Norms of valence, arousal, and dominance for 13,915 English lemmas

Amy Beth Warriner [1] Victor Kuperman [1,*] Marc Brysbaert [2]

[1] McMaster University, Canada

[2] Ghent University, Belgium

Keywords: emotion, semantics, gender differences, age differences, crowdsourcing

*Corresponding author: Victor Kuperman, Ph.D.

Department of Linguistics and Languages, McMaster University
Togo Salmon Hall 626
1280 Main Street West
Hamilton, Ontario, Canada L8S 4M2
phone: 905-525-9140, x. 20384
vickup@mcmaster.ca

Abstract

Information about the affective meaning of words is used by researchers working on emotions and moods, word recognition and memory, and text-based sentiment analysis. Three components of emotions are traditionally distinguished: valence (the pleasantness of the stimulus), arousal (the intensity of emotion provoked by the stimulus), and dominance (the degree of control exerted by the stimulus). Thus far, nearly all research has been based on the ANEW norms collected by Bradley and Lang (1999) for 1,034 words. We extend the database to nearly 14 thousand English lemmas, providing researchers with a much richer source of information, including information on gender, age and educational differences in emotion norms. As an example of the new possibilities, we included the stimuli from nearly all category norms (types of diseases, occupations, and taboo words) collected by Van Overschelde, Rawson,and Dunlosky (2004), making it possible to include affect in studies on semantic memory.

Emotional ratings of words are in high demand, because they are used in at least four lines of research. The first line concerns research on the emotions themselves: the ways in which they are produced and perceived, their internal structure, and the consequences they have on human behavior. For instance, Verona, Sprague, and Sadeh (2012) used emotionally neutral and negative words in an experiment comparing responses of offenders without a personality disorder to offenders with an antisocial personality disorder who either had additional psychopathic traits or not.

The second line of research deals with the impact that emotional features have on the processing and memory of words. Kousta, Vinson, & Vigliocco (2009) found that participants responded faster to positive and negative words than to neutral words in a lexical decision experiment, a finding later replicated by Scott, O'Donnell, and Sereno (2012) in sentence reading. According to Kousta, Vigliocco, Vinson, Andrews, and Del Campo (2011) emotion is particularly important in the semantic representations of abstract words. In other research, Fraga, Pineiro, Acuna-Farina, Redondo, and Garcia-Orza (2012) reported that emotional words are more likely to be used as attachment sites for relative clauses in sentences such as "Someone shot the servant of the actress who …".

A third approach uses emotional ratings of words to estimate the sentiment expressed by entire messages or texts. Leveau, Jean-Larose, Denhière, and Nguyen (2012), for instance, wrote a computer program to estimate the valence and arousal evoked by texts on the basis of word measures (see also Liu, 2012).

Finally, emotional ratings of words are used to automatically estimate the emotional values of new words by comparing them to validated words. Bestgen and Vincze (2012) gauged the affective values of 17,350 words by using rated values of words that were semantically related.

So far, nearly all studies have been based on Bradley and Lang's (1999) *Affective Norms for English Words (ANEW)* or translated versions (for exceptions see Kloumann et al., 2012; Mohammad & Turney, 2010) . These norms contain ratings for 1034 words. There are three types of ratings, in line with Osgood, Suci, and Tannenbaum's (1957) theory of emotions. The first, and most important, concerns the valence (or pleasantness) of the emotions invoked by the word, going from unhappy to happy. The second addresses the degree of arousal evoked by the word. The third dimension refers to the dominance/power of the word, the extent to which the word denotes something that is weak/submissive or strong/dominant.

The number of words covered by the ANEW norms appeared sufficient for use in small-scale factorial experiments. In these experiments, a limited number of stimuli are selected that vary on one dimension (e.g., valence) and are matched on other variables (e.g., arousal, word frequency, word length and others). However, this number is prohibitively small for the large-scale megastudies that are currently emerging in psycholinguistics. In these studies (e.g., Balota et al., 2007; Ferrand et al., 2010; Keuleers et al., 2010, 2012), regression analyses of thousands of words are used to disentangle the influences on word recognition. The ANEW norms are also limited as input for computer algorithms gauging the sentiment of a message/text or the emotional values of non-rated words.

Given the ease with which word norms can be collected nowadays, we decided to collect affective ratings for a majority of well-known English content words (for a total of 13,915). Because it can be expected that the emotional values generalize to inflected forms (e.g. *sings, sang, sung, singing* for verb lemma *sing*), we only included lemmas (these are the base forms of the words, the ones that are used as entries in dictionaries). Our sample of words (see below for the selection criteria) substantially covers the word-stock of the English language and forms a solid foundation to automatically derive the values of the remaining words (Bestgen & Vincze, 2012).

METHOD

Stimuli

Words included in our stimuli set were compiled from three sources: Bradley and Lang's (1999) Affective Norms for English Words (ANEW), Van Overschelde, Rawson, & Dunlosky (2004) Category Norms, and the SUBTLEX-US corpus (Brysbaert & New, 2009). Our final set included 1029 of the 1034 words from ANEW (5 were lost due to programmatic error) and 1060 of the participant-generated responses to 60 out of the 70 category names included in the Category Norms study (we did not include categories such as units of time and distance, or types of fish) . The remaining words were selected from the list of 30 thousand lemmas for which Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012) collected Age of Acquisition ratings. This list contains the content lemmas (nouns, verbs, and adjectives) from the 50 million-token SUBTLEX-US subtitle corpus. We only selected the highest-frequency words known by 70% or more of the participants in Kuperman et al. (2012), given that affective ratings are less valid/useful for words not known to most participants. Our final set included 13,915 words, 22.5% of which are most often used as adjectives (Brysbaert, New, & Keuleers, 2012), 63.5% as nouns, 12.6% as verbs, and 1.4% as other or unspecified parts of speech. The mean word frequency of the set was 1,056 (SD = 8464, range = 1:314232, median = 87) in the 50 million-token SUBTLEX-US corpus: 152 words or 1% had no frequency data. For each word in our set, we collected ratings on three dimensions using a 9 point scale.

The stimuli were distributed over 43 lists of 346 to 350 words each. Each list consisted of 10 calibrator words, 40 control words from ANEW, and a randomized selection of non-ANEW words. The calibrator words were drawn from ANEW and were chosen separately for each of the three dimensions with the goal of giving participants a sense of the entire range of stimuli they would encounter[1]. Participants always saw these calibrator words first. The remaining ANEW words were divided into sets of 40 and

[1] Calibrator words for the respective dimensions were as follows (in the increasing order of ratings): Valence: jail (1.91), invader (2.23), insecure (2.30), industry (5.07), icebox (5.67), hat (5.69), grin (7.66), kitten (7.58), joke (7.88), free (8.25). Arousal: statue (2.82), rock (3.14), sad (3.49), cat (4.50), curious (5.74), robber (6.20), shotgun (6.55), assault (6.80), thrill (7.19), sex (7.60). Dominance: lightning (4.00), mildew (4.19), waterfall (5.34), wealthy (6.11), lighthouse (6.24), honey (6.39), treat (6.66), mighty (6.85), admired (6.94), liberty (7.04)

served as controls for the estimation of correlations between our data and the ANEW norms. This meant that a selection of these words appeared in more than one list and that the lists used for each of the three dimensions were mostly, but not completely, identical. The control words and the non-ANEW words were randomly mixed together in each list. Once created, the words in each list were always presented in a fixed order following the calibrator words.

Data Collection

Participants were recruited via Amazon Mechanical Turk's crowdsourcing website. Responders were restricted to those who self-identified as current residents of the US and to completing any given list only once. This completion of a single list by a given participant is henceforth referred to as an assignment. Each assignment involved rating words on a single dimension only, in contrast to the ANEW study where participants rated each word on all three dimensions. The instructions given were minor variations on the instructions in the ANEW project and are given below with the respective changes to the wording for the separate dimensions indicated in square brackets.

You are invited to take part in the study that is investigating emotion, and concerns how people respond to different types of words. You will use a scale to rate how you felt while reading each word. There will be approximately 350 words. The scale ranges from 1 (happy [excited; controlled]) to 9 (unhappy [calm; in control]). At one extreme of this scale, you are happy, pleased, satisfied, contented, hopeful [stimulated, excited, frenzied, jittery, wide-awake, or aroused; controlled, influenced, cared-for, awed, submissive, or guided]. When you feel completely happy [aroused; controlled] you should indicate this by choosing rating 1. The other end of the scale is when you feel completely unhappy, annoyed, unsatisfied, melancholic, despaired, or bored [relaxed, calm, sluggish, dull, sleepy, or unaroused; in control, influential, important, dominant, autonomous, or controlling]. You can indicate feeling completely unhappy [calm; in control] by selecting 9. The numbers also allow you to describe intermediate feelings of pleasure [calmness/arousal; in/under control], by selecting any of the other feelings. If you feel completely neutral, neither happy nor sad [not excited nor at all calm; neither in control nor controlled], select the middle of the scale (rating 5).

Please work at a rapid pace and don't spend too much time thinking about each word. Rather, make your ratings based on your first and immediate reaction as you read each word.

On average, assignments were completed in approximately 14 minutes. Participants received 75 cents per completed assignment. After reading an informational consent statement and the instructions, participants were asked to indicate their age, gender, first language(s), country/state resided in most between birth and age 7, and educational level. Subsequently, they were reminded of the scale anchors and presented with a scrollable page in which all words in the list were shown to the left of nine numbered radio buttons. Although we did not incorporate the self-assessment manikins (SAM) that were used in the ANEW study, we did anchor our scales in the same direction with Valence ranging from happy to unhappy, Arousal from excited to calm, and Dominance from controlled to in control. In the Results section, we show that our numerical ratings correlate highly with the SAM ratings from ANEW demonstrating that the methods are roughly equivalent. Once finished, participants clicked 'Submit' to complete the study.

Lists were initially presented to 20 respondents each. However, missing values due to subsequent exclusion criteria resulted in some words having less than 18 valid ratings. Several of the lists were re-posted until the vast majority of words reached at least this threshold. Data collection began March 14, 2012 and was completed May 30, 2012.

RESULTS AND DISCUSSION

Data Trimming

Altogether, 1,085,998 ratings were collected across all three dimensions. Around 3% of the data were removed due to missing responses, lack of variability in responses (i.e. providing the same rating for all words in the list), or the completion of less than 100 ratings per assignment. Valence and Arousal ratings were reversed post-hoc to maintain a more intuitive low to high scale (e.g. sad to happy rather than happy to sad) across all three dimensions. Means and standard deviations were calculated for each word. Ratings in assignments with negative correlations between a given participant's rating and the mean for that word were reversed (9%). This was done based both on empirical evidence that higher numbers intuitively go with positive anchors (Rammstedt and Krebs, 2007) and an examination of these participants' responses which revealed unintuitive answers (e.g. indicating that negative words such as 'jail' made them very happy). Any remaining assignments with ratings correlating with mean ratings per items at less than .10 were removed and the means and standard deviations were re-calculated. The final data set consisted of 303,539 observations for Valence (95% of the original data pool), 339,323 observations for Arousal (89% of the original data pool) and 281,735 observations for Dominance (74% of the original data pool). A total of 1827 responders contributed to this final data set with 362 of them completing assignments for 2 or more dimensions. 144 participants completed two or more assignments within a single dimension.

For Valence, 51 words received less than 18 (but more than 15) valid ratings. For Arousal, this number was 128. For Dominance, 564 words had between 16 and 17 ratings and 17 words had between 14 and 15 ratings each. In all three cases, more than 87% of words had between 18 and 30 ratings per word. 50 words in each dimension received more than 70 ratings each due to the doubling up of ANEW words and the re-running of lists. To illustrate how our data enriches the set of words available in the ANEW, Table 1 provides examples of words that are not included in the ANEW list and show very high or very low ratings in any of the three dimensions.

Table 1: Words at the extreme of each dimension that were not included in ANEW.

	Valence		Arousal		Dominance	
Lowest	pedophile	1.26	grain	1.60	dementia	1.68
	rapist	1.30	dull	1.67	Alzheimer's	2.00

	AIDS	1.33	calm	1.67	lobotomy	2.00
	leukemia	1.47	librarian	1.75	earthquake	2.14
	molester	1.48	soothing	1.91	uncontrollable	2.18
	murder	1.48	scene	1.95	rapist	2.21
Highest	excited	8.11	motherfucker	7.33	rejoice	7.68
	sunshine	8.14	erection	7.37	successful	7.71
	relaxing	8.19	terrorism	7.42	smile	7.72
	lovable	8.26	lover	7.45	completion	7.73
	fantastic	8.36	rampage	7.57	self	7.74
	happiness	8.48	insanity	7.79	incredible	7.74

Demographics

Of the 1827 valid responders, approximately 60% were female in all three cases (419 Valence, 448 Arousal, and 505 Dominance). Their ages ranged from 16 to 87 years with 11% younger than 20 years old; 45% between 21 and 30; 21% between 31 and 40; 11% between 41 and 49; and 12% age 50 or older. 24 (3.3%), 32 (4.3%), and 23 (2.7%) participants in each dimension respectively reported a native language other than English while 10 (1.4%), 12 (1.6%), and 12 (1.4%) participants respectively reported more than one native language, including English. Table 2 shows the number of participants at each of the seven possible education levels. Most had some college or a bachelor's degree.

Table 2: Reported education levels within each dimension

Education Level	Number of Participants		
	Valence (%)	Arousal (%)	Dominance (%)
Some High School	28 (4)	32 (4)	28 (3)
High School Graduate	96 (13)	98 (13)	117 (14)
Some College – No Degree	237 (33)	252 (34)	298 (35)
Associates Degree	82 (11)	79 (11)	93 (11)
Bachelors Degree	212 (29)	222 (30)	218 (26)

Masters Degree	55 (8)	53 (7)	78 (9)
Doctorate	13 (2)	9 (1)	13 (2)
TOTAL	*723*	*745*	*845*

Note: The numbers across all three columns add up to more than 1827 as some people contributed to more than one dimension.

Descriptive Statistics

Table 3 reports descriptive statistics for the three distributions of ratings. Distributions of both valence and dominance ratings are negatively skewed (G_1= -0.28 and -0.23 respectively) with 55% of the words rated above the median of the rating scale for both dimensions, see Figure 1. The Mann-Whitney one-sample median test indicates that the medians of both the valence and dominance distributions are not significantly different from rating 5, which is the median of the scales (both $p > 0.1$). The tendency for more words to make people feel happy and in control goes along with numerous former findings that there is a positivity bias in English and other languages (see Augustine, Mehl, & Larsen, 2011 and Kloumann et al., 2012). The positivity bias – or the prevalence of positive word types in English books, Twitter messages, music lyrics and other genres of texts – is argued to reflect the preference of humankind for pro-social and benevolent communication. Arousal, on the other hand, is positively skewed (G_1 = 0.47), meaning that only a relatively small proportion of words (20% above rating 5) make people feel excited.

Figure 1: Distributions of valence (green), arousal (red) and dominance (blue) ratings. Dotted lines represent the medians of respective distributions.

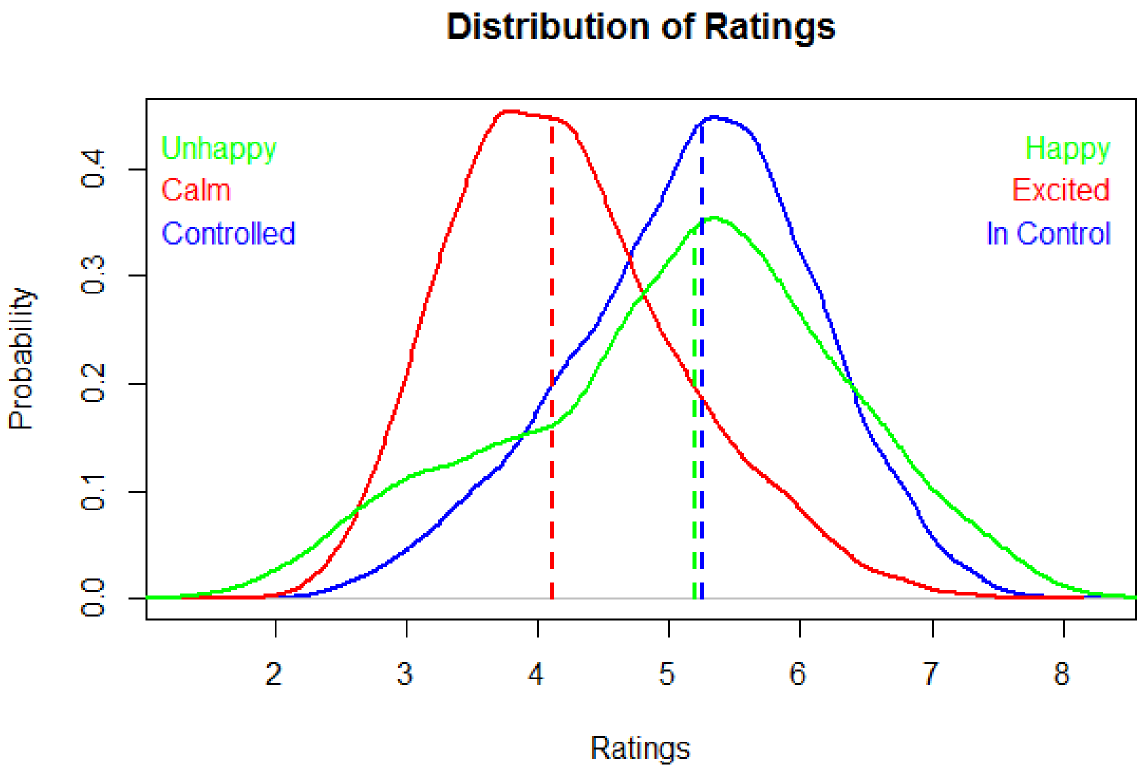


Table 3: Descriptive statistics for the distribution of each dimensions, including the number of participants (N), number of observations, average mean and average SD.

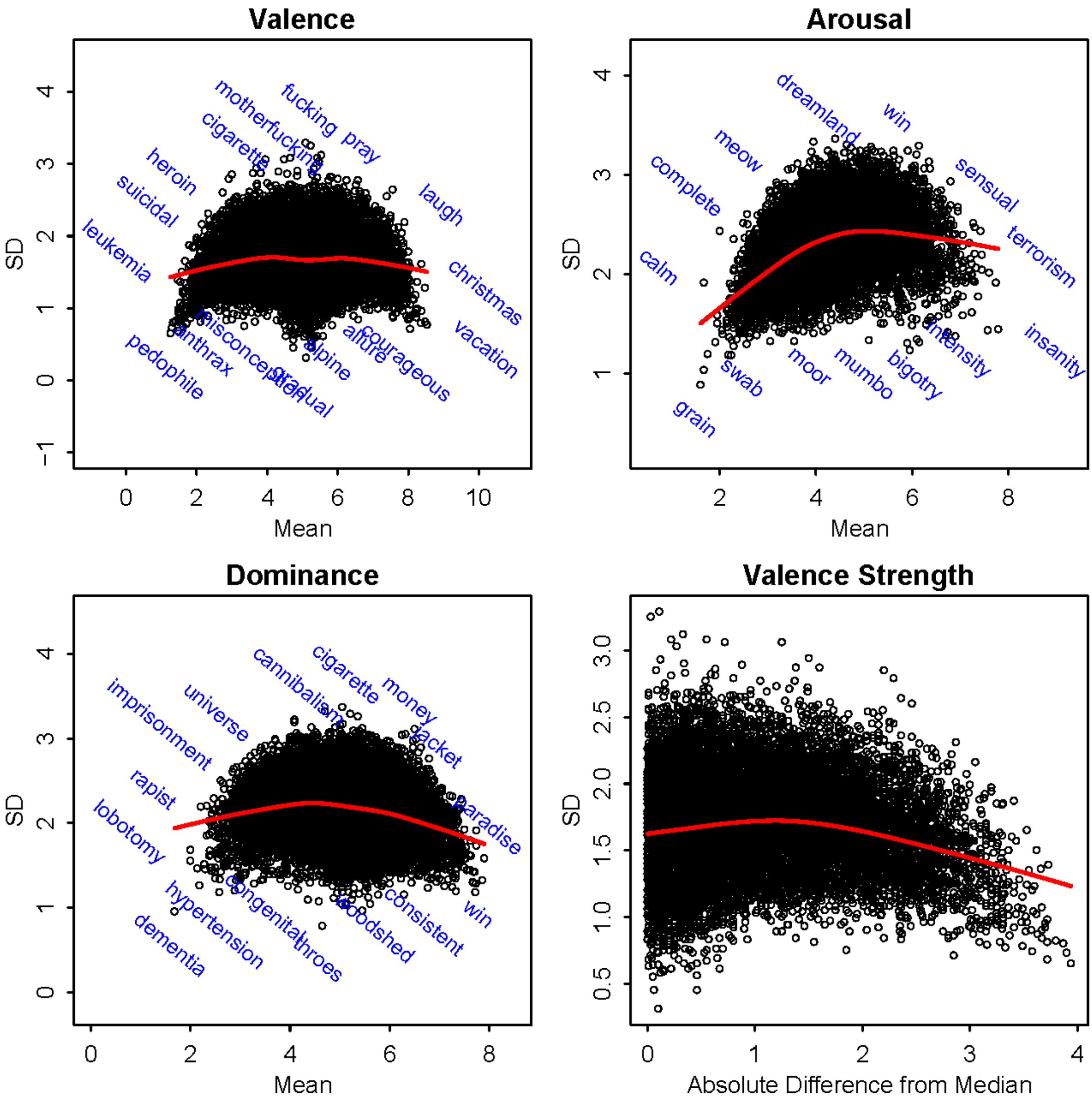
	N	# of Obs	Mean	Avg SD
Valence	723	303,539	5.06	1.68
Arousal	745	339,323	4.21	2.30
Dominance	845	281,735	5.18	2.16

Ratings of Valence are relatively consistent across participants while Arousal and Dominance are much more variable. This is indicated by the difference between average standard deviations of dimensions: 1.68 for Valence but 2.30 and 2.16 for Arousal and Dominance respectively. In addition, the split-half reliabilities were .914 for Valence, .689 for Arousal, and .770 for Dominance: see below for other examples of a higher variability of dominance and arousal ratings. Figure 2a-c shows, for the three emotional dimensions, the means of the ratings for each word plotted against their standard deviations, with the scatterplot smoother lowess line demonstrating the overall trend in the data (red solid line).

For illustrative purposes, each plot is supplied with selected examples of words that are substantially more or less variable than other words with the given mean rating. Swear words, taboo words and sexual terms account for a disproportionally large number of words that elicit more variable ratings of valence and arousal than expected given the words' mean ratings (shown as words in blue above the red lowess line in Figures 2a-c respectively), in line with Kloumann et al. (2012). Below we demonstrate that the exceeding variability in such words may be due to gender differences in norms.

For valence, the scatterplot in Figure 2a (top left) is symmetrical about the median, with relatively positive or negative words associated with a smaller variability in ratings across participants as compared to valence-neutral words (see Moors et al., in press, for a similar finding in Dutch). The same holds for the pattern observed in dominance ratings, Figure 2c (bottom left). The plot of valence *strength* (absolute difference between the valence rating and the median of valence ratings, Figure 2d) corroborates the tendency of more extreme (positive or negative) words to be less variable in ratings than neutral ones. In contrast, for arousal in Figure 2b (top right), words that make people feel calm generally elicit more consistent ratings than those that make people feel excited. To sum up, in terms of variability of ratings, valence and dominance pattern together and are best considered in terms of their magnitude (how strong is the feeling) rather than their polarity (sad vs happy, or controlled-by vs in-control); polarity, however, determines the variability in arousal ratings.

Figure 2: Standard deviation of ratings for valence (a, top left), arousal (b, top right), dominance (bottom left) and valence strength (d, bottom right) plotted against respective mean ratings. Panels a-c also provide examples of words with disproportionately large and small standard deviations given their mean.



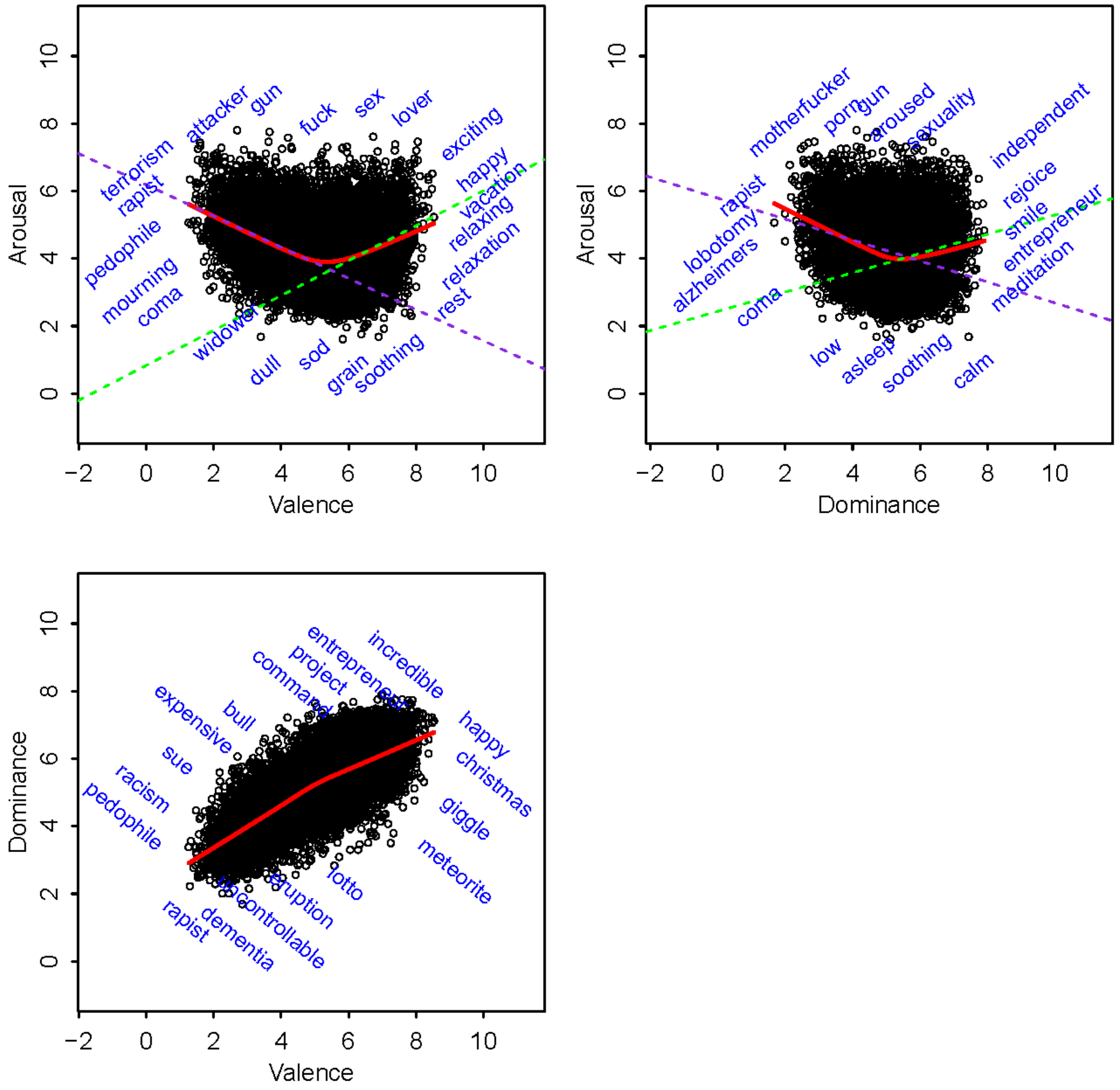
Correlations between Dimensions

We found the typical U-shaped relationship between arousal and valence (see Figure 3a; Bradley and Lang, 1999; Redondo et al., 2007; Soares, et al., 2012. Words that are very positive or very negative are more arousing than those that are neutral. This is corroborated by the positive correlation between valence and arousal for positive words (mean valence rating > 6, r = .273, p < .001) and the negative correlation between valence and arousal for negative words (mean valence rating < 4, r = -.293, p < .001). The relationship between valence and dominance is linear, with words that make people feel happier also making them feel more in control (see Figure 3b). There is another U-shaped relationship between arousal and dominance (see Figure 3c) corroborated by the positive correlation between dominance and arousal for high rated dominance words (mean rating > 6, r = .139, p < .001) and a negative correlation between dominance and arousal for low rated dominance words (mean rating < 4, r = -.193, p < .001). Table 4 shows that a quadratic relationship between arousal and valence and between arousal and dominance explains more of the variance than a linear relationship. However, this does not rule out the possibility that the high and low levels of these associations might better be explained by a regression with the breakpoint at the median of the scale (see Figure 3)The relationship between dominance and valence, however, is fitted better by a linear model.

Table 4: Pearson's correlations, linear and quadratic coefficients and the quadratic R^2 for each dimension. For both arousal/valence and arousal/dominance, the quadratic relationship explains more variance than the linear function.

	R	Linear Coefficient	Quadratic Coefficient	R^2
Arousal and Valence	-0.185	-0.130	34.883	0.143
Dominance and Valence	0.717	0.974	-	0.518
Arousal and Dominance	-0.180	-0.172	21.842	0.075

Figure 3: Scatterplots of dimensions (top left, arousal vs valence; top right, arousal vs dominance; bottom left, Dominance vs Valence) along with lowess lines (in red) showing the functional relationships and regression lines for arousal as predicted by high (in green) and low (in purple) valence and dominance. Sample words have also been included.



The strength of the correlation between dominance and valence casts doubt on the claim that the three dimensions under consideration here are genuinely orthogonal affective states. This assumption was the basis of the original ANEW study (Bradley and Lang, 1999), stemming from original factor analyses done by Osgood, Suci, & Tanenbaum (1957). Future research will have to demonstrate that dominance explains unique variance over and above valence in the language processing behavior. The fact that extreme values of valence and dominance are more arousing point again at the utility of considering valence/dominance strength (how different is the word from neutral) rather than polarity as the explanatory variable. We return to this point below.

Reliability

We compared our ratings with several smaller sets of ratings that had been collected previously by other researchers, including the ANEW set from which we drew our control words. The correlations are listed in Table 5.

Table 5: Correlations of present ratings with similar studies across languages

Data Set				Correlations		
Source	Language	N (source)	N (overlap)	Valence	Arousal	Dominance
a	English	1040	1029	.953	.759	.795
b	Dutch	4299	3701	.847	.575	N/A
c	Spanish	1034	1023	.924	.692	.833
d	Portuguese	1040	1023	.924	.635	.774
e	Finnish	213	203	.956	N/A	N/A
f	English	10222	4504	.919	N/A	N/A

Sources: [a] Bradley & Lang (1999); [b] Moors et al., (in press) – English glosses; [c] Redondo, Fraga, Padrón, & Comesaña, (2007) – English glosses; [d] Soares, Comesaña, Pinheiro, Simões, & Frade (2012) – English glosses; [e] Eilola & Havelka (2010) – English glosses; [f] Kloumann, Danforth, Harris, Bliss, & Dodds (2012); All studies except Moors et al., (in press) utilized a 9-point scale in acquiring their ratings. Moors et al., (in press) used a 7-point scale.

Valence appears to generalize very well across studies and languages, as evidenced by high correlations. Both arousal and dominance showed more variability across languages and studies as reflected in the lower correlations. Note that these studies themselves (those that reported the information – c, d, and e) also found a lower correlation between their arousal and dominance ratings and the arousal and dominance ratings reported in other papers (arousal range: .65 to .75; dominance range: .72 to .73). Importantly, however, cross-linguistic correlations were stronger (range of Pearson's r for arousal was .575 - .759) than those between gender, age and education groups within our study (range of Pearson's r was .467-.516), see Table 8 below. This observation clearly indicates the validity of using emotional ratings to English glosses of words in a language which does not have an extensive set of ratings at the researcher's disposal. This seems to be more the case for valence and dominance than for arousal.

Correlations with Lexical Properties

As known for other subjective ratings of lexical properties (cf. Baayen, Feldman, & Schreuder, 2006), judgments of the emotional impact of a word are likely to be affected by other aspects of the words' meaning. Table 6 reports correlations of valence, arousal and dominance with a range of available semantic variables. In the remainder of the paper, words, rather then the trial-level data, were chosen as units of correlational analyses.

Table 6: Correlations between emotional dimensions and semantic variables reported in prior studies (degrees of freedom are based on the number of datapoints reported as N (overlap)).

Source	Measure	N (source)	N (overlap)	Valence	Arousal	Dominance
a	Imageability	5,988	5,125	0.161	-0.012	0.031
b	Imageability	326	318	-0.037	0.099 $^{+}$	-0.160
	Concreteness	326	318	0.109 $^{+}$	-0.244	-0.019
	Context Avail.	326	318	0.196	-0.147	0.044
c	Concreteness	1,944	1,567	0.105	-0.258	0.009
d	Imageability	3,394	2,906	0.152	-0.045	0.006
	Familiarity	3,394	2,906	0.206	-0.028	0.215
e	AoA	30,121	13,709	-0.233	-0.062	-0.187
	% Known	30,121	13,709	.094	0.078	0.103
f	Sensory Exp	5,857	5,007	0.067	0.228	-0.044
g	Body-Object	1,618	1,398	0.203	-0.143	0.172
h	Familiarity	559	503	0.272	-0.193	0.329
	Pain	559	503	-0.456	0.579	-0.343
	Smell	559	503	0.139	0.052	-0.043
	Color	559	503	0.401	0.052	0.081
	Taste	559	503	0.309	-0.102	0.084
	Sound	559	503	-0.176	0.407	-0.286
	Grasp	559	503	0.024	-0.121	0.252
	Motion	559	503	-0.113	0.328	-0.328
i	Sound	1,402	1,283	-0.04	0.311	-0.121
	Color	1,402	1,283	0.322	-0.072	0.100

	Manipulation	1,402	1,283	0.070 *	0.026	0.255
	Motion	1,402	1,283	0.011	0.335	-0.140
	Emotion	1,402	1,283	0.902	-0.206	0.658
j	Log Frequency	74,286	13,763	0.182	-0.033	0.167

Note 1: The overlapping words in this study represent a biased sample due to the fact that words in the current study were restricted to only include words that were known by 70% or more participants in the study cited here.

Note 2: Since we chose words to fill our quota that were higher in frequency, the overlap here is also biased towards the upper range.

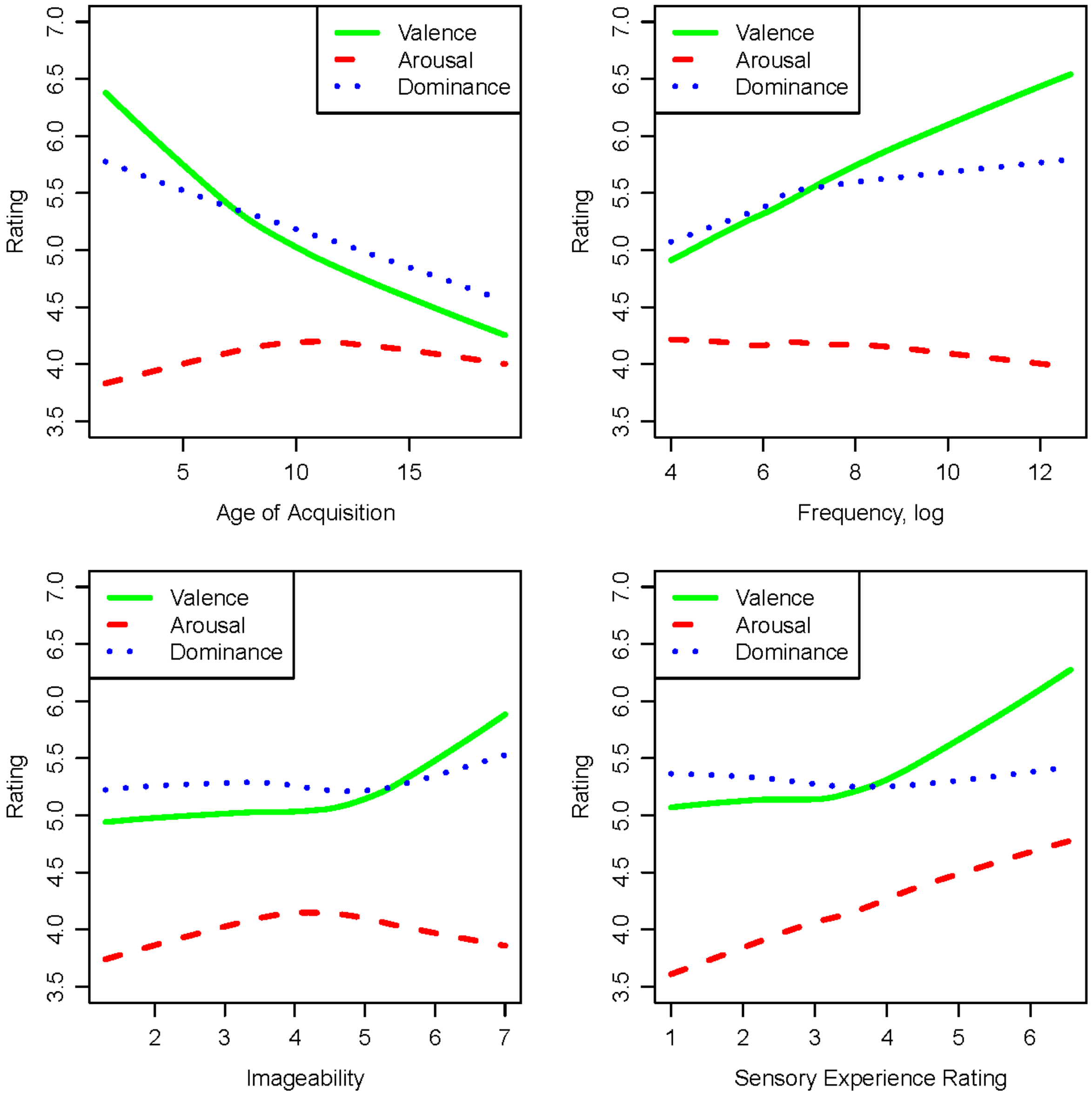
Sources: [a] Cortese & Fugett (2004); [a] Schock, Cortese, & Khanna (2012); [b] Altarriba, Bauer, & Benvenuto (1999); [c] Gilhooly & Logie (1980); [d] Stadthagen-Gonzalez & Davis (2006); [e] Kuperman et al. (2012); [f]extended dataset of Juhasz, Yap, Dicke, Taylor, & Gullick (2011) and Juhasz and Yap (in press) ; [g] Tillotson, Siakaluk, & Pexman (2008); [h] Amsel, Urbach, & Kutas (2012); [i] Medler, Arnoldussen, Binder, & Seidenberg (2005); [j] Brysbaert & New (2009)

Most correlations that emotional ratings show with other semantic properties are weak to moderate, with the exception of correlations with variables that directly tap into emotional states (h and i in Table 6). Specifically, words that make people happy are easier to picture (r = .161, df = 5123, p <.001), more concrete (r = .105, df = 1565, p <.001), familiar (r =.206, df = 2904, p < .001), context rich (r = .196, df = 316, p <.001), easy to interact with (r = .203, df = 1396, p < .001), are of high frequency (r = .182, df = 13763, p < .001) and learned at an early age (r = -.233, df = 13707, p < .001). They are also associated with low pain (r = -.456, df = 501, p < .001), intense smell (r = .139, df = 501, p < .01), vivid color (r = .322, df = 1281, p < .001), pleasant taste (r = .309, df = 501, p < .001), quiet sounds (r = -.176, df = 501, p < .001), and stillness (r = -.113, df = 501, p < .05). Virtually all these properties are also associated with words that make people feel in control, i.e. they correlate in the same way with dominance ratings.

Words that make people feel excited are more ambiguous (r = -.258, df = 1565, p < .001), unfamiliar (r = -.193, df = 501, p < .001), context impoverished (r = -.147, df = 316, p < .01), and difficult to interact with (r = -.143, df = 1396, p <.001). They are also associated with strong general sensory experience (r = .228, df = 5005, p < .001), specifically with high pain (r = .579, df = 501, p < .001), unpleasant taste (r = -.102, df = 501, p < .05, intense sounds (r = .407, df = 501, p < .001), motion (r = .335, df = 1281, p < .001), and an inability to be grasped (r = -.121, df = 501, p < .01).

As correlations do not reveal the form of the functional relationships, Figure 4 below zooms in on functional relationships between the three emotional dimensions and selected semantic properties of interest.

Figure 4: Relationship between the three dimensions and Age of Acquisition, Imageability, and Sensory Experience Ratings, presented as scatterplot smoother lowess trend lines.



The top left panel of Figure 4 reveals that early words are maximally positive, strong and calm. Words become more negative and weak (controlled-by) on average as the age-of-acquisition increases. The peak of arousal is reached in the words learned around age of 10, while later-acquired words are less exciting. It is tempting to interpret these results as an average developmental timeline of vocabulary acquisition in North American children, with (a) earliest happy and calm words learned in a risk-averse environment protecting a child from negativity and excitement and (b) excitable words like sexual terms, taboo words and swear words learned in the early school age. Yet it is more likely that the age-of-acquisition patterns of emotional words are at least partly due to how often they occur in English, and thus how likely children are to encounter and learn them early. Figure 4 top right demonstrates that the more frequent a word is, the happier, stronger and calmer it tends to be. The observed linear relationship between log frequency of occurrence and valence is reasonably strong: the Pearson's correlation coefficient is 0.18, and the increase in valence between the least and most frequent words is on the order of 2 points on the 9-point scale. This corroborates the finding of Garcia, Garas and Schweitzer (2012) and runs counter to the claim of Kloumann et al. (2012) that the positivity bias in English words is only observed in word types (there are more positive than negative words) and that correlations between frequency and valence, if any, are corpus-specific and small. The discrepancy may be due to the much broader range of frequency that we consider here, with fourteen thousand words from the top of the frequency list rather than five thousand words in each of the corpora considered in Kloumann et al. (2012). We leave the verification of the positivity bias over a broader frequency range to further research.

Only highly imageable words are emotionally colored (Figure 4, bottom left): as imageability increases from rating 5 on the 7-point scale, words become more positive and strong (in-control). Again, arousal stands out in these patterns: words that are hardly imageable at all or very imageable are calm, while those in the middle of the imageability range raise excitement.

The increasing strength of the sensory experience (Figure 4, bottom right) varies strongly with arousal: the more tangible the word is, the more exciting it is. This suggests that abstract notions are less powerful in agitating human readers than material objects. The functional relationship with valence is only observed in the top half of the sensory experience range: more tangible words induce increasingly positive emotions. No reliable relationship is observed between sensory experience ratings and dominance.

Interactions Between Demographics and Ratings

Participants were naturally divided into two genders. In addition, we divided them into two age ranges using the median split – younger (less than 30) and older (30 or greater). We also dichotomized education level into higher (those who had an Associate's degree or greater) and lower (some college or less). All three dimensions showed slightly but significantly higher average ratings for younger vs. older, and for lower education vs. higher education. Also, males gave slightly but reliably higher ratings in all dimensions than females. Separate independent t-tests showed that this difference was significant for Valence and Arousal but not for Dominance. The means, standard deviations, and independent t-test significance levels of each group division are listed in Table 7.

Table 7: Group differences in emotional dimensions. Reported are the number of raters (N), the number of observations (# of Obs) and the percent of total observations in each group (in brackets), the group mean and the average standard deviation and, in the last column, the p-value of a two-tailed independent t-test comparing group means.

		Male				Female			
	N	# of Obs	Mean	Avg SD	N	# of Obs	Mean	Avg SD	p

Valence	301	116,819 (38%)	5.13	1.60	419	184,636 (61%)	5.00	1.64	<.001
Arousal	291	119,658 (37%)	4.38	2.27	448	197,648 (62%)	4.10	2.28	<.001
Dominance	336	149,329 (44%)	4.83	2.15	505	188,433 (55%)	4.81	2.13	n.s.
		Old				Young			
	N	# of Obs	Mean	Avg SD	N	# of Obs	Mean	Avg SD	p
Valence	346	158,067 (52%)	5.04	1.61	382	147,892 (48%)	5.10	1.68	<.001
Arousal	373	174,402 (54%)	4.13	2.27	374	146,021 (46%)	4.31	2.31	<.001
Dominance	384	153,581 (45%)	4.80	2.04	464	187,137 (55%)	4.88	2.17	<.001
		High Edu				Low Edu			
	N	# of Obs	Mean	Avg SD	N	# of Obs	Mean	Avg SD	p
Valence	362	136,280 (45%)	5.10	1.57	361	167,259 (55%)	5.04	1.70	<.05
Arousal	363	142.151 (45%)	4.28	2.17	382	177,213 (55%)	4.14	2.33	<.001
Dominance	402	154,590 (46%)	5.17	2.02	443	184,733 (54%)	5.20	2.22	<.05

Note: Numbers of observations do not always equal 100% due to a small number of participants who declined to answer the relevant demographic questions.

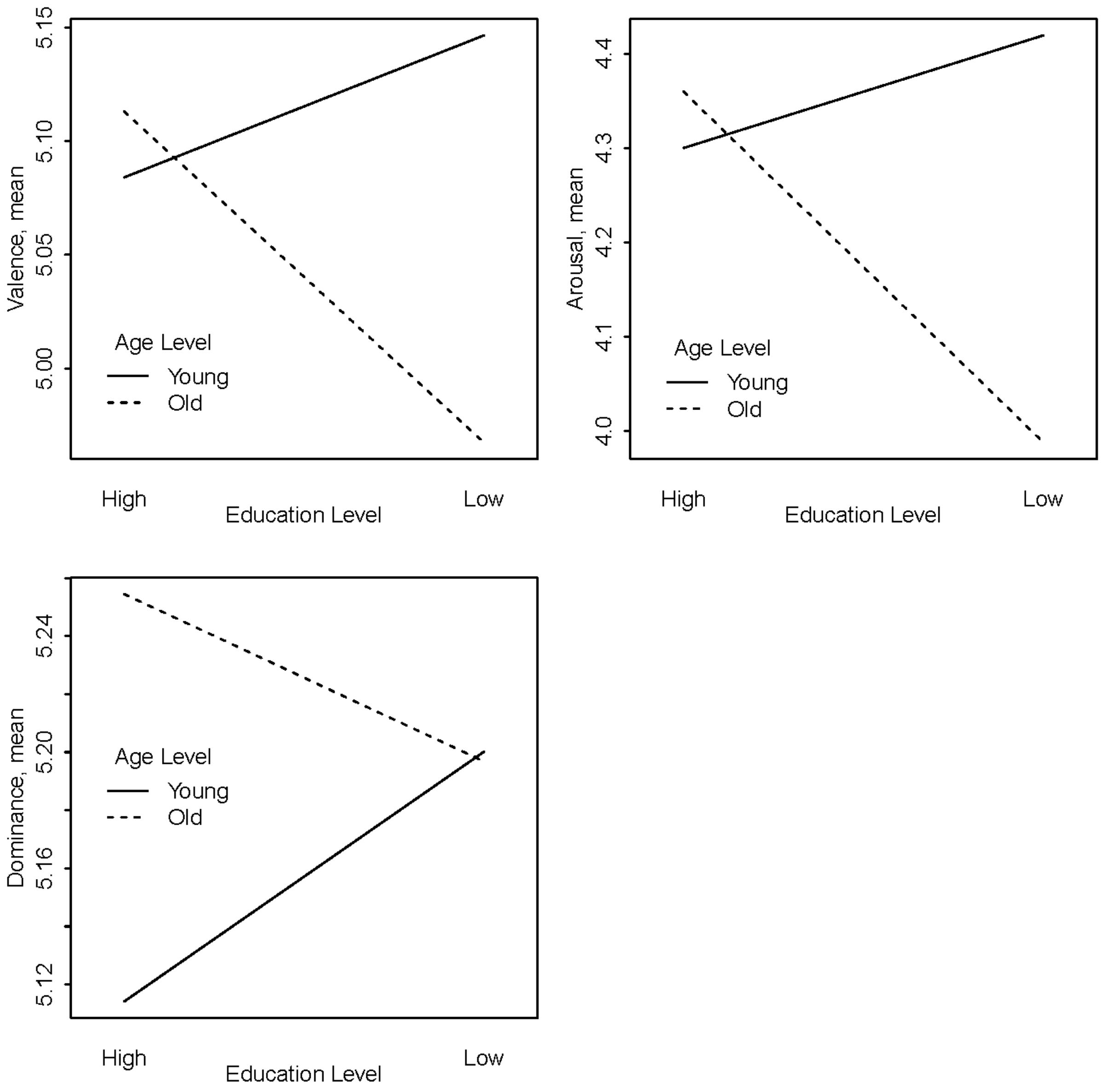
Table 8 reports correlations between groups of participants and demonstrates substantial variability in the ratings they provide: as with the overall data in Table 5, Arousal and Dominance elicit less agreement in judgments than Valence does.

Table 8: Correlations between Groups

	Valence	Arousal	Dominance
Male and Female	.789	.516	.593
Old and Young	.818	.500	.591
High Edu and Low Edu	.831	.467	.608

We ran a series of multiple regressions looking at age, gender, and education (all dichotomized as described above) as predictors. All main effects were significant at $p < .001$ and each variable made a unique contribution to the variance in the collected ratings. In addition, most of the two- and three-way interactions for all three dimensions were significant, likely due to the large number of data points available. However, the actual ranges of effects tended to be small. One exception was the interaction between age and education level for all three dimension (see Figure 5). For valence and arousal, highly educated people rated words similarly, regardless of age. For those with less education, age strongly affected ratings with the younger group providing higher ratings, on average, than the older. For dominance, the opposite pattern holds. Age affected those in the higher education group with older providing higher ratings than younger, but did not have an effect in the lower education group.

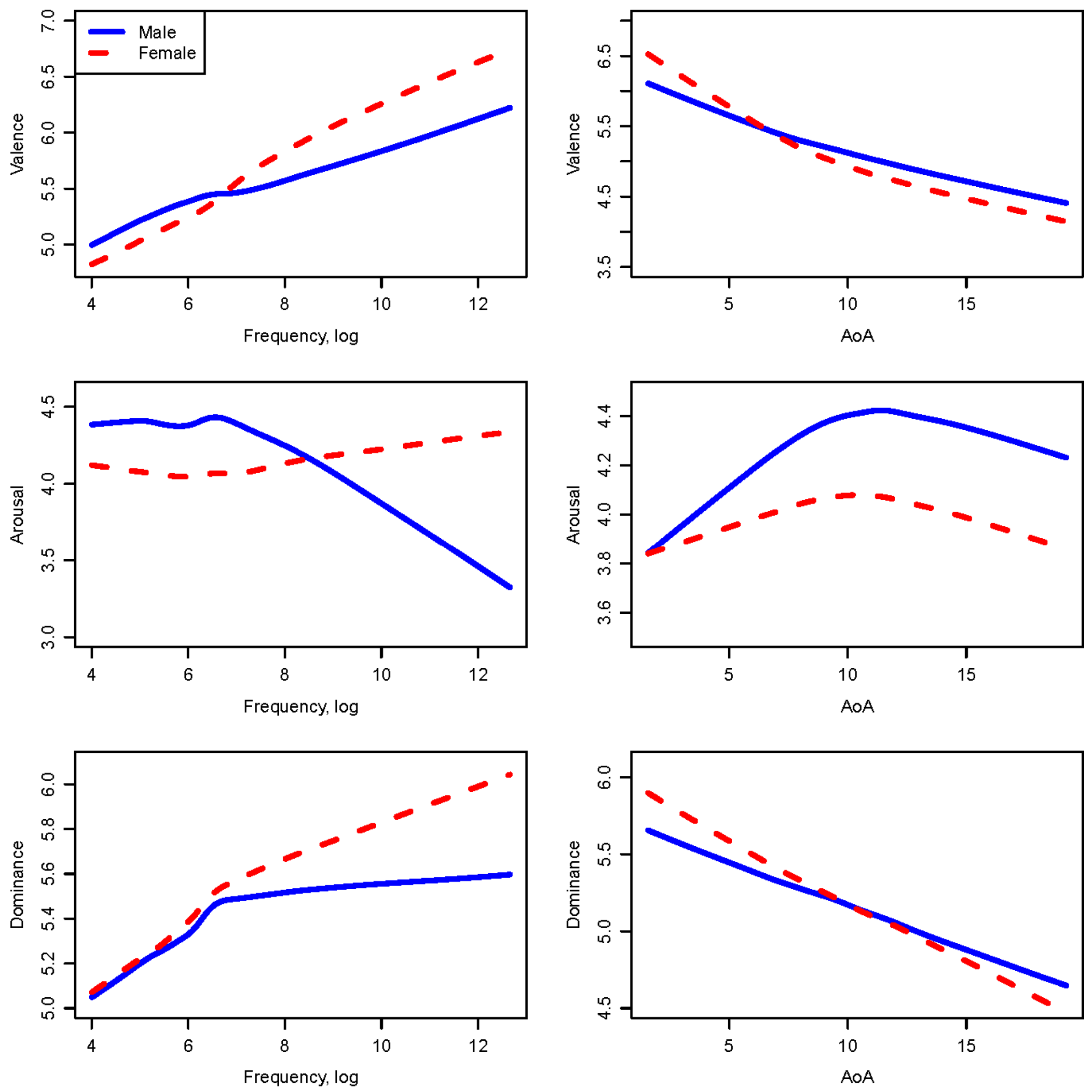
Figure 5: Interactions between dichotomized education and age levels for all three dimensions. All interactions are significant at p < .001.



Gender Differences

In what follows, we concentrate on gender differences. Effects of well-established lexical properties on emotion norms varied by gender. Figure 6 presents interactions of gender with frequency of occurrence and age of acquisition as predictors of emotional ratings. All interactions reached significance in multiple regression models with each set of ratings, separately, as a dependent variable: all ps < 0.01.

Figure 6: Interactions of gender with frequency (left) and age-of-acquisition (right) as predictors of mean ratings of valence (top), arousal (middle) and dominance (bottom). Interactions are presented with gender-specific lowess trend lines.

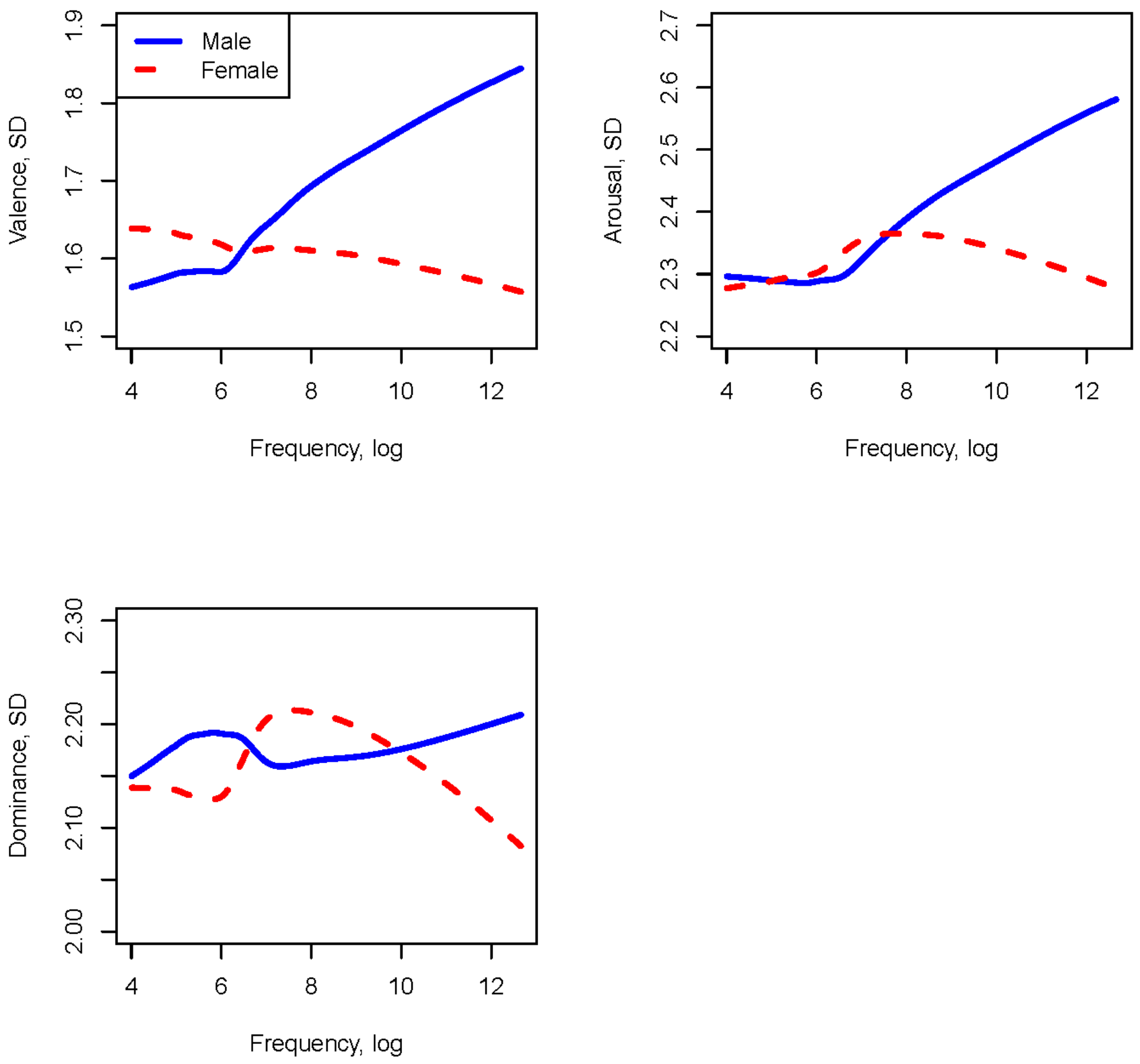


Interactions reveal that female raters provide more extreme negative/weak ratings for lowest-frequency and more extreme positive/strong ratings for higher-frequency words, yielding a broader range of values for both valence and dominance. The same holds for the more extreme ratings given by females to earliest- and latest-learned words, as compared to males.

Quite the opposite pattern was observed in the ratings of arousal (Figure 6, middle row). Female raters show a weak relationship between either frequency or AoA and arousal, with slightly higher arousal words in the higher-frequency band and in the mid-range of AoA. Conversely, male raters reveal a strong tendency to find higher-frequency and earlier-learned words less exciting than relatively late and infrequent words.

Variability in ratings also varied by gender, see Figure 7. Male raters disagree increasingly more on all ratings to higher frequency words, while variance in ratings by female participants was increasingly attenuated with the increase in word frequency.

Figure 7: Interactions of gender with frequency as a predictor of standard deviations of ratings of valence (top left), arousal (top right) and dominance (bottom left). Interactions are presented with gender-specific lowess trend lines.

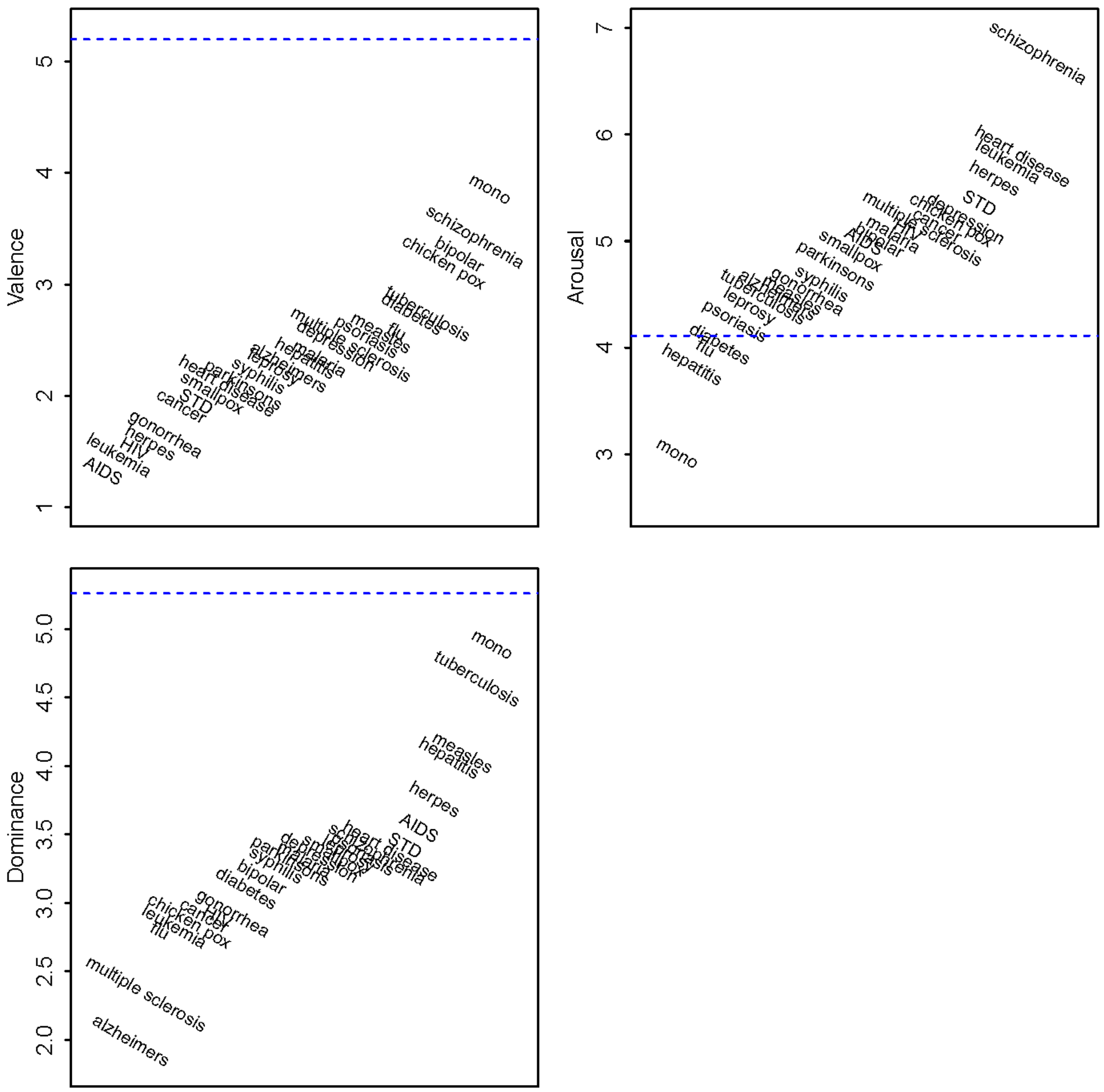


While pinning down the origin of these differences is an issue for further investigation, here we note the necessity of research into emotion words to take into account these interactions as potential sources of systematic error.

Semantic Categories

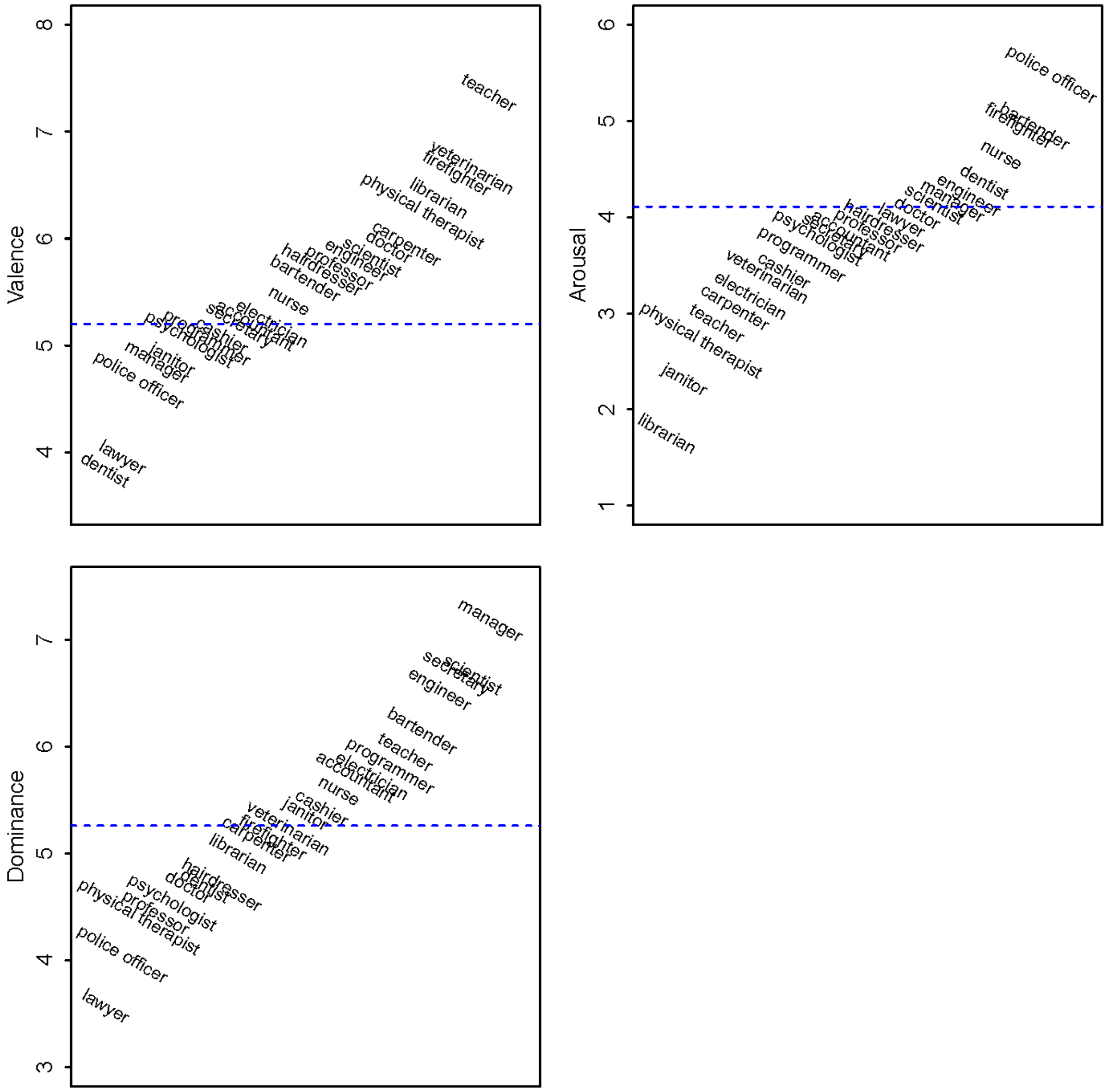
An interesting aspect of emotional ratings is their use to quantify attitudes and opinions toward physical, psychological and social phenomena either in the population at large or in specific target groups. We showcase here emotional ratings to the semantic categories of “disease” (Figure 8) and “occupation” (Figure 9), based on Van Overschelde et al.'s (2004) Category norms with occasional additions of semantically similar words. As Figure 8 suggests, all diseases are rated as words evoking negative feelings, high arousal, and feelings of being controlled, i.e. all ratings were below the median of valence/dominance and above the median of arousal in the entire dataset (shown as dotted line). Sexually transmitted diseases are judged among the most negative and the most anxiety-provoking entries in the subset. This is generally in line with surveys of attitudes that list sexually transmitted diseases among the most stigmatized medical conditions (e.g. Brems, Johnson, Warner & Roberts, 2010). The most feared medical conditions -- cancer, Alzheimer’s, heart disease, stroke (listed by the decreasing percentage of respondents who feared it; YouGov, 2011; MetLife Foundation, 2011) – are also among the most negative, the least controllable and the most anxiety-provoking diseases.

Figure 8: Ratings of words denoting disease. Dotted lines represent median ratings of respective emotional dimensions in the entire dataset.



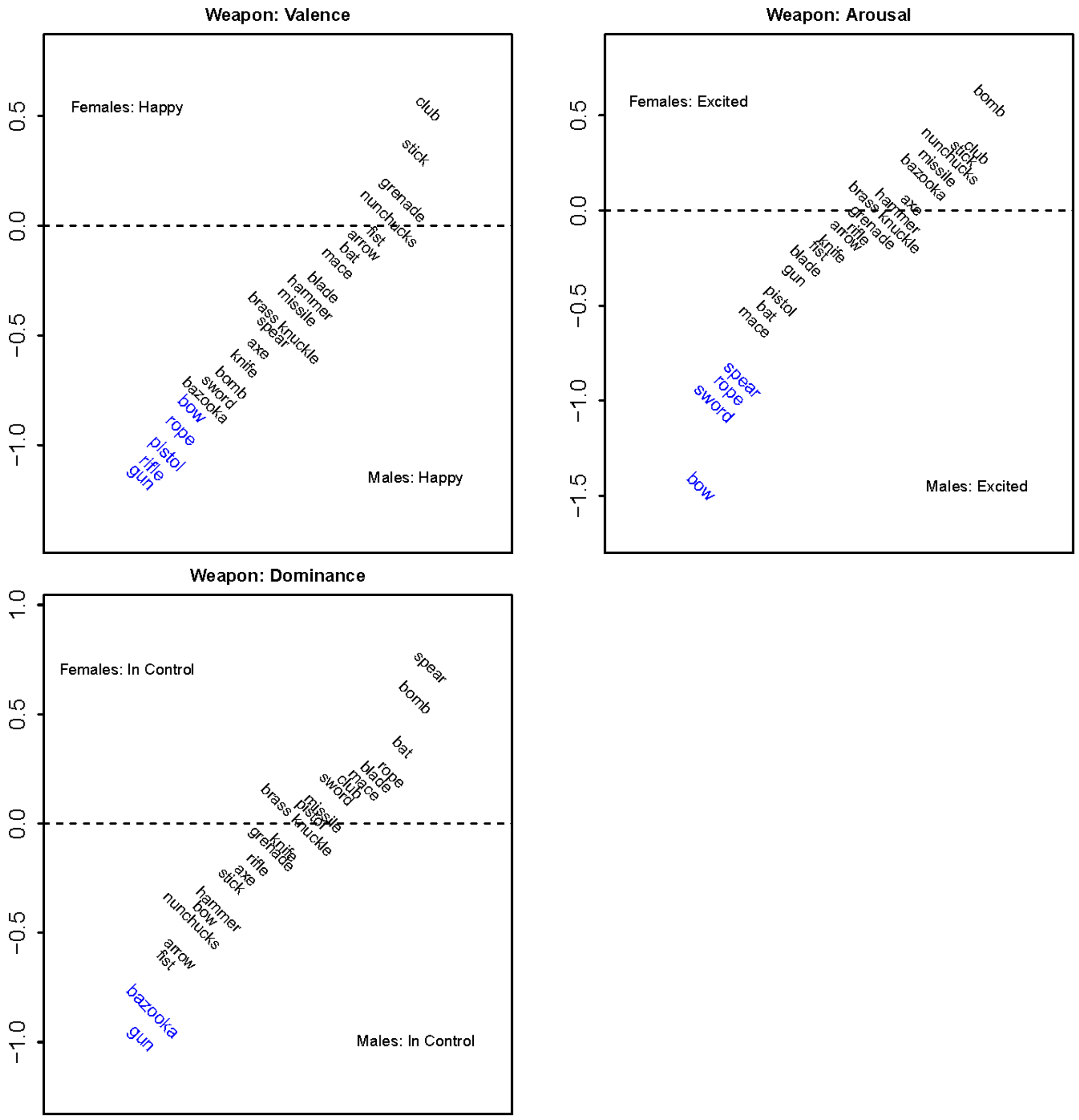
Ratings of valence to occupations reveal that the best-paying professions in the list are judged as the most negative, below the median in the overall dataset: cf. *lawyer*, *dentist*, and *manager*. The correlation between the average income as reported by the Bureau of Labor Statistics (2011) and mean valence is indeed negative, but does not reach significance (r = -.167, p = .434), possibly due to a reduced statistical power (*df* = 22). Some interesting contrasts can be seen that might prove interesting to social scientists. For example, both the words *police officer* and *firefighter* are rated as highly arousing, but police officer is viewed negatively while *firefighter* is viewed positively. In contrast, *librarian* is a positive but completely unarousing occupation term.

Figure 9: Ratings of words denoting occupations. Dotted lines represent median ratings of respective emotional dimensions in the entire dataset.



Emotional ratings are also a useful tool for studying gender differences in attitudes and beliefs. Figure 10 reports gender differences in ratings to terms denoting weaponry, with the difference between ratings of female and male responders on the y-axis. Upper parts of plots in Figure 10 show words that were given higher valence, arousal or dominance ratings by female responders; dotted lines represent the no-difference line. Words in blue color stand for items for which the difference in ratings between gender groups reached significance at the 0.01-level in the two-tailed independent t-test.

Figure 10: Gender differences in ratings for weapon related words.



All three emotional dimensions showed a significantly greater number of ratings in the lower parts of the plots (all p-values in chi-squared tests < 0.01). This indicates that male responders generally have a happier, more aroused and more in-control attitude towards weapons, especially fire weapons and the bow for which the gender difference in ratings reached significance.

A similar bias towards higher valence, arousal and dominance is observed in ratings of male responders to taboo words and sexual terms. As Figure 11 and 12 demonstrate, most lexical items in this subset are located below the dotted line, revealing overall higher ratings to taboo words in male responders (marked in blue if reaching significance) and in rare cases in female responders (marked in red if reaching significance). Observed discrepancies in attitudes are corroborated by Janschewitz, 2008, Newman, Groom, Handelman, and Pennebaker, 2008, and Petersen and Hyde, 2010. The discrepancies also explain the disproportionate presence of sexual terms and taboo words among lexical items with exceedingly variable ratings (see highlighted words in Figure 2 with the standard deviation larger than the value predicted from their mean).

Figure 11: Gender differences in ratings for taboo words.

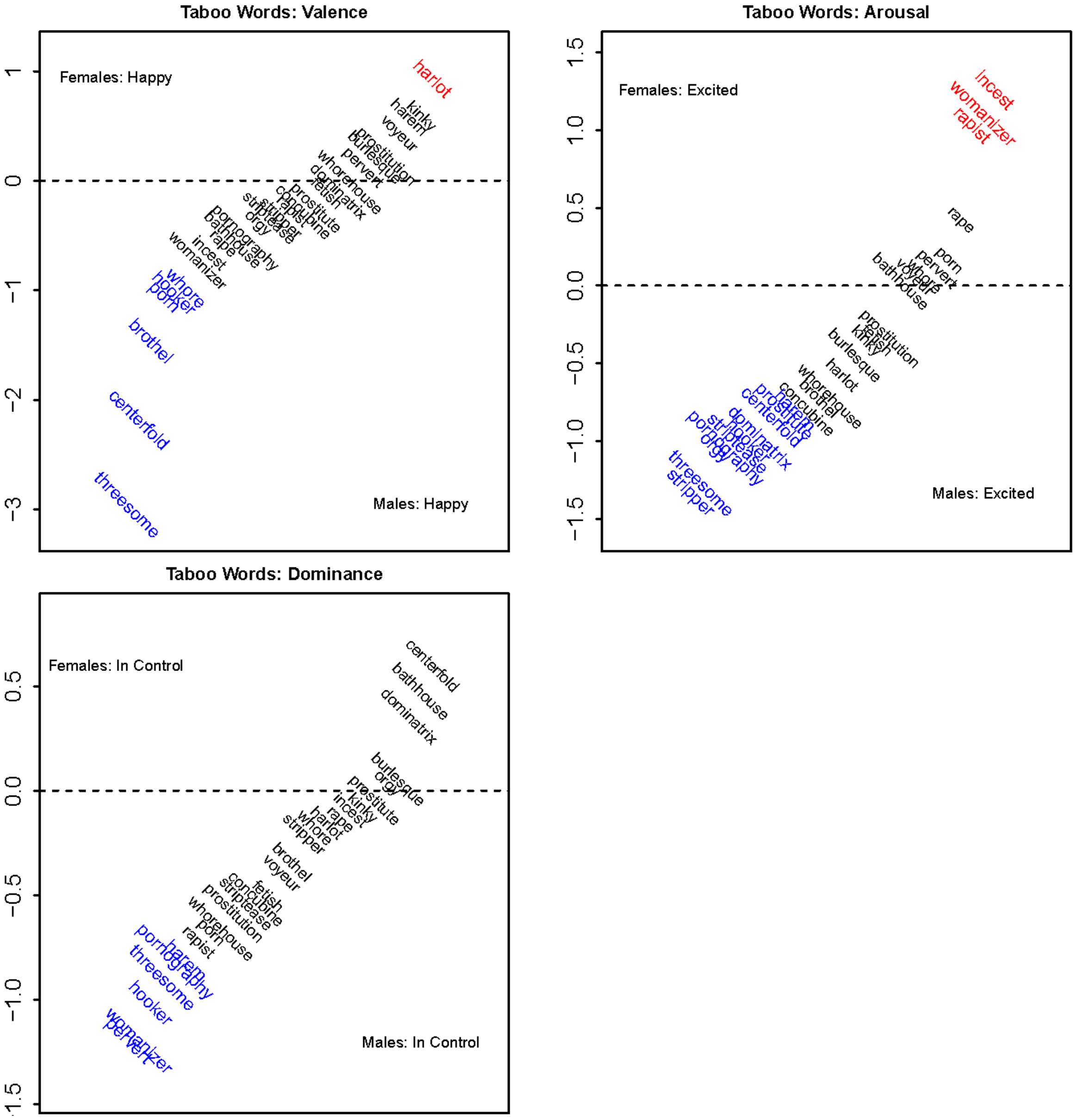
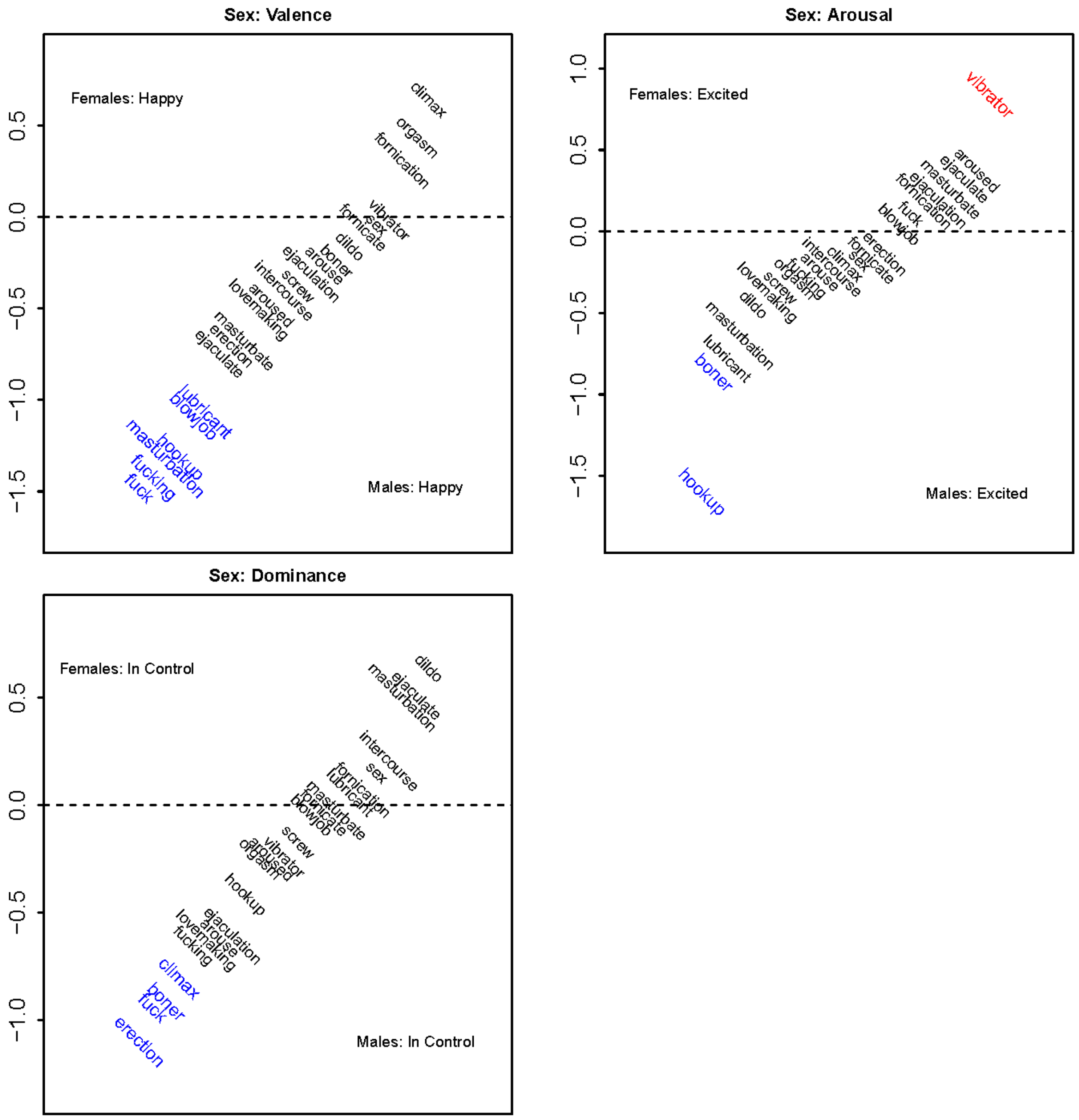


Figure 12: Gender differences in ratings for sex related words.



GENERAL DISCUSSION

Technological advances are rapidly changing the tools language researchers have at their disposal. Two main, complementary developments are (1) the collection of large sets of human data through crowdsourcing platforms, and (2) the automatic calculation of word characteristics on the basis of relationships between words. In the former case, current means of digital communication are used to reach a large audience at an affordable price. The current study is a typical example of this: Instead of having to limit the list of words to a few hundred because of a lack of human respondents, we extended the list to nearly 14 thousand (see Kuperman et al., 2012, for another example of a large sample rating obtained via crowdsourcing). Our collection of primary demographic information, such as age, gender and education, additionally enables refined analyses of both the central tendency and variability in each of the emotional dimensions. Likewise, it paves the way for characterization of attitudes and opinions in the population at large, as well as specific groups of respondents.

The derivation of word features by means of counting word co-occurrences is an approach that is likely to expand considerably in the coming years. Arguably the showcase at the moment is the derivation of word meanings by establishing which words co-occur in texts and bits of discourse. Estimates based on word co-occurrences correlate reasonably well with human-generated word associations and semantic similarity ratings. The approach was initiated by Landauer and Dumais (1997) and Burgess (1998). Recent reviews and extensions can be found in Shaoul and Westbury (2010) and Zhao, Li, and Kohonen (2011). The enterprise critically depends on algorithms that automatically extract word information from collections of texts and calculate various measures of co-occurrence.

Bestgen and Vincze (2012) applied this approach to the affective dimensions of words. They calculated affective norms for over 17 thousand words by comparing each word to the thousand words from the ANEW list. The score of each word was derived from the ANEW norms of the words with the closest distance in the semantic space. Bestgen and Vince (2012) observed that performance was best when the 30 closest neighbors of the target word were used. This led to correlations of r= .71 between the automatically derived values of valence and the human ratings, r = .56 for arousal, and r = .60 for dominance. All things equal, these correlations depend on the number of so-called “seed words”, words with known values to which the new words can be compared. The more seed words, the better the estimates for the remaining words. On the other hand, the more seed words for which there are human data, the less need for the automatic extraction of such information. Our extensive dataset clearly contributes to the accuracy of such computational estimates. Additionally, it introduces the opportunity to make estimates of textual sentiment for specific reader profiles: low-educated men, older women, or highly educated youngsters. This in turn may inform the creation of texts that are made more or less emotionally appealing or arousing to specific target populations.

To sum up, our collection of emotion norms for nearly 14 thousand words gives computational and experimental researchers of language use a much wider selection for their studies. Depending on the size of a person’s vocabulary, this is estimated to be between one half and one quarter of the words known to individuals. Reliable ratings of affective states invoked by this number of words will advance the study of the interplay between language and emotion.

AVAILABILITY

Our ratings are available as supplementary materials to this article and provided in .csv format. Every value is reported three times, one for each dimension, prefixed with V for valence, A for arousal, and D for dominance. For each word, we report the overall mean (Mean.Sum), standard deviation (SD.Sum), and number of contributing ratings (Rat.Sum). We also report these values for group differences, replacing the suffix .Sum with the following (.M = male; .F =

female; .O = older; .Y = younger; .H = high education; .L = low education). Words are presented in alphabetical order.

We note that group differences (gender, education level, and age, while interesting, are actually quite limited. Taking a conservative p < .01 as our definition of significantly different, there are less than 100 words per dimension that meet this criteria (education and arousal include more with nearly 200 words each). In terms of gender, the differences seem to occur primarily in categories related to sex, violence, and other taboo topics. When these stereotypical domains are under investigation, we do advise people to consider gender differences in ratings. The semantic categories for other group differences were more difficult to define. In general, unless there is an already established reason to consider group differences, using the overall .Sum ratings is, we feel, completely valid.

Acknowledgements

This study was supported by the Odysseus grant awarded by the Government of Flanders (the Dutch-speaking Northern half of Belgium) to MB, and by the Insight Development grant of the Social Sciences and Humanities Research Council and the Discovery grant of the Natural Sciences and Engineering Research Council of Canada to VK. We thank Dan Wright and Barbara Juhasz for insightful comments to the previous draft.

References

Altarriba, J., Bauer, L. M., & Benvenuto, C. (1999). Concreteness , context availability , and image ability ratings and word associations for abstract , concrete , and emotion words. *Behavior Research Methods, 31*(4), 578–602.

Amsel, B. D., Urbach, T. P., & Kutas, M. (2012). Perceptual and motor attribute ratings for 559 object concepts. *Behavior Research Methods*. Advance online publication. doi:10.3758/s13428-012-0215-z

Augustine, A.A., Mehl, M.R., & Larsen, R.J. (2011). A positivity bias in written and spoken English and its moderation by personality and gender. *Social Psychological and Personality Science, 2*(5), 508-515.

Baayen, R. H., Feldman, L. F. and Schreuder, R. (2006) Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language,* , 496-512.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., and Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*, 445–459.

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

Bradley, M.,M. & Lang, P. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

Brems, C., Johnson, M.E., Warner, T.D., Roberts, L.W. (2010). Health care providers' reports of perceived stigma associated with HIV and AIDS in rural and urban communities. *Journal of HIV/AIDS & Social Services, 9*(4), 356-370.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–90.

Bureau of Labor Statistics (May 2011). *National occupational employment and wage estimates: United States.* Retrieved August 31, 2012 from http://www.bls.gov/oes/current/oes_nat.htm#00-0000

Burgess, C. (1998). From simple associations to the building block of language: Modeling meaning in memory with the HAL model. *Behavior Research Methods, Instruments, and Computers, 30*, 188–198.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112 (1)*, 155–159.

Cortese, M. J., & Fugett, A. (2004). Imageability ratings for 3,000 monosyllabic words. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 384–7.

Eilola, T. M., & Havelka, J. (2010). Affective norms for 210 British English and Finnish nouns. *Behavior Research Methods*, *42*(1), 134–40.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., & Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods, 42*(2), 488-496.

Fraga, I., Pineiro, A., Acuna-Farina, C., Redondo, J., & Garcia-Orza, J. (2012). Emotional nouns affect attachment decisions in sentence completion tasks. *Quarterly Journal of Experimental Psychology, 65*, 1740-1759.

Garcia, D., Garas, A., Schweitzer, F. (2012). Positive words carry less information than negative words. *EPJ Data Science, 1*(3). Open Access. doi: 10.1140/epjds3.

Gilhooly, K. J., & Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods*, *12*(4), 395–427.

Janschewitz, K. (2008). Taboo, emotionally valenced, and emotionally neutral word norms. *Behavior Research Methods, 40*(4), 1065-1074.

Juhasz, B. J., Yap, M. J., Dicke, J., Taylor, S. C., Gullick, M. M. (2011). Tangible words are recognized faster: the grounding of meaning in sensory and perceptual systems. *Quarterly Journal of Experimental Psychology, 64*, 1683–1691.

Juhasz, B. J., & Yap, M. J. (in press). Sensory experience ratings for over 5,000 mono-and disyllabic words. Behavior Research Methods.

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods, 42*(3), 643-650.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods, 44*, 287-304.

Kloumann, I. M., Danforth, C. M., Harris, K. D., Bliss, C. A., & Dodds, P. S. (2012). Positivity of the English language. *PloS one*, *7*(1), e29484. doi:10.1371/journal.pone.0029484

Kousta, S.T., 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.

Kousta, S.T., Vinson, D.P., & Vigliocco, G. (2009). Emotion words, regardless of polarity, have a processing advantage over neutral words. *Cognition, 112*, 473-481.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30 thousand English words. *Behavior Research Methods*, *44*, 978-990.

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.

Leveau, N., Jhean-Larose, S., Denhière, G. & Nguyen, B. (2012). Validating an interlingual metanorm for emotional analysis of texts. *Behavior Research Methods, 44,* 1007-1014.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Medler, D., Arnoldussen, A., Binder, J., & Seidenberg, M. (2005). *The Wisconsin perceptual attribute ratings database.* Retrieved from http://www.neuro.mcw.edu/ratings/

MetLife Foundation (2011). *What America thinks: MetLife Foundation Alzheimer's survey*. Retrieved August 31, 2012 from http://www.metlife.com/assets/cao/contributions/foundation/alzheimers-2011.pdf

Mohammad, S.M., & Turney, P.D. (2010). Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, LA, California.

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., van Schie, K., Van Harmelen, A., De Schryver, M., et al. (in press). Norms of valence, arousal, dominance, and age of acquisition for 4300 Dutch words. *Behavior Research Methods*. Retrieved from https://lirias.kuleuven.be/handle/123456789/351830

Newman, M.L., Groom, C.J., Hamdelman, L.D. & Pennebaker, J.W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45, 211-236.

Osgood, C.E., Suci, G.J., & Tannenbaum, P. (1957). *The Measurement of Meaning*. University of Illinois Press.

Petersen, J.L. & Hyde, J.S. (2010). A meta-analytic review of research on gender differences in sexuality, 1993 – 2007. *Psychological Bulletin, 136*(1), 21-38.

Rammstedt, B. & Krebs, D. (2007). Does response scale format affect the answering of personality scales? European *Journal of Psychological Assessment, 23*(1), 32-38.

Redondo, J., Fraga, I., Padrón, I., & Comesaña, M. (2007). The Spanish adaptation of ANEW (affective norms for English words). *Behavior Research Methods*, *39*(3), 600–5.

Schock, J., Cortese, M. J., & Khanna, M. M. (2012). Imageability estimates for 3,000 disyllabic words. *Behavior Research Methods*, *44*(2), 374–9.

Scott, G.G., O'Donnell, P.J., & Sereno, S.C. (2012) Emotion words affect eye fixations during reading. *Journal of Experimental Psychology: Learning, Memory and Cognition, 38*, 783-792.

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

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), 256–69.

Stadthagen-Gonzalez, H., & Davis, C. J. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, *38*(4), 598–605.

Tillotson, S. M., Siakaluk, P. D., & Pexman, P. M. (2008). Body-object interaction ratings for 1,618 monosyllabic nouns. *Behavior Research Methods*, *40*(4), 1075–8.

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, *50*(3), 289–335.

Verona, E., Sprague, J., & Sadeh, N. (2012). Inhibitory control and negative emotional processing in psychopathy and antisocial personality disorder. *Journal of Abnormal Psychology, 121*, 498-510.

YouGov (2011). *Cancer Britons most feared disease*. Retrieved August 31, 2012 from http://yougov.co.uk/news/2011/08/15/cancer-britons-most-feared-disease/

Zhao, X., Li, P. & Kohonen, T. (2011). Contextual self-organizing map: Software for constructing semantic representation. *Behavior Research Methods*, 43, 77-88.