoc selection of the measure that best fits it. However, this is tautological in exactly the sense that it is tautological to say that the width of a sheet of paper (a measure of length) is 21.6 cm (another measure of length). Unless we agree to use some fixed unit of measure that has an empirically accessible definition, we cannot even compare different ratings in an intelligible way. One way of conceiving of the work presented here is that it attempts to define such a fixed standard estimate of arousal and valence, based on co-occurrence patterns in massive amounts of human-generated text rather than comparatively small samples of human judgements.

Assessing validity

We can assess the convergent validity of our computed values by checking to see whether they show expected correlations with other measures. It has previously been shown that imageability and valence judgements are positively correlated (i.e., words judged to be low imageability are also judged to have lower valence; Altarriba, Bauer, & Benvenuto, 1999; Kousta et al., 2011). In a recent paper using the same methods as those that we have used in this paper, an eight-item affect term co-occurrence measure (different from those discussed in this paper) correlated with 3700 human imageability judgements drawn from a variety of sources at $r = .55$ ($p < .00001$), collapsing over the development set and validation set items (Westbury et al., 2013).

Our computed valence measure correlated with same 3700 imageability judgements at $r = .10$ ($p < .00001$). To put this into context, we directly compared it to the correlation between the imageability judgements and Warriner et al. (2013) valence judgements for the 2532 words that appeared in both data sets. Across that subset of words, the correlation of imageability judgements with the Warriner et al. (2013) valence judgements was $r = .15$ ($p < .00001$), compared to a correlation of $r = .12$ ($p < .00001$) for the computed valence judgements. These two correlations were not reliably different by Fisher's r-to-z test ($Z = 1.09$, $p = .13$). Thus our computed valence measures show convergent validity by predicting imageability as well as human judgements of valence.

Note, however, that the radical and highly reliable difference in predictive power between the imageability-specific affect model developed in Westbury et al. (2013) ($r = .57$ for the same 2532 words) and the two valence measures suggests that it is not valence per se that is predictive of imageability, but rather that valence is correlated with a different measure of affective force that is itself very highly predictive of imageability.

The relation between arousal and imageability has not been previously studied. Our computer estimates of arousal ratings were correlated with the full set of 3700 imageability ratings at $r = -.05$ ($p = .002$) and with the subset rated by Warriner et al. (2013) at $r = -.14$ ($p < .00001$). By comparison, the Warriner et al. (2013) arousal ratings for the subset of 2532 words were not reliably

correlated with imageability ratings ($r = .02$, $p > .05$). In this case there is a dissociation between our estimates, which suggest a small but reliable negative correlation between arousal and imageability (imageable words are less arousing than nonimageable words), and the human arousal estimates, which suggest that there is no correlation at all.

Based on the relationship between imageability and valence and the fact that there is a well-known positive correlation between imageability and logged orthographic frequency ($r = .10$, $p < .00001$ for all 3700 imageability-rated words), it can be predicted that there should be a positive correlation between valence and logged frequency. Across all 3700 words, we do see this, both for valence ($r = .16$, $p < .00001$) and for arousal ($r = .20$, $p < .00001$). For the 2532 human-rated words only, there are statistically indistinguishable correlations with logged frequency of the computer-generated valence measures ($r = .17$, $p < .00001$) and the human valence ratings ($r = .18$, $p < .00001$; Fisher's r-to-z test: $Z = 0.37$, $p > .3$). However, the arousal ratings for these words show very different patterns, with logged frequency correlating at $r = -.06$ with the human arousal ratings and at $r = .16$ with the computer-generated arousal ratings (Fisher's r-to-z test: $Z = 7.8$, $p < .00001$). This is similar to the correlations with imageability ratings, with computer-generated valence looking very similar to human valence ratings, but arousal ratings looking quite different. Because arousal ratings are not very reliable, and we do not know which ratings we should take as a gold standard (as discussed above), it is not possible to make any sensible interpretation of these differences.

A more qualitative way of assessing the computational models is to judge their face validity by examining the words selected by the prediction equation as having both high and low valence or arousal. We applied the valence and arousal estimating equations above to a set of 23,211 words that had orthographic frequencies between 0.5 and 500 per million, excluding words outside that range because co-occurrence models are less reliable with very common or very uncommon words. We also excluded any words in that range that appeared in one of the predictor equations, for the reason noted earlier.

The 50 words rated highest and lowest on arousal are reproduced in the Supplemental Material, Appendix B. The words rated most arousing are GREED, HATRED, CHAOS, FEAR, VIOLENCE, BRUTALITY, REVENGE, MAYHEM, LOOTING, and PANICKED. These struck us as high-arousal words, especially compared to the least arousing words: NICE, HAPPY, WARM, I'VE, PORCH, BESIDE, COMFORTABLE, MILES, SAD, and CONDOLENCES. These low-arousal words suggest that arousal is characterized not so much by positive and negative extremes as by presence or absence (see discussion of Figure 3 below).

This contrasts with valence, which is characterized by positive and negative extremes. The 50 words rated highest and lowest on valence are provided in the Supplemental Material, Appendix C. The 10 most strongly valenced words are ENJOY, WONDERFUL, ROBUST, STRONGER, ENJOYING, SOLID, CELEBRATION, DELIGHT, ENJOYED, and STRENGTH, clearly all positive words associated with positive desire. The 10 least valenced words are RIOT, VIOLENCE, CHRONIC, PROBLEMS, RIOTING, SECTARIAN,

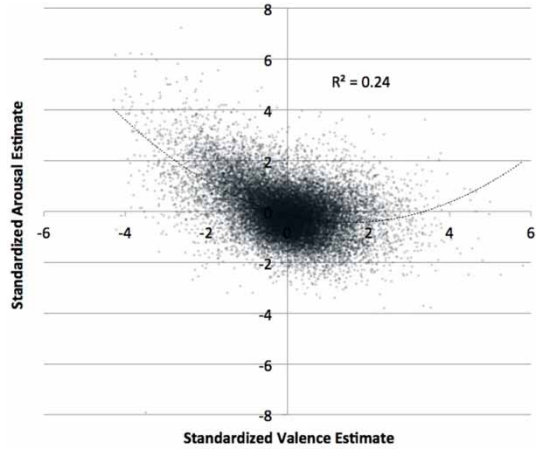

Figure 3. *Relationship estimated arousal and estimated valence of 23,211 words with orthographic frequencies between 0.5 and 500 per million.*

VIOLENT, OUTRAGE, INSECURITY, and TROUBLES, which are all undesirable and negative (or associated with undesirable and negative things).

Figure 3 graphs the relationship between estimated arousal and valence for all 23,211 words. As visual inspection and the fitted polynomial suggests, the relationship between arousal and valence is not a simple linear relationship; rather, we found the typical U-shaped relationship (Võ et al., 2009; see Larsen et al., 2008, for an analysis of the behavioural effects of this relationship). There are many very extreme words in the high-arousal, low-valence quadrant. There are a few, less extreme words in the high-arousal, low-valence category. Almost no words fall into the extreme end of the low-arousal quadrants, especially the low-arousal, low-valence quadrant, in part because (as mentioned above) arousal is characterized by absence or presence, being much more extreme at the high end (many extreme high-arousal words) than at the low end (few extreme low-arousal words).

As a final qualitative validity check, we judge the face validity of the interaction of estimated arousal and valence. Limiting the selection to extreme words (at least 2.0*z* on both dimensions) and selecting by the absolute average estimates of the two dimensions, the three highest valence, highest arousal words (addressing the question: *What are the best good things?*) are PASSION, EXCITEMENT, and EROTIC. The three words at the opposite extreme—low valence, low arousal (*What are some bad, not very exciting things?*)—are SADDENED, SORRY, and REGRET. The three most extreme words in the low-valence, high-arousal category (*What are the worst bad things?*) are VIOLENCE, HATRED, and GREED. Finally, the three most extreme words in the high-valence, low-arousal category (*What are some good, not very exciting things?*) are ENJOY, WONDERFUL, and HAPPY. These words all have face validity for their categories, offering reasonable, automatically generated answers to their associated questions.

Although tangential to the main focus of this paper, it is interesting to look at the lexical decision target words that come out highest on the co-occurrence-variable-only estimates (i.e., ignoring the lexical predictors also in that equation) for predicting LDRT, which consisted only of weighted co-occurrence similarities to two emotion labels: PLEASANT and UNPLEASANT. The top 20 words on this measure include the 12 words *inn*, *dining*, e*legant, cafe, rustic, restaurant, breakfast, dinner, wine, patio, picnic*, and *lounge*, all clearly related to dining out. This pattern continues for some time, with many words in the following 100 words also related to dining (e.g., *trendy, restaurants, dessert, elegance, tasty, chef, bistro, menu, tables*, and *cuisine*, to name just a few). One of the claims implicit in this line of research is that lexical semantics can be largely explained in terms of proximity to emotion words, or their proxies (i.e., nonemotion words that have emotional connotation due to their co-occurrence proximity to emotion terms). It remains a major challenge to understand what the appropriate weightings are for identifying different semantic categories. However, it is noteworthy that a simple weighting of just the two terms PLEASANT and UNPLEASANT suffices for picking out many words belonging to the very specific category of "having to do with dining out".

Computed valence and arousal judgements for the 23,211 words considered in this paper are available for download (Westbury, 2014).

GENERAL DISCUSSION

We began this paper with a discussion of "basic emotions" and a suggestion that one might gain insight into them by analysing how co-occurrence similarities from different sets of emotion labels predict valence ratings, arousal ratings, and their behavioural effects in lexical access. By these measures, only one of the 12 models that we considered was optimal: Co-occurrence distances from a subset of emotion labels in Wundt's model turned out to be better than any other model we could find at predicting LDRTs. For predicting arousal and valence, which are widely accepted as basic components of emotion, we derived emotion label sets that were significantly better than any of the 12 proposed sets of basic emotions. These models complicate and question the idea

that there may be "basic emotions", suggesting that the question of which emotions are basic needs to be more concretely specified by asking another question: *Basic for what?*

The best model we found for predicting arousal included three strongly weighted co-occurrence similarities to emotion labels related to automatic emotions: LUST, HUMILATION, and PANIC. These three words refer to very different states elicited by very different stimuli. This suggests that the construct of "arousal" might be broken down more finely, into arousal due to *sexual stimuli*, *social stimuli*, and *dangerous stimuli*.

Our best model of valence was very different and had a clear structure suggesting that valence consists of the four dimensions: *potency*, *happiness*, *approachability*, and *association with anger*. This also offers some suggestive routes to further studies of the construct of valence.

Finally, LDRT was best predicted by a very simple model that only included similarities to the emotion labels PLEASANT and UNPLEASANT, along with the interaction of these two measures. Although this seems very similar to valence, we found that the valence model could not be satisfactorily substituted for this two-predictor model and provided some evidence that the two models define distinct semantic dimensions.

In Westbury et al. (2014) we undertook a similar exercise to what we have undertaken here, using co-occurrence distances from the same set of emotion labels to predict human judgements of imageability compiled from a variety of sources. We met with a similar degree of success, accounting for a validated 58% of the variance in a set of nearly 2000 imageability judgements drawn from a variety of sources, and for 100% of the variance that was attributable to the imageability judgements in a lexical decision task. Similarly, Westbury (2013) used co-occurrence distance from a small set of emotion labels to predict about 50% of the variance in 1526 judgements of subjective familiarity and went on to show that for a new (validation) set of 699 words the regression model's estimates of subjective familiarity were statistically indistinguishable from independently collected human subjective familiarity judgements for the same words. The present findings thus add to a set of converging evidence that disparate sets of human judgements can be modelled using co-occurrence distances from emotion labels.

This work raises some general questions about how best to model effects of new predictors. Experimental psychologists have become used to being told that we must control for the effects of various (but idiosyncratically chosen) "well-known" variables before we can study the effects of any new predictors in which we may be interested. However, there is no way to do this that is immune from reasonable criticism. If we "fix" the effects of the control variables by regressing them out from the target variable before we try to predict that target variable with new predictors, we run the risk of having drawing erroneous conclusions, as we noticed when we examined the effect on LDRT residuals of distances from the emotion labels PLEASANT and UNPLEASANT in Experiment 2. The "true" effect of these predictors is to decrease LDRTs, as we can clearly see when we correlate them directly with those LDRTs: Distance from the emotion label PLEASANT is correlated with the 10,931 raw LDRTs from our subset of the Warriner et al. (2013) norms at $r = -.31$ ($p < .00001$), and distance from the word UNPLEASANT is correlated with those same measures at $r = -.24$ ($p < .00001$). However, recall that the relationship between the co-occurrence distance from UNPLEASANT and the residualized LDRTS (i.e., after controlling for lexical variables) was in the opposite direction. Controlling for variables by removing their effects in advance and then looking at correlations only with the resulting residuals can lead to misleading conclusions if those variables we control for do not have a uniform relationship to our predictors of interest.

If, on the other hand, we do not "fix" the effects of variables we do not want to study by removing them in advance, then we run into other related problems that are well known: namely, that our new predictors are probably going to be correlated with and interact in different ways with the variables that we are trying to control for, in which

case there is also no real sense in which those variables "have been controlled for". There is in fact no way to truly "control" for the effects of variables we want to ignore, except when they are either wholly uncorrelated with our new predictors (something so rare in psycholinguistics as to be virtually impossible) or when we have reason to believe that their effects are identical on all new predictors in which we may be interested.

It is trivially possible to make effects of any specific predictor larger or smaller by including or failing to include other predictors in our regression models. There is no consensus in the field of what variables are ontologically primal and which are not, in part because the question fades into incomprehensibility since almost all lexical predictors are correlated to some degree. To consider a highly salient example, researchers have agreed for years that orthographic frequency is "real" and is the most important predictor of human lexical access behaviour. However, Baayen (2010) showed that the effect of individual word frequency on lexical access was negligible or absent if one controlled (as Baayen, 2010, argued we should) for word co-occurrence probabilities before entering individual word frequency. So, *is there a word frequency effect, or not?* There is no unassailably objective answer to this question. Any answer one might give depends entirely on whether one believes that word co-occurrence has more claim to ontological primacy than single word frequency.

So long as we wish to retain our named predictors (rather than converting them to abstract orthogonal components whose theoretical meaning may be opaque), there are also no simple solutions to these problems, which point to some inherent limitations of correlational studies such as this one. We have tried to partially address these problems here by looking at our predictors from dual perspectives, both as overly collinear predictors and as uncorrelated residuals. We find much consistency in the general pattern of results across both methods: They both agree that it is much easier to predict valence from co-occurrence distance than to predict arousal, suggesting that valence is more lexicalized than arousal is; they both agree that distance from almost any emotion label seems to be predictive of valence judgements; and, they both agree that the small amount of variance that is predictable in arousal ratings from co-occurrence distances is due to distance from emotion labels associated with autonomic nervous system activity, most notably activity that underlies lust, anger, and fear. Both methods were extremely consistent in giving "rational" signs to emotion terms (i.e., positive signs for distance from positive emotion labels in predicting valence and negative signs for distance from negative emotion labels). This consistency provides some reason to believe that there is a true relationship between co-occurrence distances to emotion labels, and to emotional processing, especially with respect to valence.

Although further work using other methods will be required to validate any of the findings in this paper as being genuine insights into the structure of emotion, we offer this work as a methodological exercise in trying to ground the identification of the basic components of emotion in empirical data drawn from a large corpus of ordinary language. We have shown that this method is able to strongly predict human ratings and that it has convincing face validity. If indeed, as this work suggests, latent semantic emotionality differentials can be reliably detected in massive corpora of text generated by vast numbers of humans—particularly when the composition of any given corpus can have its content customized to reflect contributions exclusively from samples with particular population characteristics, something not considered in this study—it seems plausible to consider our method, or derivations thereof, as a useful way of estimating human emotionality judgements (perhaps with the added ability to capture group differences among subpopulations). This work demonstrates that mining large corpora of human-generated text for latent semantic information using co-occurrence distances is a useful methodology in the study of emotions.

Supplemental Material

Supplemental content is available via the "Supplemental" tab on the article's online page (http://dx.doi.org/10.1080/13506285.2014.970204).

Original manuscript received 17 January 2014
Accepted revision received 24 August 2014

REFERENCES

Adelman, J. S., & Estes, Z. (2013). Emotion and memory: A recognition advantage for positive and negative words independent of arousal. *Cognition*, *129*, 530–535.

Altarriba, J., Bauer, L. M., & Benvenuto, C. (1999). Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words. *Behavioral Research Methods, Instruments, & Computers*, *31*, 578–602.

Aristotle. (200 BCE/1941). *The basic works of aristotle*. (R. McKeon, Trans.). New York: Random House.

Baayen, H. R. (2010). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon*, *5*, 436–461.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*, 445–459.

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings (Tech.Rep. No. C-1)* CITY: University of Florida, The Center for Research in Psychophysiology, Gainesville, FL.

Briesemeister, B. B., Kuchinke, L., & Jacobs, A. M. (2014). Emotion word recognition: Discrete information effects first, continuous later?. *Brain Research*, *1564*, 62–71. doi:10.1016/j.brainres.2014.03.045

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

Bullinaria, J. A., & Levy, J. P. (2012). Extracting semantic representations from word co- occurrence statistics: Stop-lists, stemming, and SVD. *Behavior Research Methods*, *44*, 890–907.

Burgess, C., & Lund, K. (2000). The dynamics of meaning in memory. In E. Dietrich and A. B. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines* (pp. 117–156). Mahwah, NJ: Lawrence Erlbaum Associates.

Darwin, C. (1872). *The expressions of emotions in man and animals*. New York, NY: Philosophical Library.

Durda, K., & Buchanan, L. (2008). Windsors: Windsor improved norms of distance and similarity of representations of semantics. *Behavior Research Methods*, *40*, 705–712.

Ekman, P. (1980). Biological and cultural contributions to body and facial movement in the expression of emotions. In A. O. Rorty (Ed.), *Explaining emotions* (pp. 73–102). Berkeley, CA: University of California Press.

Ekman, P. (1999). Basic emotions. In T. Dalgleish, & M. Power (Eds.), *Handbook of cognition and emotion* (pp. 45–60). Chichester, UK: John Wiley and Sons.

Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science*, *164*, 86–88.

Fehr, B., & Russell, J. A. (1984). Concept of emotion viewed from a prototype perspective. *Journal of Experimental Psychology: General*, *113*, 464–486.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review*, *103*, 518–565.

Hofmann, M. J., Kuchinke, L., Biemann, C., Tamm, S., & Jacobs, A. M. (2011). Remembering words in context as predicted by an Associative Read-Out Model. *Frontiers in Psychology*, *2*, 1–11, article no. 252.

Izard, C. E. (1977). *Human emotions*. New York, NY: Plenum.

Izard, C. E. (1992). Basic emotions, relations amongst emotions and emotion–cognition relations. *Psychological Review*, *99*, 561–565.

Jacobs, A. M., & Grainger, J. (1994). Models of visual word recognition: Sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 1311–1334.

Johnson-Laird, P. N., & Oatley, K. (1989). The language of emotions: An analysis of a semantic field. *Cognition and Emotion*, *3*, 81–123.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*, 1–37.

Kassam, K. S., Markey, A. R., Cherkassky, V. L., Loewenstein, G., & Just, M. A. (2013). Identifying emotions on the basis of neural activation. *PLoS ONE*, *8*, e66032. doi:10.1371/journal.pone.0066032

Kousta, S., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General*, *140*, 14–34.

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*, 211–240.

Larsen, R. J., Mercer, K. A., Balota, D. A., & Strube, M. J. (2008). Not all negative words slow down lexical decision and naming speed: Importance of word arousal. *Emotion*, *8*, 445–452.

Levenson, R. W. (2003). Autonomic specificity and emotion. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 212–224). Oxford: Oxford University Press.

Lifchitz, A., Jhean-Larose, S., & Denhière, G. (2009). Effect of tuned parameters on an LSA multiple choice questions answering model. *Behavior Research Methods*, *41*, 1201–1209.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*, 203–208.

Morris, C. W. (1946). *Signs, language, and behavior*. New York: Prentice Hall.

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin*, *49*, 197–237.

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

Panksepp, J. (1982). Toward a general psychobiological theory of emotions. *Behavioral and Brain Sciences*, *5*, 407–467.

Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions*. New York, NY: Oxford University Press.

Panksepp, J. (2005). Affective consciousness: Core emotional feelings in animals and humans. *Consciousness and Cognition*, *14*, 30–80.

Panksepp, J. (2007). Neurologizing the psychology of affects: How appraisal-based constructivism and basic emotion theory can coexist. *Perspectives on Psychological Science*, *2*, 281–296.

Panksepp, J. (2008a). Cognitive conceptualism – Where have all the affects gone?. Additional corrections for Barrett et al. (2007). *Perspectives on Psychological Science*, *3*, 305–308.

Panksepp, J. (2008b). The power of the word may reside in the power of affect. *Integrative Psychological and Behavioral Science*, *42*, 47–55. doi:10.1007/s12124-007-9036-5.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In R. Plutchik, & H. Kellerman (Eds.), *Emotion: Theory, research, and experience: Vol. 1. Theories of emotion* (pp. 3–33). New York: Academic.

Reisenzein, R. (2009). Emotional experience in the computational belief-desire theory of emotion. *Emotion Review*, *1*, 214–222.

Rhode, D. L. T., Gonnerman, L. M., & Plaut, D. C. (2007). *An improved method for deriving word meaning from lexical co-occurrence*. Unpublished manuscript. Cambridge, MA: Massachusetts Institute of Technology. Retrieved April 20, 2007, from http://tedlab.mit.edu/~dr/

Robinson, M. D., Storbeck, J., Meier, B. P., & Kirkeby, B. S. (2004). Watch out! That could be dangerous: Valence-arousal interactions in evaluative processing. *Personality and Social Psychology Bulletin*, *30*, 1472–1484.

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, *4*, 328–350.

Rosch, E. H. (1978). Principles of categorization. In E. Margolis, & S. Laurence (Eds.), *Concepts: Core readings* (pp. 189–206). Cambridge, MA: MIT Press.

Ryle, G. (1949). *The concept of mind*. London: Hutchinson.

Scaratino, A., & Griffiths, P. (2011). Don't give up on basic emotions. *Emotion Review*, *3*, 444–454.

Shaoul, C., & Westbury, C. (2006a). USENET Orthographic Frequencies for 111,627 English Words. (2005–2006) Edmonton, AB: University of Alberta. Retrieved from http://www.psych.ualberta.ca/~westburylab/downloads/wlfreq.download.html

Shaoul, C., & Westbury, C. (2006b). Word frequency effects in high-dimensional co-occurrence models: A new approach. *Behavior Research Methods*, *38*, 190–195.

Shaoul, C., & Westbury, C. (2008). *HiDEx: The high dimensional explorer*. Edmonton, AB. Retrieved from http://www.psych.ualberta.ca/~westburylab/downloads.html

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods*, *42*, 393–413.

Shaoul, C., & Westbury, C. (2011). HiDEx: The high dimensional explorer. In P. McCarthy, & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution*, 230–246. IGI Global.

Shaver, P., Schwartz, J., Kirson, D., & O'Connor, C. (1987). Emotion and emotion knowledge: Further explorations of a prototype approach. *Journal of Personality and Social Psychology*, *52*, 1061–1086.

Stevenson, R. A., Mikels, J. A., & James, T. W. (2007). Characterization of the affective norms for English words by discrete emotional categories. *Behavior Research Methods*, *39*, 1020–1024.

Tomkins, S. S. (1962). *Affect, imagery, consciousness: Vol. I: The positive affects*. New York: Springer.

Tomkins, S. S. (1963). *Affect, imagery, consciousness: Vol. II: The negative affects*. New York: Springer.

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*, 534–538.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*, 1191–1207.

Westbury, C. (2013). Y16ou can't drink a word: Lexical and individual emotionality affect subjective familiarity judgments. *Journal of Psycholinguistic Research*, *43*, 631–649. doi:10.1007/s10936-013-9266-2

Westbury, C. (2014). *Human judgments estimated from co-occurrence with affect terms*. Retrieved from http://www.psych.ualberta.ca/~westburylab/downloads/AffectEstimates.download.html

Westbury, C. F., Briesemeister, B. B., Hofmann, M. J., & Jacobs, A. M. (2014). Giving meaning to emotional valence: Co-occurrence models can help. Manuscript submitted for publication.

Westbury, C. F., Shaoul, C., Hollis, G., Smithson, L., Briesemeister, B. B., Hofmann, M. J., & Jacobs, A. M. (2013). Now you see it, now you don't: On emotion, context, & the algorithmic prediction of human imageability judgments. *Frontiers in Psychology*, *4*, 991. doi:10.3389/fpsyg.2013.00991.

Wundt, W. (1896). *Grundriss der Psychologie [Outlines of psychology]*. Leibzig: Engelmann.