THE INTERPLAY OF LANGUAGE AND EMOTION

THE INTERPLAY OF LANGUAGE AND EMOTION: USING AFFECTIVE NORMS TO EXPLORE WORD RECOGNITION, MOTIVATION, AND LEXICON

By AMY BETH WARRINER, B.A. (hons.)

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

McMaster University

DOCTOR OF PHILOSOPHY (2014) McMaster University (Psychology)

TITLE: The interplay of language and emotion: Using affective norms to explore word recognition, motivation, and lexicon

AUTHOR: Amy Beth Warriner, B.A. (hons.) (McMaster University)

SUPERVISOR: Dr. Victor Kuperman

NUMBER OF PAGES: xvii, 254

Abstract

A lack of norms limited previous work on the interplay of language and emotion. Valence and arousal are regularly dichotomized affecting generalizability and accuracy. Important questions remain unexplored such as the interaction between these dimensions along with individual and group differences. Chapters 2 and 3 report collections of affective and concreteness norms. In Chapter 4, these norms are used to reveal that valence is negatively and arousal is positively correlated with reaction time, both monotonically. Previously, it has been argued that people categorically distinguish between positive and negative or prioritize emotional over neutral stimuli. We demonstrate that this automatic vigilance must be graded. Chapter 5 introduces a method for measuring approach and avoidance in proportion to valence and arousal. A previously demonstrated congruency effect between valence and approach and avoidance movements is categorical. We showed that people choose distances proportionally to word valence and that responses are affected by word frequency, gender, and personality. Finally, Chapter 6 combines the distribution of affect with word frequency information to reveal how language is organized around communicative needs. A compound bias toward high-arousal emotional and low-arousal, mid-valence word types along with more frequent use of positive words suggest that humans need tools to talk about danger and thrills as well as the mundane, while fostering relationships by focusing on the positive. Thus, this dissertation provides important resources – large sets of norms – for the extension of studies on emotion and language. It shows the value of these norms in revisiting past studies of word processing, enabling new methods for testing the motivations behind emotional effects, and considering how the distribution of emotion across language informs our understanding of these motivations. Throughout each chapter, group and individual differences are explored.

Acknowledgements

First and foremost, I want to thank my supervisor, Victor Kuperman. He took a chance on me partway through my degree and turned the daunting task of starting over on a new project into an amazingly fruitful learning adventure. No matter how discouraged I got, he remained unfailingly optimistic. His continued belief in me and in our work kept me going. His enthusiasm for research and his excitement over new data inspired me. I have learned so much over the past two years and will forever be grateful.

In addition to Victor, I want to thank Marc Brysbaert who is a co-author on three of the included papers. His involvement and financial assistance made these studies possible. While I only met him once, I am grateful he was a part of our team.

I am also thankful for David Shore and Louis Schmidt. They agreed to be a part of my committee partway through my journey even though my research area was outside their normal scope. Their participation, despite their extremely busy schedules, has made my timely completion of this dissertation possible. Thanks go also to my previous committee members, Scott Watter, Jen Ostovich and Michael Kliffer.

Karin Humphreys, my former supervisor, was the one who first saw my potential as an undergraduate and opened up opportunities for me to explore research. Her encouragement set me on this path and her belief in me brought me back to McMaster. I am grateful for her support, both personal and professional. The collegial atmosphere of the Cognitive Science Lab gave me my first taste of what working with a friendly and motivated team could be like. My thanks goes to each and every member who assisted with my research, brainstormed with me during lab meetings, or simply shared a stress-relieving laugh.

So many fellow students have contributed to my success, by helping run studies, giving feedback on presentations, or providing encouragement just when I needed it. In particular, I want to thank Emma Bridgwater for having coordinated data collection over the past year along with Daniel Schmidtke, Kaitlin Falkauskas, and Noor Al-Zanoon for including my stimuli in their eye-tracking studies and even helping with some of the data preparation.

Life is not lived, nor is research completed, in isolation. From strangers who found my short research descriptions fascinating thereby re-igniting passion, to friends who reminded me to enjoy the present, to family who pushed me when I felt like giving up – each shares in the success represented by this dissertation. I am particularly grateful for Catherine Anderson whose office door was never closed when I needed to chat, for Janice Anderson who patiently listened to my ever-changing career plans, and for Kelly Wilk Ricard who encouraged me to take care of myself.

And last but, certainly not least, I want to express my most sincere gratitude for my family. My mom, dad, brother and sister-in-law have been unfailing supporters, always believing I could do anything I put my mind to and being available anytime I needed encouragement. I couldn’t have done this without them.

I also want to acknowledge that my success would not have been possible without financial support from SSHRC in the form of a Joseph-Armand Bombardier Canada Graduate Scholarship and from the Margo Wilson and Martin Daly Ontario Graduate Scholarship.

Table of Contents

Declaration of Academic Achievement

This is a ‘sandwich’ thesis. Three of the empirical chapters have been published in peer-reviewed journals. One has been accepted pending revisions. And one is in preparation. The following outlines the status of each chapter and my contribution to each manuscript.

CHAPTER 2: This is a reprint of **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. The concept for this study came from VK and MB. My role included data collection from participants on the crowdsourcing platform Amazon Mechanical Turk and data analysis under the supervision of VK. I was also the primary writer. VK helped with editing, and contributed to the section on gender and semantic differences and the general discussion.

CHAPTER 3: This is a reprint of Brysbaert, M., **Warriner, A.B.**, & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*(3), 904-911. The concept for this study came from MB. I collected the data from participants on Amazon Mechanical Turk and did the data analysis. MB wrote the manuscript and I helped edit.

CHAPTER 4: This is a reprint of Kuperman, V., Estes, Z., Brysbaert, M., & **Warriner, A.B.** (2014). Emotion and language: Arousal and valence affect word recognition. *Journal of Experimental Psychology: General, 143*(3), 1065-1081. I participated in the conceptualization of this study and in discussions about the form and scope of the manuscript. Data analysis and writing was done primarily by VK and ZE. I helped edit.

CHAPTER 5: This is a manuscript in preparation. Co-authors include Victor Kuperman, David Shore and Louis Schmidt. I came up with the concept for this study which was refined through discussions with VK, DS, and LS. I collected the data with assistance from undergraduate students in the Reading Lab, did the data analysis and wrote the introduction, methods and discussion. VK assisted with writing the results and editing. DS and LS also helped edit the manuscript.

CHAPTER 6: This paper has been accepted pending minor revisions as Warriner, A.B. and Kuperman, V. Emotion is like a boomerang: Revisiting affective biases in the English language. *Cognition and Emotion.* The concept arose from discussions between VK and me I performed the initial data analysis and wrote the first draft of the paper. The manuscript has been altered significantly through revisions to which both VK and I contributed.

List of Figures

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

List of Tables

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 1: Introduction

The purpose of this dissertation is to set the groundwork for inquiries into the interplay of language and emotion, specifically how emotion impacts word processing and how language (primarily, the lexicon) reflects the structure of emotion. I will show how studies in this area have been hampered by small sets of lexical affective norms and introduce data from large norming studies designed to fill this gap. I will demonstrate how the availability of such a large set of norms changed our understanding of how affective properties of words influence such established measures of word recognition effort as lexical decision and naming times. I will then explore how patterns within this large, and thereby more representative set of norms, combined with information about word frequency in English, can reveal the underlying communicative needs that motivate language use in general. Finally, I propose and test a method for measuring how the emotionality of words affects the degree to which they evoke an approach or avoidance motivation. In the discussion I consider how this foundation can be built upon.

I begin by exploring what emotion is and how it might be expected to impact word processing. I review the ways in which this impact has been studied to date and the limitations of this past research. I then introduce the idea that not only can a consideration of the affective properties of words inform studies of word processing but that the structure of the lexicon can increase our understanding of human emotional experience. I conclude this introduction with an explanation of how this dissertation addresses those limitations and fills in some critical gaps in the literature.

What is emotion?

While emotion is a commonly understood word, its exact definition is notoriously difficult to pin down, and has been debated since the time of Plato and Aristotle. Definitions have changed over time and across domains (for a history, see Kagan, 2007), however, all include a subjective, physiological, and behavioral response to the appraisal of a personally significant event (Mauss & Robinson, 2009; Mulligan & Scherer, 2012) and contrast emotion with moods, preferences, dispositions, and attitudes which are all more durative and less event-specific (Russell 2003; Scherer, 2005).

Emotions have been modeled in several different ways. Historically, emotions were viewed as discrete and researchers searched for their basic or natural kinds. The belief was that each emotion would have a different pattern of subjective, physiological, and behavioural responses that could distinguish it from every other emotion (Dolan, 2002; Lench, Flores & Bench, 2011). Each emotion was viewed as an evolved set of instructions for the coordination of cognitive and physiological systems in response to events that were repeatedly encountered in our ancestral history (Cosmides & Tooby, 2000; Plutchik, 1980). Lists differed from researcher to researcher, but the five commonly agreed upon basic emotions included happiness, sadness, fear, anger, and disgust (Ekman, 1992). Physiologically, heart rate, skin conductance, and finger temperature were found to distinguish between anger, fear and disgust (Cacioppo, Berntson, Larsen, Poehlmann, & Ito, 2000). With regards to brain regions, two recent meta-analyses both found support for a connection between fear and the amygdala and between disgust and the basal ganglia but disagreed on the localization of anger, happiness, and sadness (Murphy, Nimmo-Smith, & Lawrence, 2003; Phan, Wager, Taylor, & Liberzon, 2002). Behaviorally, the strongest evidence for the existence of discrete emotions has come from facial perception studies in which people consistently identify others emotional expressions, even cross-culturally (Elfenbein & Ambady, 2002) and show high correlations between self-reported experience and their own facial expressions (Fridlund, Ekman, & Oster, 1987; Mauss, Levenson, McCarter, Wilhelm, & Gross, 2005; Ruch, 1995).

Many researchers, however, found that the coherence between these response domains for a given emotion category wasn't consistently high enough (Bradley & Lang, 2000; Russell, 2003; Shweder, 1993). Conscious self-report became a critical proving ground as subjective experience was thought to have the clearest connection to whatever causal mechanism was in play (Barrett, 2006a). However, factor analysis of self-reports failed to support the idea of discrete emotions, instead showing high correlations between positive and high correlations between negative states (e.g., Boyle, 1986; Watson & Clark, 1994; Zuckerman & Lubin, 1985). This led researchers to propose that emotion may best be measured via its underlying dimensions, primarily positivity or pleasure referred to as valence (e.g. Waston, Clark, & Tellegen, 1988) and activation or intensity referred to as arousal (e.g. Barrett & Russell, 1998; Mayer & Gaschke, 1988; Russell, Weiss, & Mendelsohn, 1989).

Returning to the question of physiology, valence has been found to be related to measures such as heart rate (Lang, Bradley, & Cuthbert, 1997; Cacciopo et al., 2000) and startle reflex magnitude (Bradley, Cuthbert, & Lang, 1996) while arousal has been found to be more strongly related to electrodermal responses such as skin conductance (Bradley & Lang, 2000; Lang, Greenwald, Bradley, & Hamm, 1993). Studies using word stimuli have found that arousal impacts early stages of processing and strongly activates more automated areas of the brain such as the amygdala (Hofmann, Kuchinke, Tamm, Võ, & Jacobs, 2009; Kissler, Herbert, Peyk, & Junghofer, 2007). In contrast, high valence words show an increased ERP late positive complex

(Bayer, Sommer & Schacht, 2010) and more strongly activate higher-order areas such as the orbitofrontal cortex (Lewis, Critchley, Rotshtein, & Dolan, 2007): for a thorough review, see Citron (2012). When the stimuli involved pictures, valence and arousal were found to interact in both early and late stages (Feng, Courtney, Mather, Dawson, & Davison, 2011; Lane, Chua, & Dolan, 1999). The actual region activated may depend more on the type of stimuli being processed than its specific emotional content as affect and cognition appear to be highly integrated (Ghashghaei & Barbas, 2002). For example, the arousal of words showed more activity in the left amygdala while arousal of pictures showed more activity bilaterally (Kensinger & Schacter, 2006).

Behaviorally, valence, in particular, has been associated with approach and avoidance motivation. Researchers adopting an evolutionary perspective have argued that emotions are primarily an adaptive response, a preparation for action that enhances survival (Damasio, 1998; LeDoux, 1996). They suggest that emotions can be divided into two primary reflexes – approaching positive stimuli and avoiding negative stimuli (Bradley & Lang, 2000; Bradley, Codispoti, Cuthbert, & Lang, 2001). Several studies have shown that people are indeed faster to make congruent movements such as pulling a lever in response to positive pictures or words than incongruent movements such as pulling a lever in response to negative pictures or words (Chen & Bargh, 1999). The actual action matters less then the end result, the distance between the subject and affective stimuli either being increased or decreased (van Dantzig, Pecher, & Zwaan, 2008; Krieglmeyer, De Houwer, & Deutsch, 2011). Relatedly, automatic vigilance is the theory that avoiding negative stimuli is of particular importance to survival (Pratto & John, 1991) and thus such stimuli capture and engages attention to a greater degree than positive or neutral stimuli (e.g. Fox, Russo, Bowles, & Dutton, 2001; Horstmann, Scharlau, & Ansorge, 2006;

McKenna & Sharma, 2004). This explains why people sometimes respond more slowly to negative than to positive stimuli in a variety of tasks (e.g. Algom, Chajut, & Lev, 2004; Pratto & John, 1991; Wentura, Rothermund, & Bak, 2000).

There are of course, other ways of modeling emotion and disagreements within the field. Some suggest the additional property of dominance (Russell & Mehrabian, 1977) based on Osgood, Succi, and Tennenbaums's (1957) factor analysis of conceptual meaning and/or an interpersonal property (Mesquita & Markus, 2004). One debate focuses on which factors best explain emotional experience. Some propose that valence and arousal are not independent and that the two dimensions should be what they call negative activation and positive activation (Watson, Wiese, Vaidya, & Tellegen, 1999). Similarly, in an attempt to derive an evolutionarily driven explanation for emotion patterns, Wurm and Vakoch (2000) combined low valence/high arousal/high dominance into a dimension called danger and high valence/high arousal or high valence/high dominance into a dimension called usefulness. Another debate concerns whether pleasantness and unpleasantness should be measured separately via unipolar scales (Lewis, et al.; Russell & Carroll, 1999).

Despite these disagreements, I would argue that the dimensional model including valence and arousal is currently the most productive engendering the most research. Numerous ratings studies with high correlations have shown that people can rate stimuli on these dimensions with fairly strong agreement (see Chapter 2). As such, this is the model adopted by this dissertation.

Where might emotion fit into a model of word processing?

Significant research has gone into determining how visual word recognition takes place (see Adelman, 2012; Balota, Yap, & Cortese, 2006). The process is thought to start with the identification of features or lines and shapes (Quinlan, 2003) and continue to the point where the word is recognized and distinguished from alternatives (Seidenberg & McClelland, 1989). Numerous models have been proposed to explain how this takes place which generally fall into two classes. Dual route models propose that there are two routes to word recognition – one in which familiarity with a given orthographic string allows it to be mapped directly onto a lexical representation and another that allows less familiar or non-words to be assembled via an analysis of the individual letter to sound mappings (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001). Parallel distributed processing models propose orthography input nodes and pronunciation output nodes which are connected by hidden units whose weights adjust in response to learning and back propagation (Plaut, McClelland, Seidenberg & Patterson, 1996; Seidenberg & McClelland, 1989). These models perform well in some areas and less well in others. Ultimately, the goal is to account for all the possible factors that influence word recognition, thereby deriving a model that can account for all the variance in reaction times, particularly in the lexical decision and naming tasks which are typically used as basic indicators that recognition has reached that final stage. A host of factors have been proposed and have been shown to play a role – onset and rime units (Treiman & Chafetz, 1987), morphemes (Baayen & Schreuder, 2003; Feldman & Basnight-Brown, 2007), length (New, 2006), word frequency (Balota, Cortese, Sergent-Marshall, Spieler, & Yap., 2004), familiarity (Balota, Pilotti, & Cortese, 2001), age of acquisition (Juhasz, 2005), orthographic neighbourhood (Andrews, 1997), and phonological neighbourhood (Yates, Locker, & Simpson, 2004). Early on, it was thought that semantic variables played no role in word recognition as it was assumed a word had to be recognized before meaning could be identified (Balota, Ferraro, & Connor, 1991). However, recent research has been discovering a variety of semantic variables that have a significant and early influence. The concreteness of a word, its ability to be the object of a sense verb, has an effect both in lexical decision (De Groot, 1989) and naming tasks (Bleasdale, 1987). Imageability, the ease with which a person can form a picture of the word's referent in their mind, also has an effect (Strain, Patterson, & Seidenberg, 1995). Similarly, people's ratings of how strongly words evoke sensory experience predicts lexical decision times (Juhasz, Yap, Dicke, Taylor, & Gullick, 2011) as do ratings separated by sensory modality (Amsel, Urbach, & Kutas, 2012) and ratings based on how easily one can physically interact with a word's referent (Siakaluk, Pexman, Aguilera, Owen, & Sears, 2008). Even the previously mentioned danger and usefulness factors have been found to predict reaction times (Wurm & Vakoch, 2000). Despite all of the factors identified so far, much variance remains unexplained (Adelman, Marquis, Sabatos-DeVito, & Estes, 2013).

Emotion has rarely been considered in terms of its effects on word processing. It has been a stretch for researchers to consider that denotative meaning might influence word processing even prior to recognition. Emotion goes beyond denotation to connotation. First of all, one must concur that all words have affective meaning, even if that meaning is neutral. As Duncan and Barrett (2007) said, "There is no such thing as a non-affective thought". Concurrently, one must accept that affect and cognition are not separable but enmeshed together. Affective responses to stimuli occur both in subcortical areas of the brain (heretofore considered affective regions) and in anterior frontal regions (heretofore considered purely cognitive) and the two share reciprocal connections. For example, although the amygdala is considered central to emotional processing, it cannot process facial expressions without information on configuration from the visual cortex (Rolls, Tovee, Purcell, Stewart, & Azzopardi, 1994; Storbeck, Robinson, & McCourt, 2006; Rotschtein, Malach, Hadar, Graif, & Hendler, 2001). However, the agmydala can also influence sensory processing, determining what is attended to and therefore what is available for higher order cognitive functions to work on (Amaral, Behniea, & Kelly, 2003; Amaral & Price, 1984; Freese & Amaral, 2005). As such, there is no reason to exclude emotion from investigations of word processing as there are strong hints that it might play an important role in word recognition, attention capturing during linguistic tasks, and memory for words.

A few studies have looked at how the valence and/or arousal of words impacts how long it takes lexical decision and naming times, the two basic measures of word recognition. In general, negative words are identified and read slower than neutral (Algom et al., 2004). In some studies negative words were identified slower than positive (Estes & Adelman, 2008) while in others, valence showed a quadratic inverse U-shape relationship with reaction time (Kousta, Vinson, & Vigliocco, 2009). The relationship between reaction time and arousal is equally unclear. Estes & Adelman (2008) argue that the disadvantage for negative words is constant across all levels of arousal while Larsen, Mercer, Balota & Strubel (2008) argue that a slowdown only occurs with low to moderately arousing negative words. I will return to this debate in Chapter 4 with new data. Even fewer studies have looked beyond individual words to sentence processing. Scott, O'Donnell, and Sereno (2012) used eye-tracking during natural reading and found that with low frequency words, negative, high-arousal words were fixated longer than neutral words. With high frequency words, negative words were also fixated longer than positive words.

A variety of cognitive paradigms have included emotional words as stimuli and shown that they capture attention, interfering with simultaneous and subsequent tasks. In emotional Stroop tasks, negative words have been found to interfere with people's ability to respond to the color of the word while ignoring its meaning (McKenna & Sharma, 1995; Williams, Mathews, & MacLeod, 1996). Again, the effect may be more about arousal than valence as in a subsequent study, equally arousing positive and negative words caused more interference than neutral low arousing words (Dresler, Meriau, Heekeren, and van der Meer, 2009). Similarly, high arousal words have been found to cause the most interference in identifying the parity of adjacent numbers (Aquino & Arnell, 2007). With regards to subsequent tasks, identification accuracy of a second target in a rapid serial presentation stream has also been shown to be negatively impacted by the arousal of the first target (Mathewson, Arnell, & Mansfield, 2008). Within the framework of a two-stage theory of attention, some researchers have proposed that the emotion associated with words particularly monopolizes processing in Stage 2 to the detriment of other competing stimuli. For example, a high arousal cue word interfered with the identification of a neutral target word but not when the cue was masked, restricting access to Stage 2 nor when the ISI was long reducing competition (Bocanegra & Zeelenberg, 2009).

Memory for words also varies as a function of emotion. Item and source memory has been shown to be better for both positive and negative words than for neutral (Doerksen & Shimamura, 2001; Kensinger & Corkin, 2003a). Subsequent research showed this effect to be strongest in long-term memory as opposed to short-term (Kensinger & Corkin, 2003b) and largely a result of arousal with high arousal words being remembered better than neutral after a 24-hour delay (Sharot & Phelps, 2004).

As such, it appears that emotion has the potential to affect every stage of word processing from the moment that a word attracts a person’s attention perhaps competing against distractors, to the process of identifying its orthographic features and ultimately the word itself, to retrieving that word from memory.

What are the limitations of this past research?

While an important starting point, the studies cited herein have some significant limitations. In most of the above cited studies, the number of stimuli used is quite small (see Table 1for examples and numbers). This certainly restricts the generalizability of their results but it also raises a number of other issues. Because ANEW (Bradley & Lang, 1999), the largest set of affective norms until now only includes 1,034 words, after controlling for all the important factors related to word frequency such as frequency, familiarity, age of acquisition, etc., the number of words available for each condition of an emotion based study is quite small. In addition, ANEW contains an over-representation of emotionally charged words which means it does not accurately represent the full distribution of emotion across the language. To compensate, some researchers have collected their own norms for the stimuli they wish to use. The difficulty with this solution is that it can skew the emotional ratings.

There is previous evidence that using stimuli in blockeded versus mixed lists can change results (e.g. Hulme, Stuart, Brown & Morin, 2003; Raman, Baluch, & Besner, 2004). As such, it is possible that when emotional words are concentrated together in a list, they are rated differently than when they are scattered among neutral words. Other researchers have chosen to equate extremes of valence with high arousal, contrasting these two conditions with neutral. This, however, leaves out critical portions of the lexicon in that it doesn't address what happens with other combinations such as high valence, low arousal or low valence, low arousal. Additionally, dichotomizing what are considered to be continuous variables and analyzing them via an ANOVA results in a significant loss of power and interpretability (Baayen, 2010).

Table 1: List of studies cited in the Introduction that used emotional words as stimuli. The table indicates the paradigm, number and classification of words, and the type of statistical test used.

Authors	Experiment Type	Word Stimuli	Test
Algom, et al. (2004) – Exp 5	lexical decision	16 threat, 16 neutral	t-test
Larsen, et al. (2008)	naming & lexical decision	1,021 words with matches in both ANEW and ELP	regression
Estes and Adelman (2008)	naming & lexical decision	1,011 words with matches in ANEW, ELP, and CELEX	regression
Kousta, et al. (2009)	lexical decision	40 positive and 40 negative (matched by arousal), 40 neutral	ANOVA
Dresler, et al. (2009)	emotional Stroop	20 positive and 20 negative (matched by arousal), 20 neutral	ANOVA
McKenna & Sharma (1995)	emotional Stroop	5 neutral, 5 negative, 5 letter strings	ANOVA
Aquino & Arnell (2007)	digit parity	25 threat, 25 neutral, 25 sexual, 25 school-related	ANOVA
Mathewson, et al. (2008)	rapid serial visual presentation	24 neutral, 24 negative, 24 positive, 24 taboo target words; 59 neutral low-arousal distracters	regression
Bocanegra & Zeelenberg (2009)	masked visual identification with forced choice	52 mainly negative high-arousal, 52 neutral low-arousal	ANOVA
Doerksen & Shimamura (2001)	color association with free recall and recognition tests	32 pleasant, 32 unpleasant, 64 neutral	ANOVA, t-tests
Kensinger & Corkin, 2003a	backward and alphabetical word span; surprise delayed recall test	10 arousing taboo, 10 neutral	ANOVA
Sharot & Phelps (2004)	immediate and delayed recall tests for both centrally and peripherally presented words	16 negative high arousing, 15 neutral	ANOVA
Scott, O'Donnell, & Sereno (2012)	eye-tracking during natural reading	12 low frequency positive, negative, and neutral triplets, 12 high	ANOVA

		frequency (positive and negative both high in arousal compared to neutral)	
Hofmann, et al. (2009)	EEG with lexical decisions	50 positive, 50 neutral, 50 negative (all low-arousal) and 50 high-arousal negative	ANOVA
Kissler, et al. (2007)	EEG with rapid serial visual presentation	60 high-arousal pleasant, 60 high-arousal negative, 60 low-arousal neutral	ANOVA
Bayer, et al. (2010)	EEG with semantic decision task	50 high-arousal negative sentence final verbs paired with 50 low-arousal neutral; 48 pairs with matching valence but contrasting arousal	ANOVA
Citron, Gray, Critchley, Weeks, & Ferstl (2014)	fMRI with lexical decision	35 positive high-arousal, 35 positive low-arousal, 35 negative high-arousal, 35 negative low-arousal, 35 neutral low-arousal	ANOVA
Zhang et al. (2014)	EEG with rapid serial visual presentation	6 positive, 6 negative, and 6 neutral adjectives	ANOVA

Another limitation is that the studies to date have not considered group or individual differences. There has been some suggestion that the inconclusivity of research on discrete emotions may be due to ideographic differences in patterns of responding (Friedman, 2003; Wallbott & Scherer, 1991) but this avenue has not been strongly pursued (Barrett, 2006b). Some research has confirmed that dimensional explanations of emotions best explain both individual and aggregate self-reports (Barrett, 1998, 2004; Feldman, 1995). There is also evidence that people differ in how intensely they experience emotion (Barrett & Niedenthal, 2004; Conner, Barrett, & Bliss-Moreau, 2005). But none of the studies that actually evaluate how emotion affects word processing or other cognitive tasks have evaluated their results via group or individual variables.

Finally, research on what underlies emotion, how emotion is connected to basic level motivations, has been restricted to a single method. Several researchers argue that emotion, particularly in terms of valence, is actually a preparation for action – evaluating something as positive prepares one to approach it while evaluating something as negative prepares one to avoid it (Osgood, 1953; Lang, Bradley, & Cuthbert, 1990). Emotion is thought to trigger the fundamental motivational systems that drive human behaviour, sometimes referred to by the action of either approach or avoidance, sometimes referred to by the nature of the stimulus either appetitive or aversive (Carver & White, 1994; Lang, 1995). This link has been demonstrated by the fact that people are faster to approach positive vs. negative stimuli and to avoid negative vs. positive stimuli, regardless of what the actual response method is (e.g. Chen & Bargh, 1999, De Houwer, Crombez, Baeyens, & Hermans, 2001; for a review, see Krieglmeyer & Deutsch, 2010). This motivational link has also been used to explain why emotional stimuli capture attention as they activate this preparation for action thereby taking resources from other tasks. None of these studies have addressed the possible impact of arousal. Given that arousal is often considered a measure of intensity, it is possible that it moderates the degree to which valence triggers the motivational systems related to valence. However, the current methodology for eliciting the congruency effect does not allow for a measure of degree. Congruent conditions are either faster than incongruent or they are not. There is no in-between.

How does this dissertation address these limitations?

If the field is to make progress in understanding how emotion affects word-processing, it needs access to a much larger set of emotional norms. With such a resource, researchers would be able to select appropriate and more representative sets of stimuli while maintaining the necessary controls. In addition, consistent use of the same resource across multiple studies would remove the risk of sampling bias and allow for comparison of results. In Chapter 2, I present a study in which we collected valence, arousal, and dominance ratings for 13,915 lemmas (uninflected word forms), the largest collection of emotional word norms to date (Warriner, Kuperman & Brysbaert, 2013). Over 1,800 participants contributed to this set of ratings allowing us to calculate mean ratings by gender, age, and education level. This represents the first comprehensive look at group differences in emotional norms. The ratings we obtained showed high correlations with other smaller sets of norms including those in other languages, thus confirming their reliability. This dataset has been made available to researchers world-wide. It was this dataset that enabled the contributions that will be discussed in the remaining chapters. Research on another promising semantic variable, concreteness, has been similarly hampered by small sets of norms. As such, we ran another study (presented as Chapter 3) to collect concreteness ratings to nearly 40,000 English words, more than 4 times larger than any previous collection (Brysbaert, Warriner, & Kuperman, 2014). Over 4,000 people participated in this study and again we collected information about gender, age, and education level. Concreteness effects have been studied in relation to working memory (Nishiyama, 2013), long-term memory (Hanley, Hunt, Steed & Jackman, 2013), word-processing (Barber, Otten, Kousta & Vigliocco, 2013), and importantly, in relation to affective connotation (Ferre, Guasch, Moldovan, & Sanchez-Casas, 2012; Kousta, Vigliocco, Vinson, Andrews, & Del Campo, 2011). This set of norms will allow for continued research into these areas and into embodied cognition in general, which connects meaning to experience (Vigliocco, Vinson, Lewis & Garett, 2004; Andrews, Vigliocco & Vinson, 2009).

As mentioned earlier, there is a debate about how valence and arousal predict lexical decision and naming times. Estes and Adelman (2008) tested valence and arousal independently finding that increasing arousal speeded up recognition in a linear fashion while positive valence words were categorically faster than negative words. Larsen, et al. (2008) replicated these findings but also found that arousal and valence interacted in the lexical decision task such that the effects of valence were only found with low arousal words. Estes and Adelman (2008) came back with the argument that this was because Larsen et al. (2008) modeled valence as linear instead of categorical and that the interaction disappeared in the latter case. Both studies used the ANEW dataset (Bradley & Lang, 1999) in which words were selected for their emotionality and thus lacked mid-range items which are in fact, rather frequent in natural language. Kousta et al. (2009) thus supplemented ANEW with a selection of neutral words and found no effect of arousal on lexical decision times and a speeded response with emotional words regardless of valence. As one can see, the conclusions from these studies differ significantly likely due to inconsistent methodology and sampling. In addition, none of these studies considered an interaction between emotion and frequency. Frequency explains a large amount of variance in word recognition times (Yap & Balota, 2009) and interacts with other factors such as imageability and age of acquisition (e.g. Cortese & Schock, 2013; Gerhand & Barry, 1999a, 1999b) and other measures such as Stroop interference (Kahan & Hely, 2008) and fixation duration (Scott et al., 2012). In Chapter 4, we revisit this debate, using the large dataset discussed in Chapter 2 (Warriner et al., 2013), controlling for more lexical and semantic factors, and considering interactions both between valence and arousal and between these variables and word frequency. We find that reaction times decrease with increasing valence but increase with rising arousal. These effects are independent from each other but interact with word frequency such that they are largest among low-frequency words. Including valence and arousal into existing models of word recognition explains an additional 2% of unexplained variance in lexical decision times and 0.2% in naming times. We consider how these results require a revision of the automatic vigilance hypothesis.

A discussion of how emotion impacts word processing is not complete without a consideration of what motivates that impact. As stated earlier, approach or avoidance motivation is thought to be intricately linked to valence, however, little has been done to investigate what role arousal plays. In addition, the dichotomization of a congruency effect does not allow for any measure of variation. Given that both valence and arousal are continuous measures, and that there is solid evidence that they interact with other processes such as lexical decision in a continuous manner (Chapter 4), it would seem prudent to use a method that allows for variance. In Chapter 5, we introduce a new method for measuring just how close or far a person wants to be from stimuli that vary in both valence and arousal. We then test this methods sensitivity to gender and individual differences.

What can language tell us about emotion?

People appear to be readily able to associate emotion with words and there seems to be a level of consistency between individuals. The shape of the relationship between valence and arousal in word ratings studies (an inverse-U shape with arousal being higher for extremely pleasant and unpleasant words) is quite similar to the shape found in immediate and post-hoc reports of experienced emotion in daily life, and ratings to emotional pictures (Kuppens, Tuerlickx, Russell & Barrett, 2013). This highly suggests that there is sense in which emotional experience is instantiated in or reflected by the availability of words with which to communicate that experience. While this makes intuitive sense, it has not yet been investigated as such, most likely due to a paucity of normed word ratings. Previous studies have certainly suggested similar ideas when they explored what Boucher and Osgood (1969) coined as the “Pollyanna hypothesis” – an observation that positive words occur more frequently than negative words (Augustine, Mehl & Larsen, 2011; Rozin, Berman, & Royzman, 2010; Unkelbach et al., 2010). In Chapter 6, we revisit and extend this observation. We combine the affective ratings with lexical frequency information to examine the distribution of emotion across the English language both by types (individual words) and by tokens (every instantiation of a word). We also look at the distribution of arousal for the first time. By doing so, we are able to identify several compound emotional biases in English and consider how they might arise from communicative and social motivations. This is an important contribution to the field as it speaks to why language might encode emotion to begin with and therefore what might undergird the effects that emotion has on word processing.

CHAPTER 2: Norms of valence, arousal, and dominance for 13,915 English lemmas

Warriner, A.B., Kuperman, V., & Brysbaert, M. (2013).

Behavior Research Methods, 45, pp.1191-1207.

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.

Keywords: Emotion, Semantics, Gender differences, Age differences, Crowdsourcing

Introduction

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 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

[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)

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

	Number of Participants		
Education Level	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.

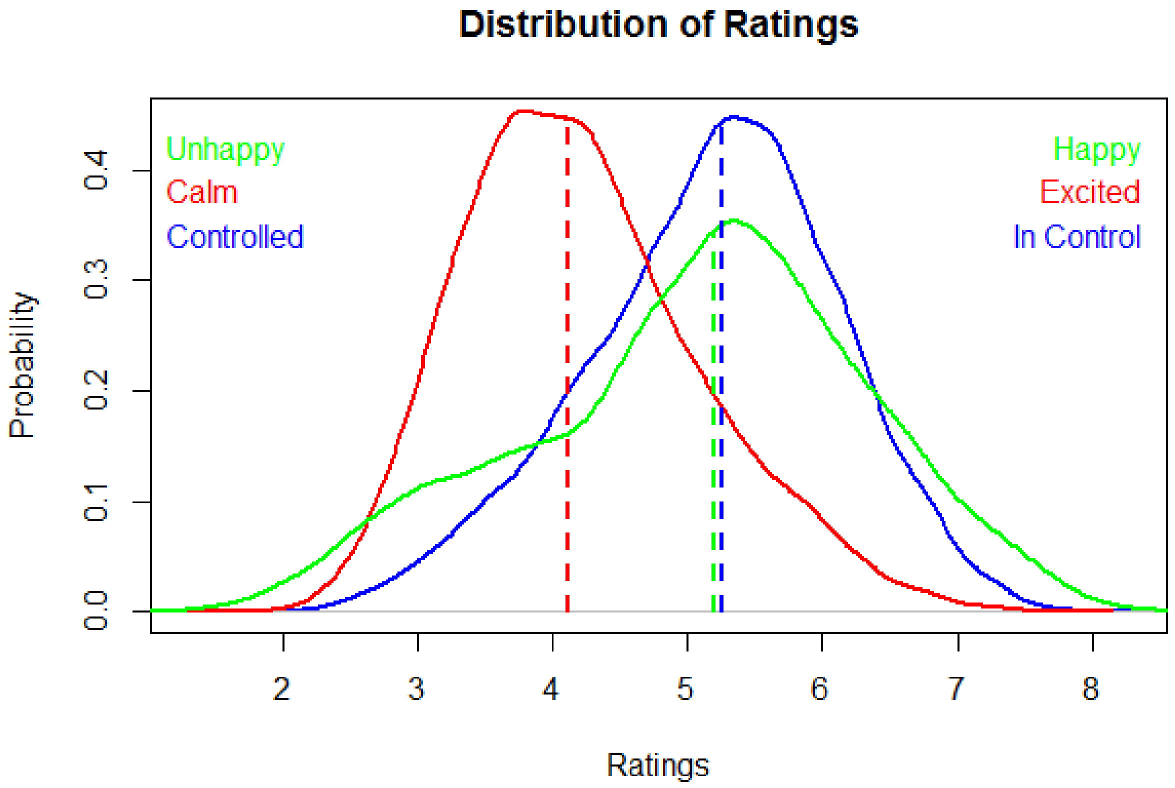


Figure 1. Distributions of valence (green), arousal (red) and dominance (blue) ratings. Dotted lines represent the medians of respective distributions.

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.

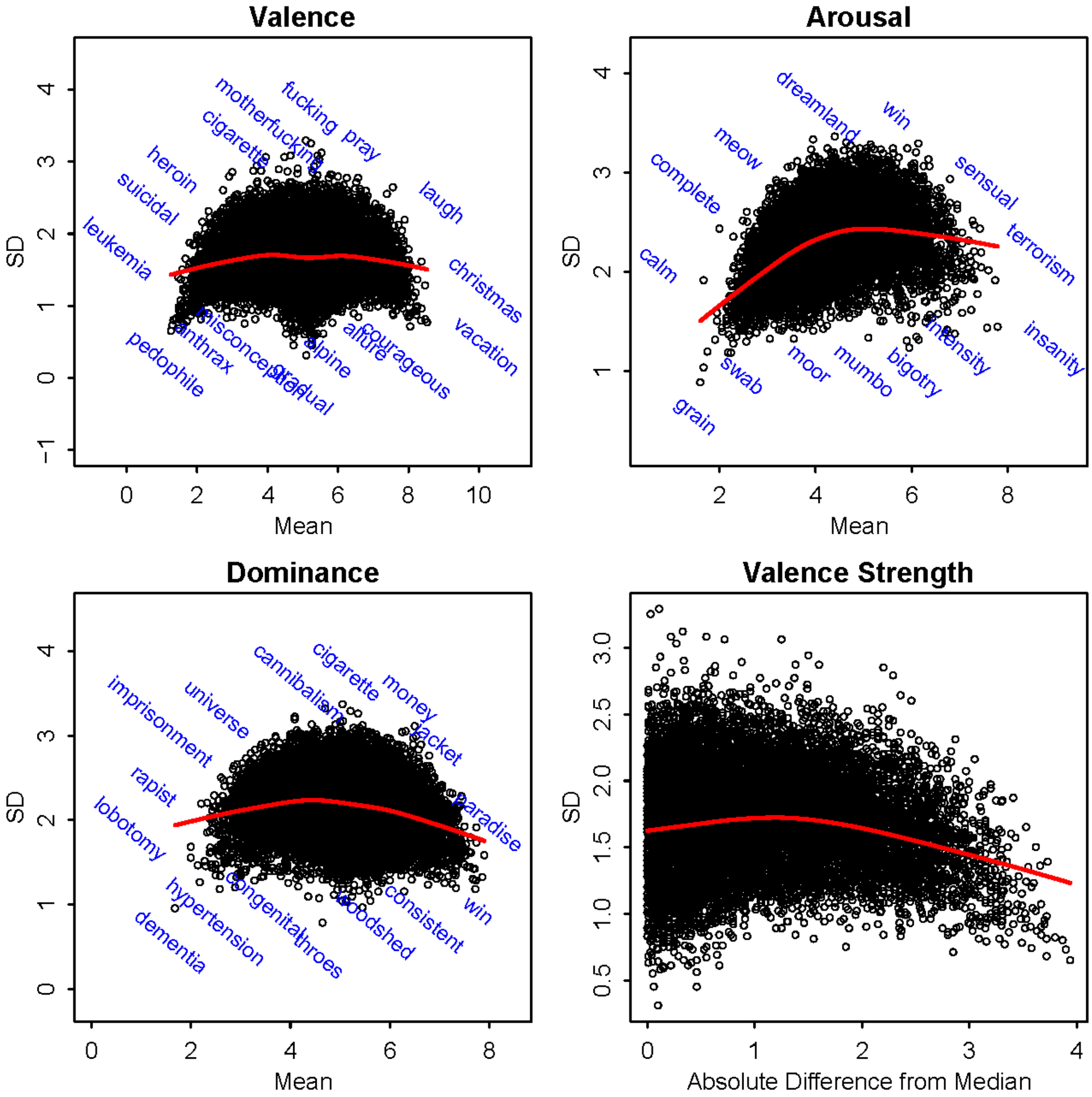


Figure 2. Standard deviation of ratings for valence (a, top left), arousal (b, top right), dominance (bottom left) and valence strength (d, bottom right) plotted against respective mean ratings. Panels a-c also provide examples of words with disproportionately large and small standard deviations given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

	R	Linear Co.	Quadratic Co.	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

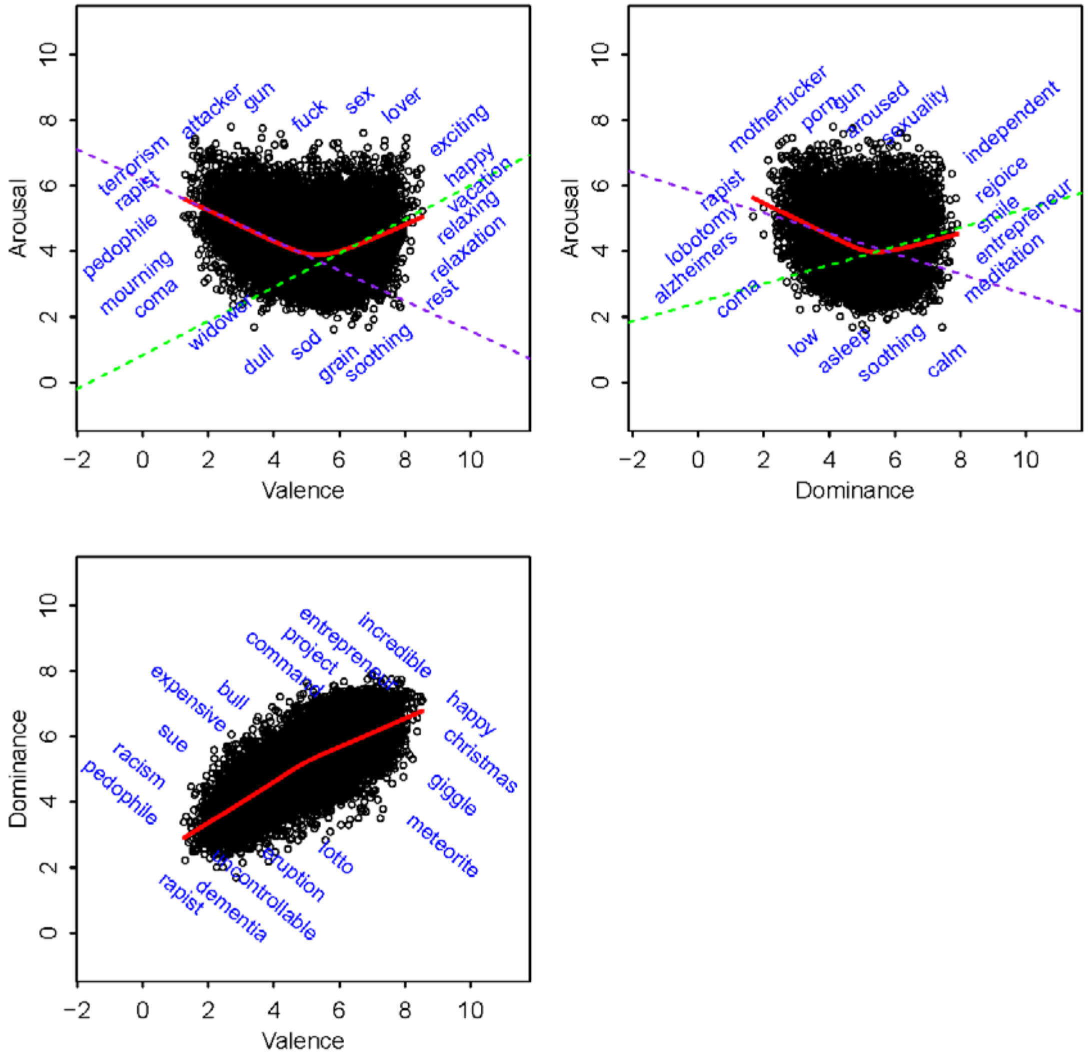


Figure 3: Scatterplots of dimensions (top left, arousal vs valence; top right, arousal vs dominance; bottom left, Dominance vs Valence) along with lowess lines (in red) showing the functional relationships and regression lines for arousal as predicted by high (in green) and low (in purple) valence and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

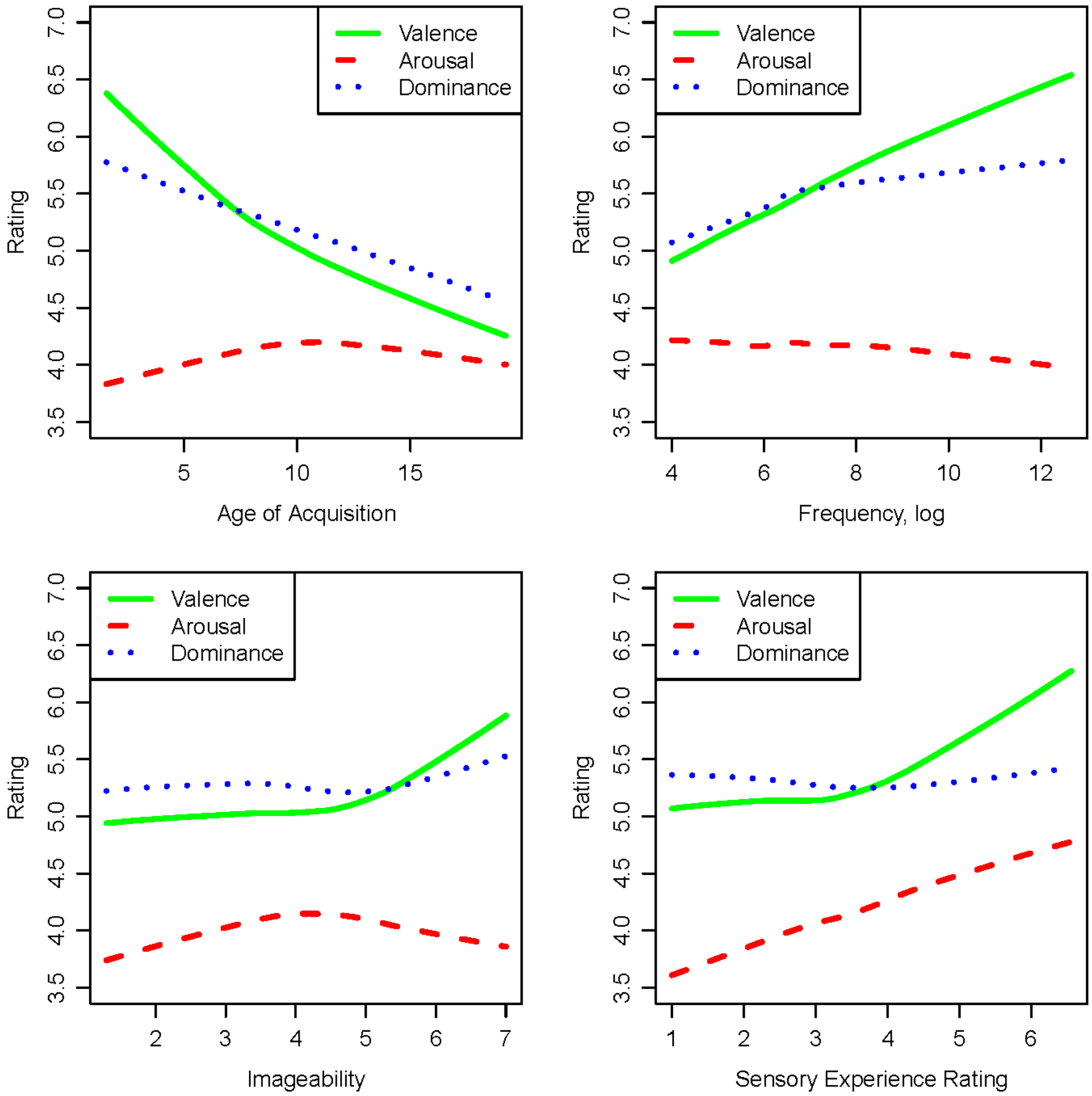


Figure 4. Relationship between the three dimensions and Age of Acquisition, Imageability, and Sensory Experience Ratings, presented as scatterplot smoother lowess trend lines.

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 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 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: 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.

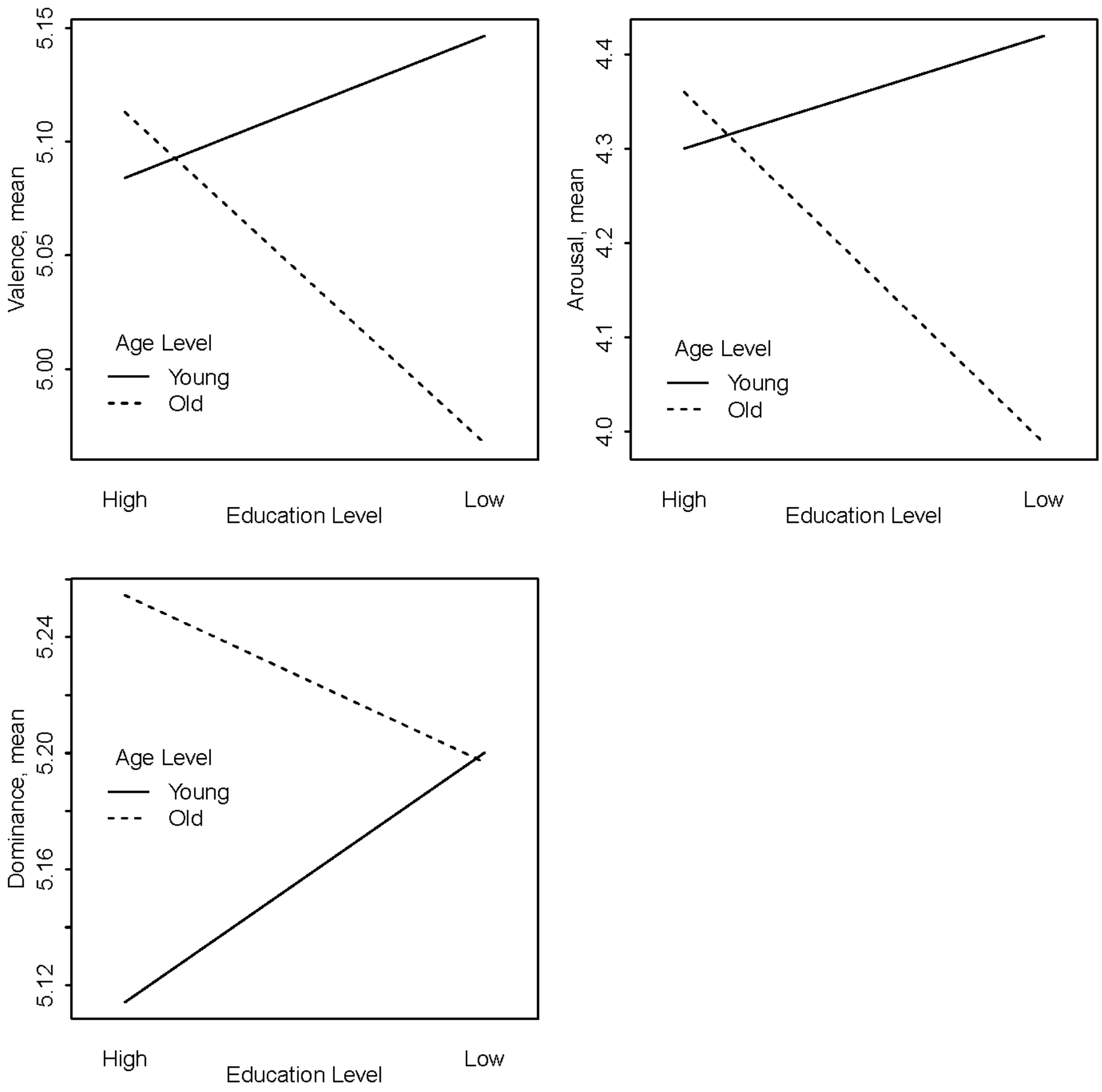


Figure 5. Interactions between dichotomized education and age levels for all three dimensions. All interactions are significant at p < .001.

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.

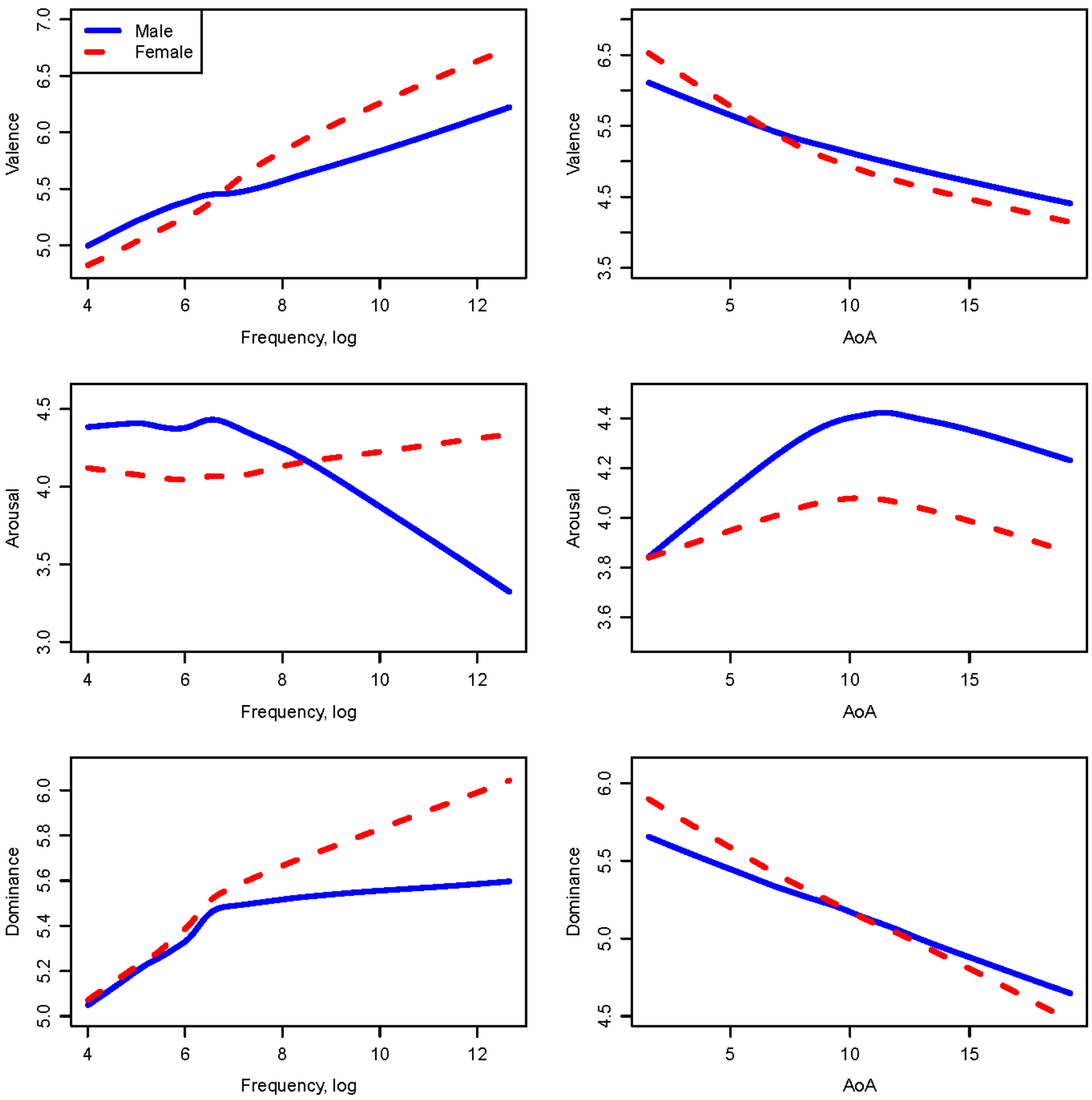


Figure 6. Interactions of gender with frequency (left) and age-of-acquisition (right) as predictors of mean ratings of valence (top), arousal (middle) and dominance (bottom). Interactions are presented with gender-specific lowess trend lines.

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.

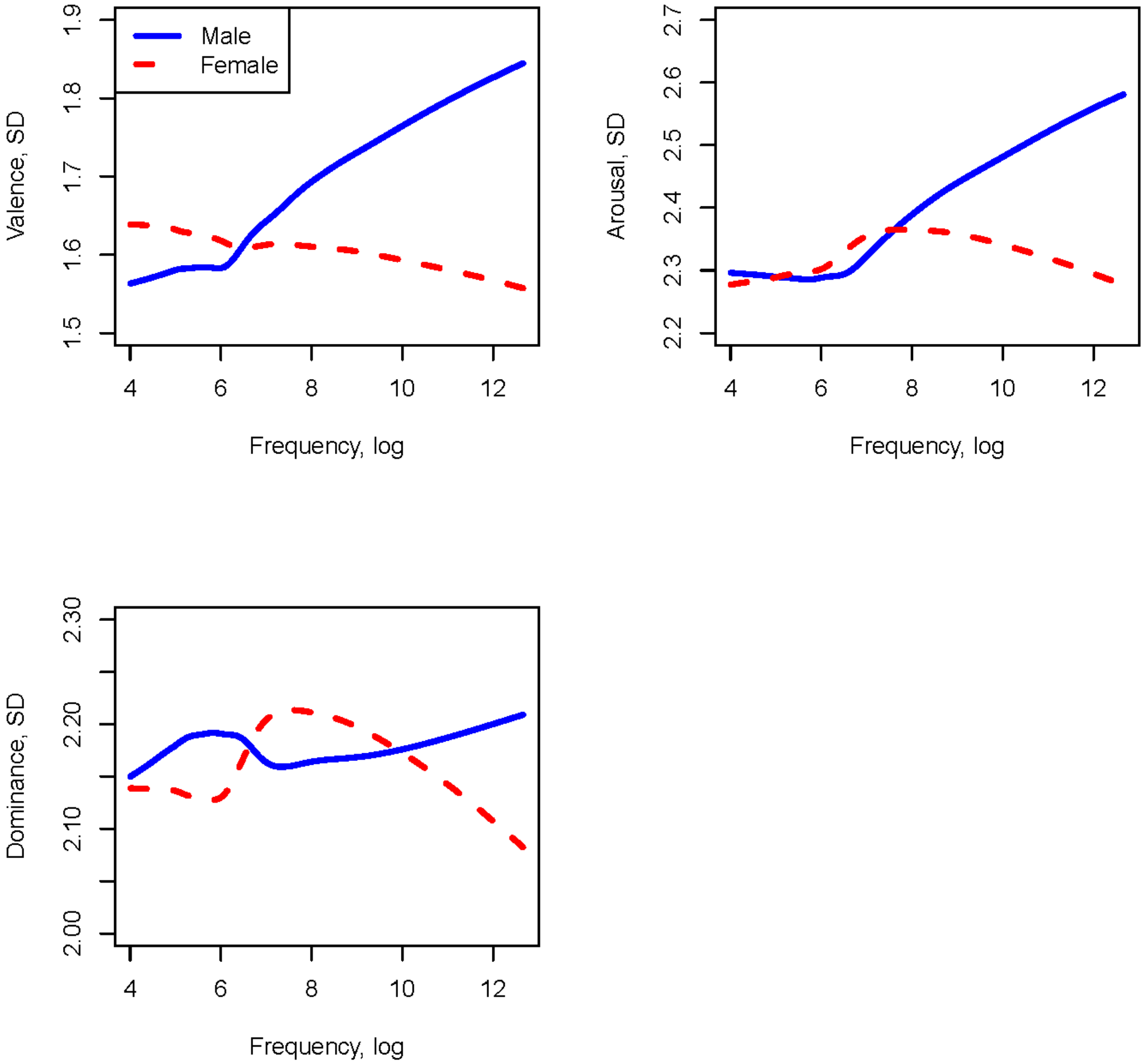


Figure 7. Interactions of gender with frequency as a predictor of standard deviations of ratings of valence (top left), arousal (top right) and dominance (bottom left). Interactions are presented with gender-specific lowess trend lines.

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.

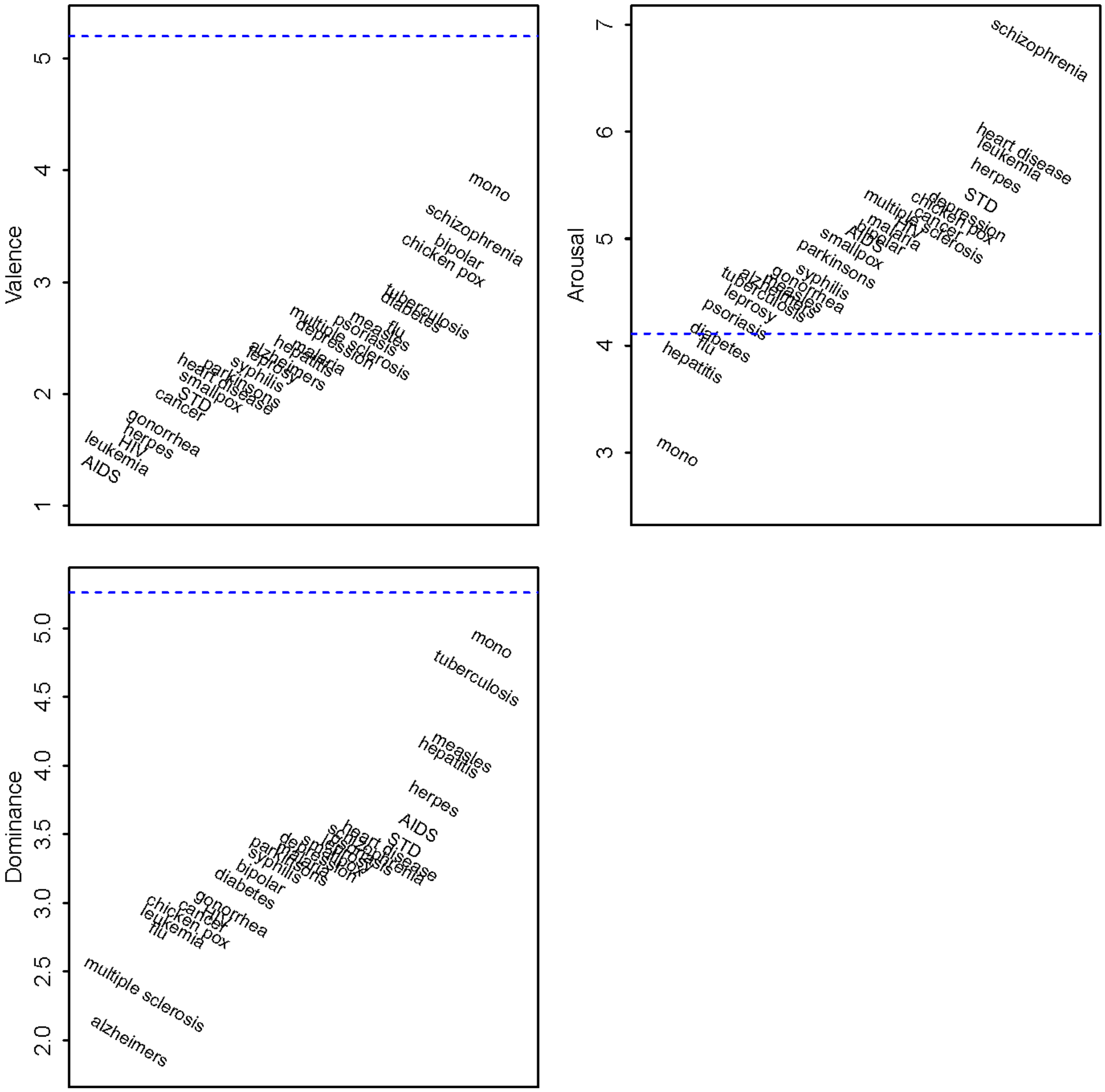


Figure 8. Ratings of words denoting disease. Dotted lines represent median ratings of respective emotional dimensions in the entire dataset.

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.

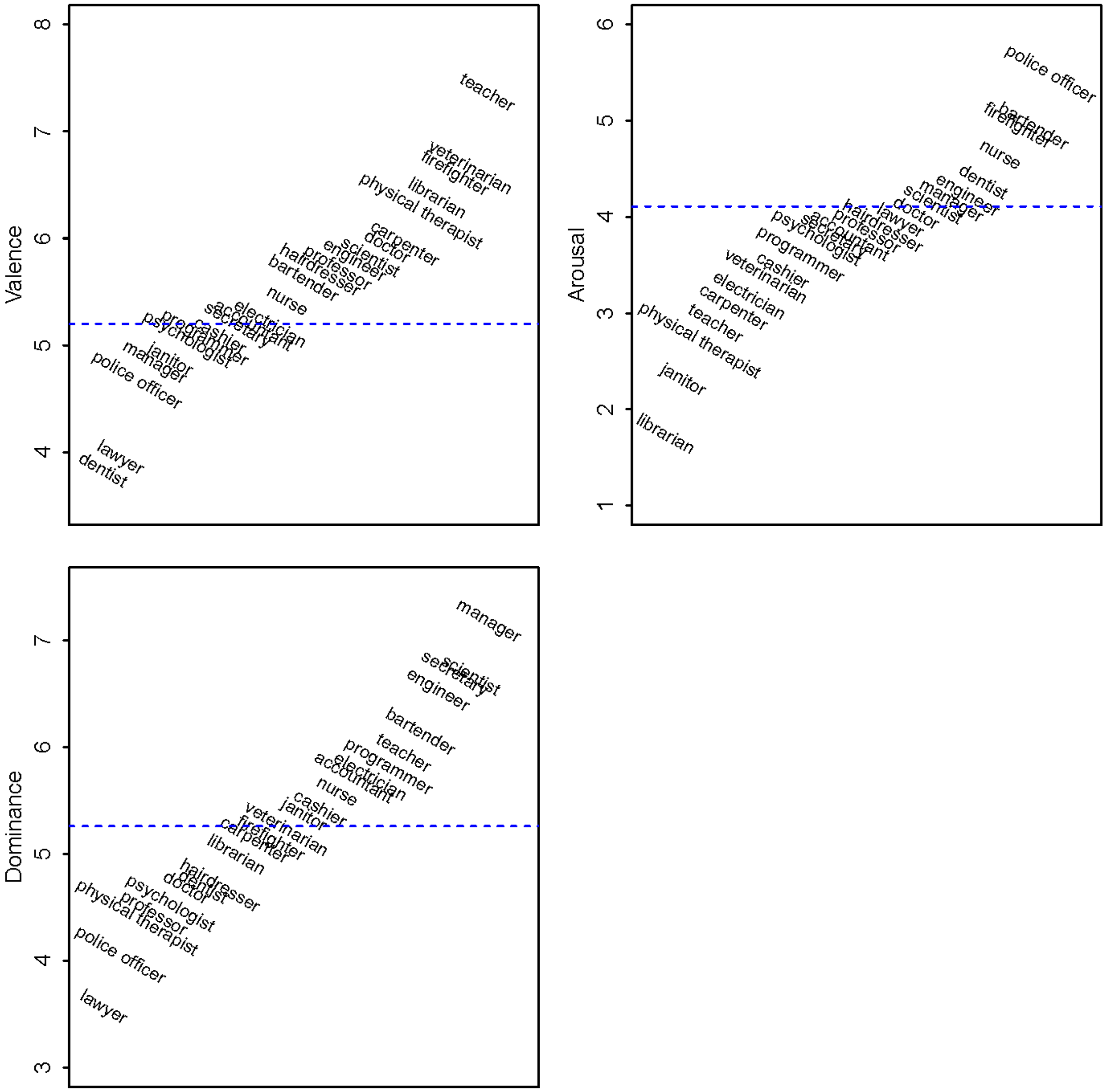


Figure 9. Ratings of words denoting occupations. Dotted lines represent median ratings of respective emotional dimensions in the entire dataset.

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.

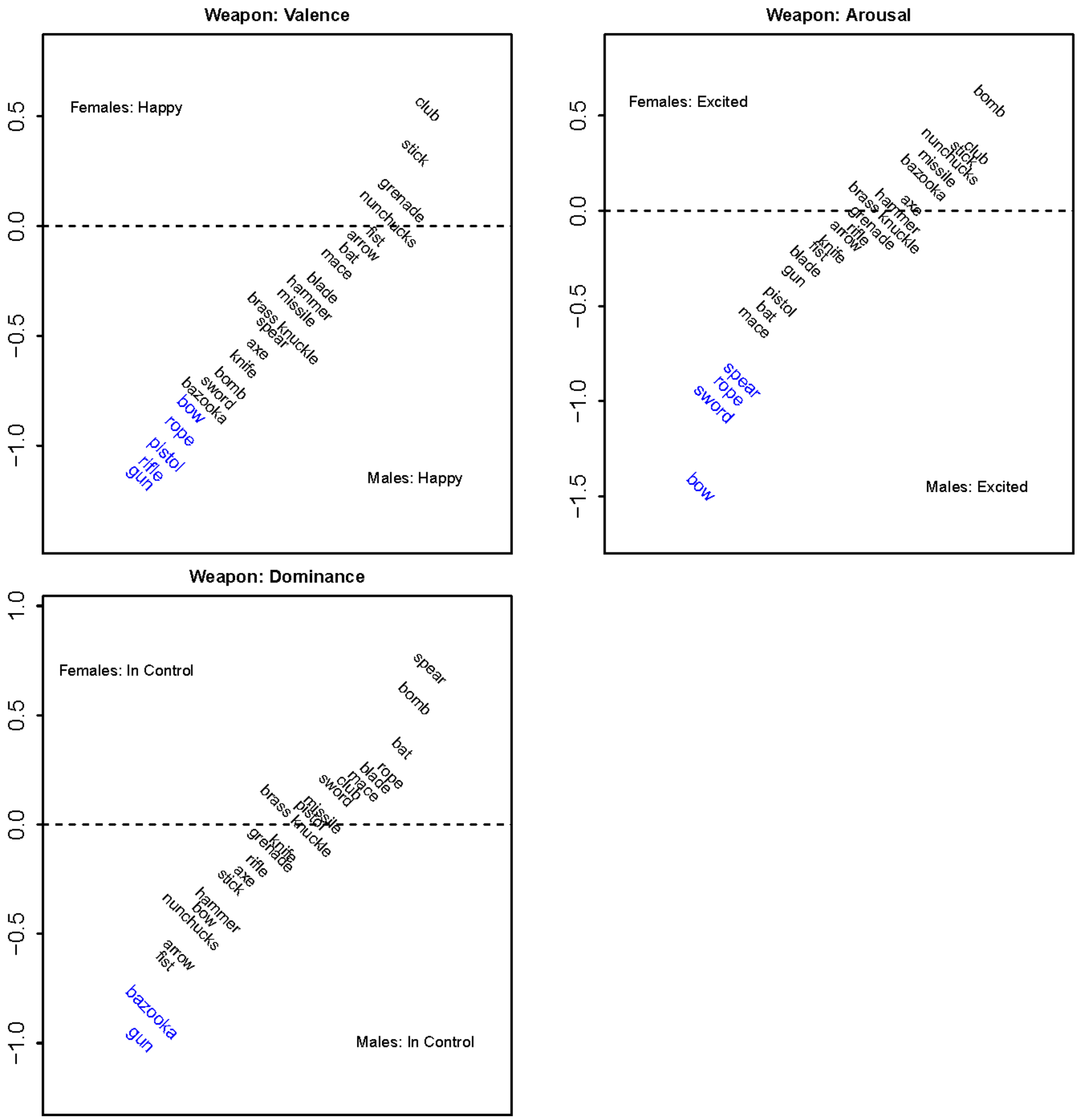


Figure 10. Gender differences in ratings for weapon related words.

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).

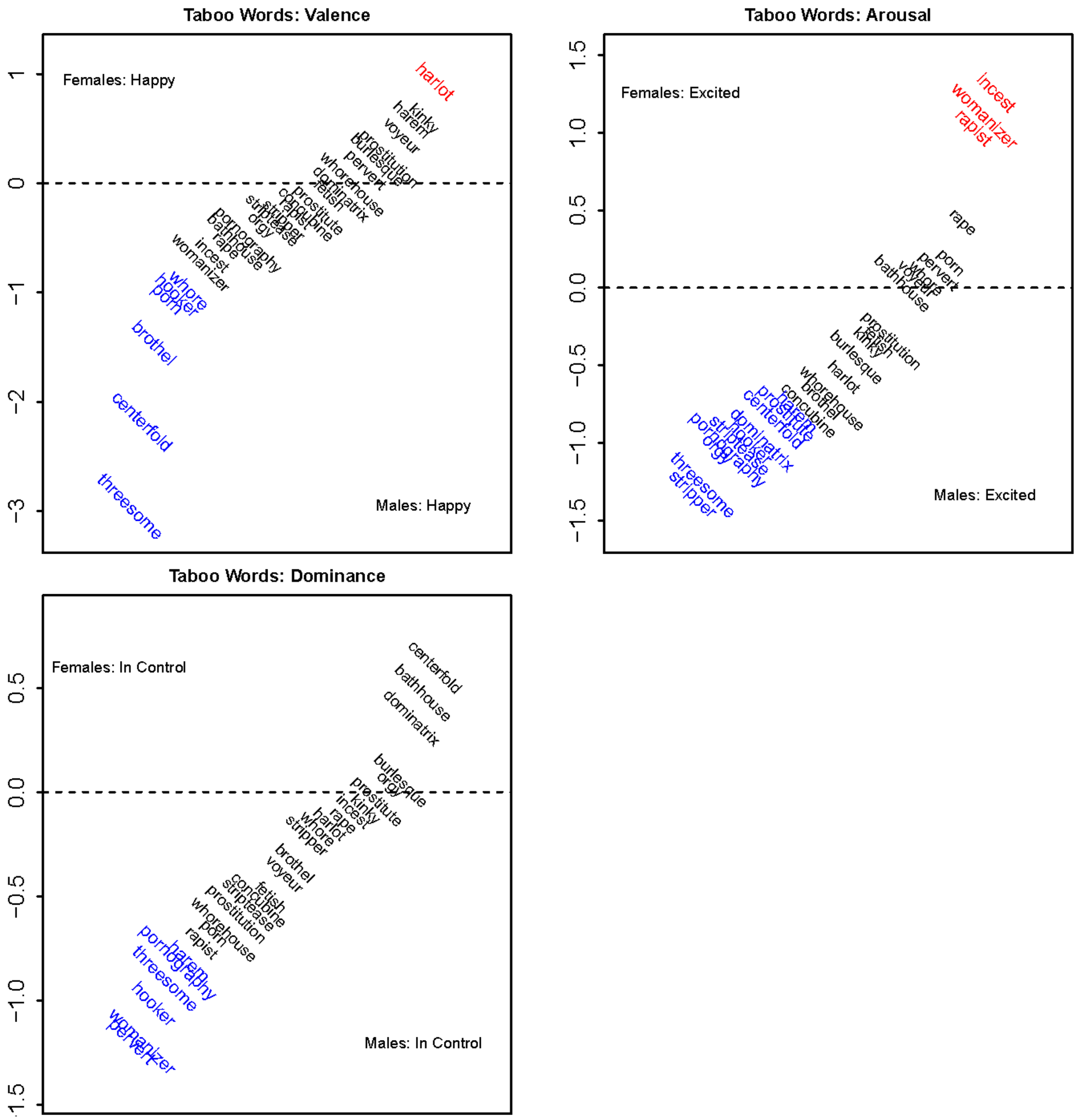


Figure 11. Gender differences in ratings for taboo words.

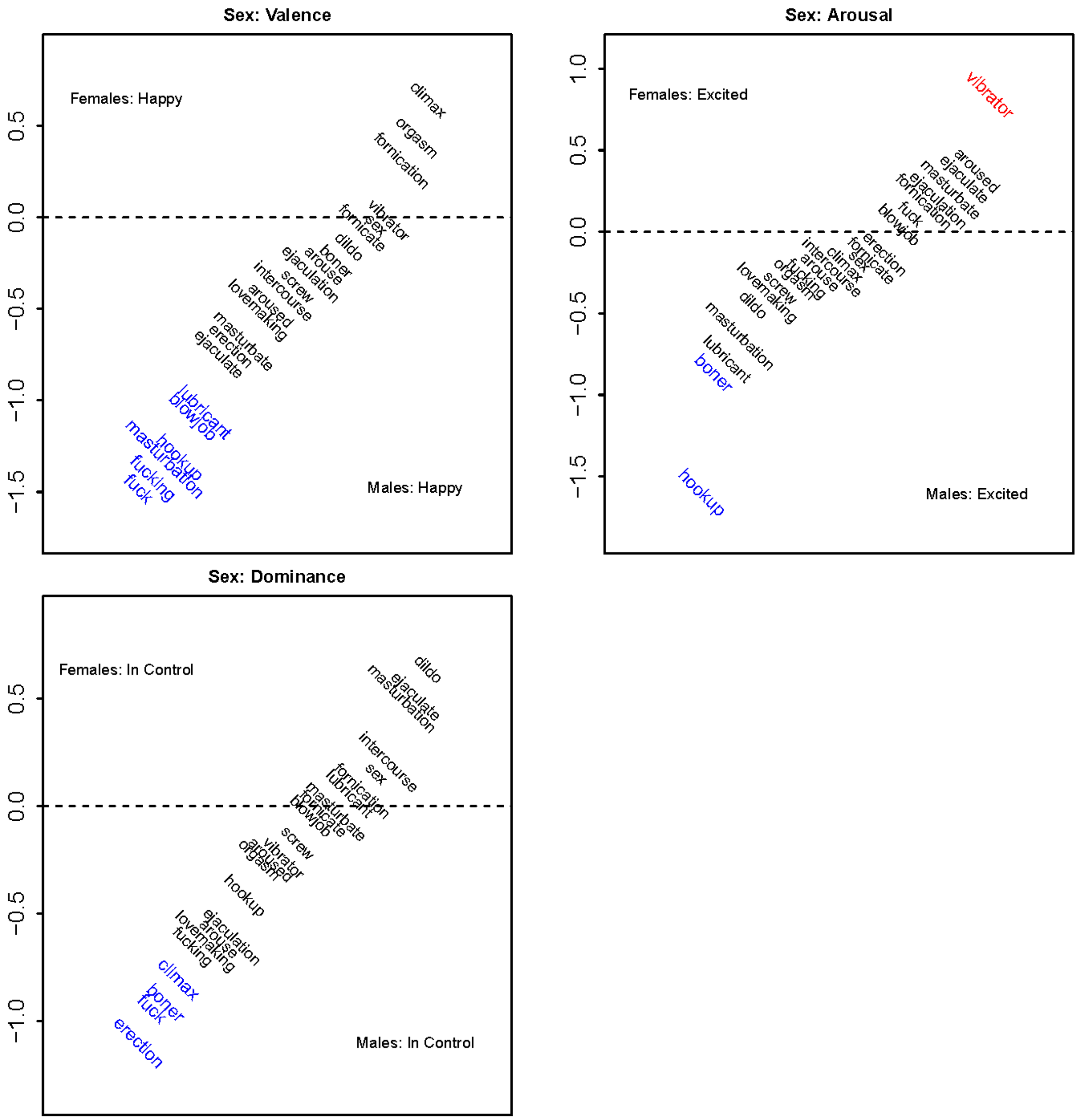


Figure 12. Gender differences in ratings for sex related words.

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.

CHAPTER 3: Concreteness ratings for 40 thousand generally known English word lemmas

Brysbaert, M., **Warriner, A.B.**, & Kuperman, V. (2014). *Behavior Research Methods, 46(3), 904-911*. doi:10.3758/s13428-013-0403-5

Abstract

Concreteness ratings are presented for 37,058 English words and 2,896 two-word expressions (such as zebra crossing and zoom in), obtained from over 4,000 participants by means of a norming study using Internet crowdsourcing for data collection. Although the instructions stressed that the assessment of word concreteness would be based on experiences involving all senses and motor responses, a comparison with the existing concreteness norms indicates that participants, as before, largely focused on visual and haptic experiences. The reported data set is a subset of a comprehensive list of English lemmas and contains all lemmas known by at least 85 % of the raters. It can be used in future research as a reference list of generally known English lemmas.

Keywords: Concreteness, Ratings, Crowdsourcing, Word recognition

Introduction

Concreteness evaluates the degree to which the concept denoted by a word refers to a perceptible entity. The variable came to the foreground in Paivio's dual-coding theory (Paivio, 1971, 2013). According to this theory, concrete words are easier to remember than abstract words, because they activate perceptual memory codes in addition to verbal codes. Schwanenflugel, Harnishfeger, and Stowe (1988)presented an alternative context availability theory, according to which concrete words are easier to process because they are related to strongly supporting memory contexts, whereas abstract words are not, as can be demonstrated by asking people how easy it is to think of a context in which the word can be used.

The importance of concreteness for psycholinguistic and memory research is hard to overestimate. A search through the most recent literature gives the following, nonexhaustive list of topics related to concreteness. Are there hemispheric differences in the processing of concrete and abstract words (Oliveira, Perea, Ladera, & Gamito, 2013)? What are the effects of word concreteness in working memory (Mate, Allen, & Baqués, 2012; Nishiyama, 2013)? How are concrete and abstract concepts stored in and retrieved from long-term memory (Hanley, Hunt, Steed, & Jackman, 2013; Kousta, Vigliocco, Vinson, Andrews, & Del Campo, 2011;Paivio, 2013)? Does concreteness affect bilingual and monolingual word processing (Barber, Otten, Kousta, & Vigliocco, 2013; Connell & Lynott, 2012; Gianico-Relyea & Altarriba, 2012; Kaushanskaya & Rechtzigel, 2012)? Do concrete and abstract words differ in affective connotation (Ferre, Guasch, Moldovan, Sánchez-Casas, 2012;Koustaet al., 2011)? Do neuropsychological patients differ in the comprehension of concrete and abstract words (Loiselle et al., 2012)?

Concreteness gained extra interest within the embodied view of cognition (Barsalou, 1999;Fischer & Zwaan, 2008; Wilson, 2002)—in particular, when neuroscience established that words referring to easily perceptible entities coactivate the brain regions involved in the perception of those entities. Similar findings were reported for action-related words, which coactivate the motor cortex involved in executing the actions (Hauk, Johnsrude, & Pulvermüller, 2004). On the basis of these findings, Vigliocco, Vinson, Lewis, and Garrett (2004; see also Andrews, Vigliocco, & Vinson, 2009)presented a semantic theory, according to which the meaning of concepts depends on experiential and language-based connotations to different degrees. Some words are mainly learned on the basis of direct experiences; others are mostly used in text and discourse. To make the theory testable, Della Rosa, Catricala, Vigliocco, and Cappa (2010) collected ratings of mode of acquisition, in which participants were asked to indicate to what extent the meaning of a word had been acquired through experience or through language. Unfortunately, to our knowledge these (Italian) norms have not yet been used to predict performance in word-processing tasks.

A final reason why concreteness has been a popular variable in psychological research is the availability of norms for a large number of words. Ratings were collected by Spreen and Schulz (1966), Paivio (both in Paivio, Yuille, & Madigan, 1968, and in unpublished data) and made available in the MRC database (Coltheart, 1981) for 4,292 words. The same database provides imageability ratings (closely related to the concreteness ratings) for 8,900 words. Throughout the years, authors have collected additional concreteness or imageability norms for specific subsets of words (e.g., Altarriba, Bauer, & Benvenuto, 1999; Schock, Cortese, & Khanna, 2012; Stadthagen-Gonzalez & Davis, 2006), which could be combined with the MRC ratings.

Impressive though the existing data sets are, developments in the past years have rendered them suboptimal. First, even 9, 000 words is a limited number when viewed in the light of recently collected megastudies. For instance, the English Lexicon Project (Balota et al., 2007) contains processing times for more than 40,000 words, and the British Lexicon Project (Keuleers, Lacey, Rastle, & Brysbaert, 2012) has data for more than 28,000 monosyllabic and disyllabic words. This means that concreteness ratings are available only for limited subsets of available behavioral data sets.

A second limitation of the existing concreteness ratings is that they tend to focus too much on visual perception (Connell & Lynott, 2012; Lynott & Connell, 2009, in press)atthe expense of the other senses and at the expense of action-related experiences. Lynott and Connell (2009) asked participant to what extent adjectives were experienced “by touch," "by hearing," "by seeing," "by smelling," and "by tasting" (five different questions). Connell and Lynott observed that these perceptual strength ratings were correlated only with concreteness ratings for vision, touch, and, to a lesser extent, smell. They were not correlated for taste and were even negatively correlated for auditory experiences. Similarly, none of the concreteness ratings collected so far includes the instruction that the actions one performs are experience based as well (and hence, concrete).

To remedy the existing limitations, we decided to collect new ratings for a large number of stimuli. This also allowed us to address another enduring problem in word recognition research—namely, the absence of a standard word list to refer to. Individual researchers use different word lists for rating studies and word recognition megastudies, mostly based on existing word frequency lists. A problem with some of these lists is that they contain many entries that, depending on the purposes of one’s study, could qualify as noise. For instance, a study by Kloumann, Danforth, Harris, Bliss, and Dodds (2012) reported affective valence ratings for the 10,000 most frequent entries attested in four corpora. Their list included items that are unlikely to produce informative affective ratings, such as spelling variants (bday, b-day, and birthday), words with special characters (#music, #tcot), foreign words not borrowed into English (cf. the Dutch words “hij” [he] and “zijn” [to be]), alphanumeric strings (a3 and #p2), and names of people, cities, and countries. The list also included inflected word forms, which is a useful design option only if one expects inflected forms to differ in rating from lemmas (e.g., runs vs. run). When we compared Kloumann et al.’slist to a large list of English lemmas (see below), only half of the stimuli overlapped (see also Warriner, Kuperman, & Brysbaert, in press). This is a serious loss of investment, which is likely to further increase for less frequent entries (where the signal-to-noise ratio is even smaller).

To tackle the problem head on, we collected concreteness ratings for a list of 63,039 English lemmas one of us (M.B.) has been assembling over the years. This list does not contain proper names or inflected forms. The latter are more difficult to define in English than would be assumed at first sight, because many inflected verb forms are homonymous (and derivationally related) to uninflected adjectives (appalling) or nouns (playing). The simplest criterion to disambiguate such cases is to verify whether the word is used more often as an adjective/noun than as a verb form. This has become possible since we collected part-of-speech-dependent word frequency measures for American English (Brysbaert, New, & Keuleers, 2012). Similarly, some nouns are used more frequently in plural form than in singular form (e.g., eyes)or have different meanings in singular and plural (glasses , aliens). For these words, both forms were included in the list. Finally, the list for the first time also includes frequently encountered two-word spaced compound nouns (eye drops, insect repellent, lawn mower) and phrasal verbs (give away, give in, give up). The latter were based on unpublished analyses of the SUBTLEX-US corpus (Brysbaert & New, 2009). By presenting the full list, we were able to see which words are known to the majority of English speakers independently of word frequency. One way often used to select words for megastudies is to limit the words to those with frequencies larger than one occurrence per million words (e.g., Ferrand et al., 2010; Keuleers, Diependaele, & Brysbaert, 2010). This is a reasonable criterion but may exclude generally known words with low frequencies, which arguably are the most interesting to study the limitations of the existing word frequency measures.

In summary, we ran a new concreteness rating study (1) to obtain concreteness ratings for a much larger sample of English words, (2) to obtain ratings based on all types of experiences, and (3) to define a reference list of English lemmas for future studies.

Method

Materials

The stimuli consisted of a list of 60,099 English words and 2, 940 two-word expressions. The list was built on the basis of the SUBTLEX-US corpus (Brysbaert & New, 2009), supplemented with words from the English Lexicon Project (Balota et al., 2007), the British Lexicon Project (Keuleers et al., 2012; if necessary, spellings were Americanized), the corpus of contemporary American English (Davies, 2009), words used in various rating studies and shop catalogs, and words encountered throughout general reading. Although it is unavoidable that the list missed a few widely known words, care was taken to include as many entries as we could find.[2]

[2] Indeed, M.B. would appreciate receiving suggestions of missing words that should have been included.

Data collection

The stimuli were distributed over 210 lists of 300 words. Each list additionally included 10 calibrator words and 29 control words. The calibrator words represented the entire concreteness range (based on the MRC ratings) to introduce the participants to the variety of stimuli they could encounter. These words were placed in the beginning of each list. They were shirt, infinity, gas, grasshopper, marriage, kick, polite, whistle, theory,and sugar. Care was taken to include words referring to nonvisual senses and actions. The control words were from the entire concreteness range as well, used to detect noncompliance with the instructions (see below). Like the calibrator words, the same set of control words were used in all lists, to make sure that we used fixed criteria throughout. Control words were scattered randomly throughout the lists.

As in our previous studies (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012; Warriner et al., in press) participants were recruited via Amazon Mechanical Turk's crowdsourcing Web site. Responders were restricted to those who self-identified as current residents of the U.S. The completion of a single list by a given participant is referred to as an assignment, given that participants were allowed to rate more than one list.

The following instructions were used:

> *Some words refer to things or actions in reality, which you can experience directly through one of the five senses. We call these words concrete words. Other words refer to meanings that cannot be experienced directly but which we know because the meanings can be defined by other words. These are abstract words. Still other words fall in-between the two extremes, because we can experience them to some extent and in addition we rely on language to understand them. We want you to indicate how concrete the meaning of each word is for you by using a 5-point rating scale going from abstract to concrete.*
>
> *A concrete word comes with a higher rating and refers to something that exists in reality; you can have immediate experience of it through your senses (smelling, tasting, touching, hearing, seeing) and the actions you do. The easiest way to explain a word is by pointing to it or by demonstrating it (e .g. To explain 'sweet' you could have someone eat sugar; To explain 'jump' you could simply jump up and down or show people a movie clip about someone jumping up and down; To explain 'couch ', you could point to a couch or show a picture of a couch).*

An abstract word comes with a lower rating and refers to something you cannot experience directly through your senses or actions. Its meaning depends on language. The easiest way to explain it is by using other words (e.g. There is no simple way to demonstrate 'justice'; but we can explain the meaning of the word by using other words that capture parts of its meaning).

Because we are collecting values for all the words in a dictionary (over 60 thousand in total), you will see that there are various types of words, even single letters. Always think of how concrete (experience based) the meaning of the word is to you. In all likelihood, you will encounter several words you do not know well enough to give a useful rating. This is informative to us too, as in our research we only want to use words known to people. We may also include one or two fake words which cannot be known by you. Please indicate when you don't know a word by using the letter N (or n).

So, we ask you to use a 5-point rating scale going from abstract to concrete and to use the letter N when you do not know the word well enough to give an answer.

Abstract (language based) Concrete (experience based)

1 2 3 4 5

N = I do not know this word well enough to give a rating.

In the instructions, we stressed that we made a distinction between experience-based meaning acquisition and language-based meaning acquisition (cf. Della Rosa et al., 2010) and that experiences must not be limited to the visual modality. We used a 5-point rating scale based on Laming’s(2004) observation that 5 is the maximum number of categories humans can distinguish consistently. When people are asked to make finer distinctions, they start using the labels inconsistently to such an extent that no extra information is obtained and the further scale precision is illusionary. In addition, we did not want to overtax the participants’ working memory, because they had to keep in mind to use the N alternative in case they did not know the word well enough to give a valid rating.

On average, assignments were completed in approximately 14 min. Participants received 75 U.S. cents per completed assignment. After reading a consent form 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 an answer box. Once finished, participants clicked 'Submit' to complete the study.

We aimed at 30 respondents per list. However, missing values due to subsequent exclusion criteria resulted in some words having less than 20 valid ratings. Several of the lists were reposted until the vast majority of words reached at least 25 observations per word. Data collection began January 25, 2013 and was completed by mid April 2013.

Results

Data trimming

Altogether 2,385,204 ratings were collected. Around 4 % 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 fewer than 100 ratings per assignment. Further cleaning involved lists for which the correlation with the MRC ratings of the control words was between − .5 and .2. (The ones with correlations below − .5 were assumed to come from participants who misunderstood the instructions and used the opposite ordering; these scores were converted. This was the case for 149 assignments or 2.5 % of the total number.) Nonnative English speakers were also removed. Finally, assignments for which the correlation across the entire list was less than .1 with the average of the other raters were removed. Of the remaining data, 1,676,763 were numeric ratings, and 319,885 were "word not known" responses. These data came from 4,237 workers completing 6,076 assignments. There were more valid data from female participants (57 %) than from male participants. Of the participants, 1,542 (36 %) were of the typical student age (17–25 years old) and 40 (1 %) were older than 65 years. The remainder came from the ages in between these two groups. The distribution across educational levels is shown in Table 1.

Table 1: Distribution of the education levels of the valid respondents.

Education Level	Number
Some high school	46
High school graduate	355
Some college – no degree	1,438
Associates degree	442
Bachelors degree	1,389
Masters degree	422
Doctorate	112
Education level not specified	33

Final List

Because ratings are only useful for well known words, we used a cutoff score of 85 % known. In practice, this meant that not more than 4 participants out of the average of 25 raters indicated that they did not know the word well enough to rate it. This left us with a list of 37,058 words and 2,896 two-word expressions (i.e., a total of 39,954 stimuli).

Validation

The simplest way to validate our concreteness ratings is to correlate them with the concreteness ratings provided in the MRC database (Coltheart, 1981). Therewere3,935 overlapping words (the nonoverlapping words were mostly words not known to a substantial percentage of participants in our study, inflected forms, and words differing in spelling between British and American

English). The correlation between both measures was r =.919, which is surprising given that our instructions emphasized—to a larger extent than the MRC instructions—the importance of action-related experiences. Also, when we look at the stimuli with the largest residuals between MRC and our ratings (Table 2), we see that they are more understandable as the outcome of different interpretations of ambiguous words than as differences between perception and action.

To further understand the essence of our ratings, we correlated them with the perceptual strength ratings collected by Lynott and Connell (2009, in press; downloaded on May 1, 2013 from http://personalpages.manchester.ac.uk/staff/louise. connell/lab/norms.html). As was indicated in the Introduction, these authors asked their participants to indicate how strongly they had experienced the stimuli with their auditory, gustatory, haptic, olfactory, and visual senses. Lynott and Connell also calculated the maximum perceptual strength of a stimulus, defined as the maximum value of the previous five ratings. Of the 1,001 words for which ratings were available, 615 had concreteness ratings in the MRC database and in our database. The correlation between the two concreteness ratings was very similar to that of the complete database (r =.898, N =615). Table 3 shows the correlations with the perceptual strength ratings. Again, it is clear that our concreteness ratings provide very much the same information as the MRC concreteness ratings, despite the differences in instructions. In particular, both concreteness ratings correlate best with haptic and visual strength and show a negative correlation with auditory strength.

Table 2: Differences between the MRC ratings and the present ratings of concreteness: The 20 words with the largest negative and positive residuals

Words much lower in our ratings			Words much higher in our ratings		
WORD	MRC	OUR	WRD	MRC	OUR
concern	509	1.70	site	408	4.56
general	408	1.62	on	262	3.25
originator	491	2.52	stop	308	3.68
outsider	468	2.33	grate	432	4.82
patient	487	2.50	flow	311	3.72
chic	454	2.26	lighter	400	4.53
master	498	2.63	himself	285	3.50
conspirator	464	2.37	pour	356	4.14
dreamer	442	2.19	devil	274	3.41
gig	525	2.89	sear	292	3.59
ally	485	2.61	their	257	3.34
gloom	399	1.86	precipitate	350	4.19
mortal	406	1.96	facility	279	3.58
religion	375	1.71	dozen	396	4.66
equality	342	1.41	month	345	4.20
buffer	509	2.89	can	365	4.55
connoisseur	483	2.70	drop	320	4.21
evaluate	388	1.85	logos	299	4.41
forelock	565	3.28	tush	287	4.45
earl	500	2.85	concert	252	4.35

Table 3 Correlations between concreteness ratings and the perceptual strength ratings collected by Connell and Lynott (2009, in press)

	Concreteness_Our	Concreteness_MRC
Auditory strength	-.259**	-.234**
Gustatory strength	.023	.054
Haptic strength	.410**	.364**
Olfactory strength	.187**	.243**
Visual strength	.449**	.399**
Maximal strength	.495**	.440**

** p < .001

Discussion

Recent technological advances have made it possible to collect valid word ratings at a much faster pace than in the past. In particular, the availability of Amazon Mechanical Turk (AMT) and the kindness of Internet surfers in providing good scientific data at an affordable price have made it possible to collect ratings for tens of thousands of words, rather than hundreds of words. In the present article, we discuss the collection of concreteness ratings for about 40,000 generally known English lemmas.

The high correlation between our ratings and those included in the MRC database (r =.92) attests to both the reliability and the validity of our ratings (for similar findings with AMT vs. lab-collected ratings, see also Kuperman et al., 2012; Warriner et al., in press). At the same time, the high correlation shows that the extra instructions we gave for the inclusion of nonvisual and action-related experiences did not seem to have much impact. Gustatory strength was not taken into account and auditory strength even correlated negatively, because words such as deafening and noisy got low concreteness ratings (1.41 and 1.69, respectively) but high auditory strength ratings (5.00 and 4.95). Apparently, raters cannot take into account several senses at the same time (Connell & Lynott, 2012).

The fact that our concreteness ratings are very similar to the existing norms (albeit for a much larger and more systematically collected stimulus sample) means that other criticisms recently raised against the ratings apply to our data set as well.[3] One concern, for instance, is that concreteness and abstractness may be not the two extremes of a quantitative continuum (reflecting the degree of sensory involvement, the degree to which words meanings are experience based, or the degree of contextual availability), but two qualitatively different characteristics. One argument for this view is that the distribution of concreteness ratings is bimodal, with separate peaks for concrete and abstract words, whereas ratings on a single, quantitative dimension usually are unimodal, with the majority of observations in the middle (Della Rosa et al., 2010; Ghio, Vaghi, & Tettamanti, 2013). As Fig. 1 shows, the bimodality of the distribution is true even for the large data set we collected, although it seems to be less extreme than reported by Della Rosa et al. Other arguments for qualitative differences between abstract and concrete concepts are that they can be affected differently by brain injury and that their representations may be organized in different ways (Crutch & Warrington, 2005; Duñabeitia, Avilés, Afonso, Scheepers, & Carreiras, 2009).

A further criticism raised against concreteness ratings is that concrete and abstract may not be basic level categories but superordinate categories (or maybe even ad hoc categories; Barsalou, 1983), which encompass psychologically more important subclasses, such as fruits , vegetables, animals,and furniture for concrete concepts and mental -state -related , emotion -related ,and mathematics -related notions for abstract concepts (Ghio et al., 2013). If true, this criticism implies that not much information can be gained from concreteness information and that more fine-grained information is needed about the basic level categories (also Mahon & Caramazza, 2011).

[3] We thank an anonymous reviewer for pointing us to this literature.

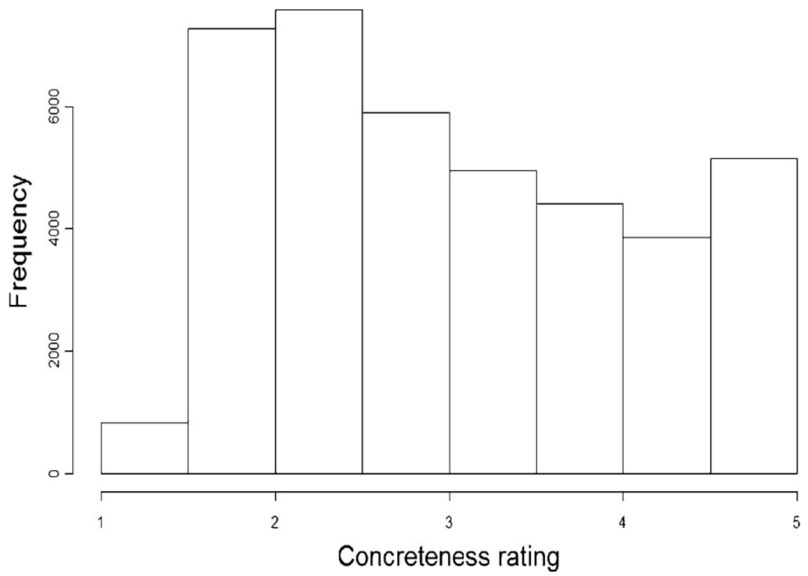


Figure 1. Distribution of the concreteness ratings (N = 39,954): 1 = very abstract (language-based), 5 = very concrete (experience-based)

The above criticisms perfectly illustrate that each study involves choices and, therefore, is limited in scope. What we won on the one hand (information about a variable for the entire set of interesting English lemmas) has been achieved at the expense of information richness on the other hand. This can be contrasted with the approach taken by Della Rosa et al. (2010), Ghio et al. (2013), Rubin (1980), and Clark and Paivio (2004), among others, who collected information about a multitude of word features, so that the correspondences between the measures could be determined. This, however, was achieved at the expense of the number of items for which information could be collected.

It is clear that our study cannot address all questions raised about concreteness norms. However, it provides researchers with values of an existing, much researched variable for an exhaustive word sample. More focused research is needed to further delineate the uses and limitations of the variable. For instance, it can be wondered how the low concreteness rating of myth (2.17) relates to the high perceptual strength rating of the same word (4.06, coming from auditory strength)[4] and what the best value is for atom, given that the concreteness rating (3.34) is much higher than the perceptual strength rating (1.37). Similarly, it may be asked what the much lower concreteness rating of loving (1.73) than of sailing (4.17) means, given that many more participants are likely to have experienced the former than the latter (remember that we defined concrete as "experience-based" and abstract as "language-based"). These examples remind us that collecting a lot of information about a variable does not by itself make the variable more "real." It only allows us to study the variable in more detail.

Next to concreteness information, the research described in this article provides us with a reference list of English lemmas for future word recognition research. To achieve this, we presented a rather exhaustive list of lemmas to our participants, so that we made no a priori selection. On the basis of our findings, we can conclude that such a big list contains about one third of words not known to enough native speakers to warrant further inclusion in rating studies (to be fair to our participants, many of these stimuli referred to little known animals and plants). For future research, it seems more efficient to focus on the 40,000 generally known words than to continue including words that will have to be discarded afterward. At the same time, our research shows that some of the well-known words have low frequencies, as measured nowadays. These obviously include all two-word expressions (which are absent in most word frequency lists), but also compound words that were concatenated in our list because this is how they were used in the study we took them from (such as birdbath and birdseed from ELP) but that, in normal text, are usually written separately. Further well-known words with low frequencies are derivations of familiar words (such as bloodlessness, borrowable, and brutalization)and, more intriguingly, some words referring to familiar objects (such as canola, lollypop, nectarine, nightshirt, thimble, wineglass , and bandanna). By focusing on these stimuli,

[4] Arguably because people hear about myths.

we can better understand the limitations of current-day word frequency measures. An interesting conceptual framework in this respect may be found in the papers of Vigliocco and colleagues (Andrews et al., 2009; Kousta et al., 2011; Vigliocco et al., 2004). Apparently, some words are well known to us because we daily experience the objects they refer to, but we rarely communicate about them, making them rather obscure in language corpora. Our database for the first time allows us to zoom in on these stimuli.

Availability

The data discussed in the present article are available in an Excel file, provided as supplementary materials. The file contains eight columns:

1 The word

2 Whether it is a single word or a two-word expression

3 The mean concreteness rating

4 The standard deviation of the concreteness ratings

5 The number of persons indicating they did not know the word

6 The total number of persons who rated the word

7 Percentage participants who knew the word

8 The SUBTLEX-US frequency count (on a total of 51 million; Brysbaert & New, 2009)

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.

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. Psychological Review, 116, 463–498.

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.

Barber, H. A., Otten, L. J., Kousta, S. T., & Vigliocco, G. (2013). Concreteness in word processing: ERP and behavioral effects in a lexical decision task. Brain and Language, 125(1), 47–53.

Barsalou, L. W. (1983). Ad hoc categories. Memory & Cognition, 11, 211–227.

Barsalou, L. W. (1999). Perceptual symbol systems. Behavioral and Brain Sciences, 22, 577–660.

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.

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding Part-of-Speech information to the SUBTLEX-US word frequencies. Behavior Research Methods, 44, 991–997.

Clark, J. M., & Paivio, A. (2004). Extensions of the Paivio, Yuille, and Madigan (1968) norms. Behavior Research Methods, Instruments & Computers, 36(3), 371–383.

Coltheart, M. (1981). The MRC psycholinguistic database. The Quarterly Journal of

Experimental Psychology, 33, 497–505.

Connell, L., & Lynott, D. (2012). Strength of perceptual experience predicts word processing performance better than concreteness or imageability. Cognition, 125, 452–465.

Crutch, S. J., & Warrington, E. K. (2005). Abstract and concrete concepts have structurally different representational frameworks. Brain, 128(3), 615–627.

Davies, M. (2009). The 385+ million word corpus of contemporary American English: Design, architecture, and linguistic insights. International Journal of Corpus Linguistics, 14(2), 159–190.

Della Rosa, P. A., Catricalà, E., Vigliocco, G., & Cappa, S. F. (2010). Beyond the abstract—concrete dichotomy: Mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 Italian words. Behavior Research Methods, 42, 1042–1048.

Duñabeitia, J. A., Avilés, A., Afonso, O., Scheepers, C., & Carreiras, M. (2009). Qualitative differences in the representation of abstract versus concrete words: evidence from the visual-world paradigm. Cognition, 110(2), 284–292.

Ferré, P., Guasch, M., Moldovan, C., & Sánchez-Casas, R. (2012). Affective norms for 380 Spanish words belonging to three different semantic categories. Behavior Research Methods, 44, 395–403.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., ... 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.

Fischer, M. H., & Zwaan, R. A. (2008). Embodied language: A review of the role of the motor system in language comprehension. The Quarterly Journal of Experimental Psychology, 61, 825–850.

Gianico-Relyea, J. L., & Altarriba, J. (2012). Word Concreteness as a Moderator of the Tip-of the-Tongue Effect. Psychological Record, 62, 763–776.

Ghio, M., Vaghi, M. M. S., & Tettamanti, M. (2013). Fine-Grained Semantic Categorization across the Abstract and Concrete Domains. PLoS ONE, 8(6), e67090. doi:10.1371/journal.pone. 0067090

Hanley, J. R., Hunt, R. P., Steed, D. A., & Jackman, S. (2013). Concreteness and word production. Memory & Cognition, 41, 365–377.

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. Neuron, 41, 301–307.

Kaushanskaya, M., & Rechtzigel, K. (2012). Concreteness effects in bilingual and monolingual word learning. Psychonomic Bulletin & Review, 19, 935–941.

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono-and disyllabic words and nonwords. Frontiers in Psychology, 1, 174. doi:10.3389/fpsyg.2010.00174

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.

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

Laming, D. (2004). Human judgement: The eye of the beholder. London: Thompson Learning.

Loiselle, M., Rouleau, I., Nguyen, D. K., Dubeau, F., Macoir, J., Whatmough, C., & Joubert, S. (2012). Comprehension of concrete and abstract words in patients with selective anterior temporal lobe resection and in patients with selective amygdalo-hippocampectomy. Neuropsychologia, 50, 630–639.

Lynott, D., & Connell, L. (2009). Modality exclusivity norms for 423 object properties. Behavior Research Methods, 41, 558–564.

Lynott, D., & Connell, L. (in press). Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. Behavior Research Methods.

Mahon, B. Z., & Caramazza, A. (2011). What drives the organization of object knowledge in the brain? Trends in Cognitive Sciences, 15, 97–103.

Mate, J., Allen, R. J., & Baqués, J. (2012). What you say matters: Exploring visual–verbal interactions in visual working memory. The Quarterly Journal of Experimental Psychology, 65, 395–400.

Nishiyama, R. (2013). Dissociative contributions of semantic and lexical-phonological information to immediate recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition, 39, 642–648.

Oliveira, J., Perea, M. V., Ladera, V., & Gamito, P. (2013). The roles of word concreteness and cognitive load on interhemispheric processes of recognition. Laterality, 18, 203–215.

Paivio, A. (1971). Imagery and verbal processes. New York: Holt, Rinchart, and Winston.

Paivio, A. (2013). Dual Coding Theory, Word Abstractness, and Emotion: A Critical Review of Kousta et al. (2011). Journal of Experimental Psychology: General, 142, 282–287.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. Journal of experimental psychology, 76, 1–25.

Rubin, D. C. (1980). 51 properties of 125 words: A unit analysis of verbal behavior. Journal of Verbal Learning and Verbal Behavior, 19, 736– 755.

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

CHAPTER 4: Emotion and language: Valence and arousal affect word recognition

Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A.B. (2014).

Journal of Experimental Psychology: General. 143(3), 1065-1081. doi: 10.1037/a0035669

Abstract

Emotion influences most aspects of cognition and behavior, but emotional factors are conspicuously absent from current models of word recognition. The influence of emotion on word recognition has mostly been reported in prior studies on the automatic vigilance for negative stimuli, but the precise nature of this relationship is unclear. Various models of automatic vigilance have claimed that the effect of valence on response times is categorical, an inverted-U, or interactive with arousal. The present study used a sample of 12,658 words, and included many lexical and semantic control factors, to determine the precise nature of the effects of arousal and valence on word recognition. Converging empirical patterns observed in word-level and trial-level data from lexical decision and naming indicate that valence and arousal exert independent monotonic effects: Negative words are recognized more slowly than positive words, and arousing words are recognized more slowly than calming words. Valence explained about 2% of the variance in word recognition latencies, whereas the effect of arousal was smaller. Valence and arousal do not interact, but both interact with word frequency, such that valence and arousal exert larger effects among low-frequency words than among high-frequency words. These results necessitate a new model of affective word processing whereby the degree of negativity monotonically and independently predicts the speed of responding. This research also demonstrates that incorporating emotional factors, especially valence, improves the performance of models of word recognition.

Keywords: arousal, automatic vigilance, emotion, lexical decision and naming, valence, word recognition.

Introduction

Emotion influences most aspects of cognition and behavior, from visual attention (Rowe, Hirsh, & Anderson, 2007) to social comparison (Estes, Jones, & Golonka, 2012). It affects how we see the world, what we think, and with whom we associate (Forgas, 1995; van Kleef, 2009). Emotions are typically characterized along two primary dimensions of arousal and valence (Russell, 2003; Russell & Barrett, 1999), which correspond respectively to Osgood and colleagues' (Osgood, Suci, & Tannenbaum, 1957) semantic factors of *activity* and *evaluation*. *Arousal* is the extent to which a stimulus is calming or exciting, whereas *valence* is the extent to which a stimulus is negative or positive. These two dimensions are theoretically orthogonal: Negative stimuli can be either calming (e.g., *dirt*) or exciting (e.g., *snake*), and positive stimuli can also be calming (e.g., *sleep*) or exciting (e.g., *sex*). Arousal and valence are also neurologically dissociable, activating distinct cortical networks (Kensinger & Corkin, 2004; LaBar & Cabeza, 2006).

The present research investigates effects of arousal and valence on word recognition. Word recognition has received considerable research attention over the last few decades, and despite a number of important theoretical advances (see Adelman, 2012), a great deal of the variance in word recognition times still remains unexplained (Adelman, Marquis, Sabatos-DeVito, & Estes, 2013). Notably, the current models incorporate a broad range of lexical factors such as word frequency (Brysbaert & New, 2009) and contextual diversity (Adelman, Brown, & Quesada, 2006), but emotional factors are conspicuously absent. So given the influence of emotion on cognition, and the lack of emotional factors in current models of word recognition, the present study examined the influence of emotion on word recognition.

Effects of emotion on word recognition

Many experiments over decades of research suggested that negative stimuli elicit slower responses than neutral stimuli on a range of cognitive tasks. For instance, negative words such as *coffin* tend to evoke slower color naming in the emotional Stroop task (for a review, see Williams, Mathews, & MacLeod, 1996), slower lexical decisions (e.g., Wentura, Rothermund, & Bak, 2000), and slower word naming (a.k.a., reading aloud; e.g., Algom, Chajut, & Lev, 2004) than neutral words such as *cotton*. This observation was attributed to a process of *automatic vigilance*, whereby humans preferentially attend to negative stimuli (Erdelyi, 1974; Pratto & John, 1991). According to this automatic vigilance hypothesis, negative stimuli engage attention longer than other stimuli (Fox, Russo, Bowles, & Dutton, 2001; Ohman & Mineka, 2001), and hence negative stimuli elicit slower responses than other stimuli. The automatic vigilance hypothesis thus assumes that emotion affects the decisional or response stage of word processing: The delayed response to negative words arises during the lexical decision or naming process, rather than during the activation of lexical or semantic representations. Alternatively, emotion could affect the activation of those lexico-semantic representations (Yap & Seow, 2013). That is, activation of negative representations may be "repressed" (Erdelyi, 1974) and/or positive representations may be activated particularly quickly. In fact, Yap and Seow recently reported evidence that valence affects both early and late stages of the word recognition process.

Those decades of experimental results, however, are critically undermined by a lack of stimulus controls (Larsen, Mercer, & Balota, 2006). Larsen et al. conducted a meta-analysis of 1033 stimulus words that were used in 32 published studies on the emotional Stroop task (i.e., color naming of emotional and neutral words). They found that the negative words used in those prior studies tended to be longer and less frequent than the neutral words (see also Warriner, Kuperman & Brysbaert, 2013). These lexical confounds, both of which are known to slow down word recognition (e.g., Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004), could parsimoniously explain the effects observed in those prior studies. And indeed, Larsen et al. found that after controlling those spurious lexical confounds, negative words no longer elicited slower responses than neutral words. Thus, the entire literature on automatic vigilance was rendered equivocal. Since Larsen et al.'s (2006) critical observation, several more recent and better controlled studies have examined the effect of emotion on word recognition, but unfortunately those studies have yielded differing conclusions.

Recent Controlled Studies. Estes and Adelman (2008a) examined the influence of valence on lexical decision and naming latencies, while controlling other important emotional and lexical factors (see Table 1). They found that arousal significantly predicted word recognition: Exciting words tended to be recognized faster than calming words. Their analyses additionally showed that, even after statistically accounting for arousal and several lexical factors, valence still explained significant variance in lexical decision and naming times. Negative words tend to be recognized more slowly than positive words. In contrast to a linear effect whereby increasingly negative and increasingly positive words elicit increasingly slow and fast response times (RTs) respectively, Estes and Adelman found that the effect of valence on word recognition times was nonlinear. Extremely negative words were recognized no slower than moderately negative words, and extremely positive words were recognized no faster than moderately positive words. This produced a step-function whereby RTs remained constant and slow across the category of negative words, decreased sharply through the neutral region of the valence scale, and then remained constant and fast across the category of positive words.

Table 1: Regression studies of the influence of emotion on word recognition latencies. EA (2008a) = Estes and Adelman (2008a); Kousta = Kousta et al. (2009); Lex Dec = lexical decision.

	EA (2008a)		Larsen et al. (2008)		Kousta
N	1011		1021		1446
	Lex Dec	Naming	Lex Dec	Naming	Lex Dec
Emotional Factors					
Arousal	***	**	*ns*	*ns*	*ns*
Valence	***	***	***	***	*
Arousal × Valence			***	*ns*	
Control Factors					
Letters	***	***	***	***	***
Syllables	*ns*	*			
Morphemes					*ns*
Frequency	**	*ns*	***	***	***
Familiarity					***
Contextual diversity	*ns*	*			
Orthographic N	*ns*	*ns*	*ns*	*ns*	***
Initial Phoneme		***			
Imageability					*ns*
Age of Acquisition					***
Bigram frequency					*ns*
Best R^2	53.24%	52.58%	58.70%	40.00%	64.55%

Note: *ns* = nonsignificant; * $p < .05$; ** $p < .01$; *** $p < .001$

Whereas Estes and Adelman (2008a) tested for independent effects of valence and arousal, Larsen, Mercer, Balota, and Strube (2008) examined whether arousal and valence have an interactive effect on word recognition. They replicated Estes and Adelman's analyses of lexical decision and naming times (except with different control factors, see Table 1), and additionally included the possible interaction between arousal and valence. Larsen et al. found a significant interaction between arousal and valence in lexical decisions (but not in naming), such that low arousal tends to slow down lexical decisions to negative words but speeds up lexical decisions to positive words (see also Robinson, Storbeck, Meier, & Kirkeby, 2004). Highly arousing words, in contrast, exhibited little or no effect of valence. Estes and Adelman (2008b) subsequently demonstrated, however, that Larsen et al.'s reported interaction of valence and arousal depended critically on the underlying form assumed for valence. When valence was entered into the regression model as a linear continuous predictor, then it interacted with arousal in predicting RTs (as in Larsen et al., 2008). However, when valence was entered into the model as a categorical predictor (as previously observed by Estes and Adelman, 2008a), the interaction reported by Larsen et al. disappeared and negative words elicited slower lexical decisions than positive words regardless of their arousal (i.e., an effect of valence was also observed among highly arousing words).

A limitation of the studies by Estes and Adelman (2008a) and Larsen et al. (2008) was their use of the Affective Norms for English Words (ANEW; Bradley & Lang, 1999) as the sole source of stimuli. ANEW is useful for sampling a limited number of emotional words, but because the words in ANEW were primarily selected for their emotionality, ANEW lacks the preponderance of emotionally neutral words that is typical of natural languages (Kousta, Vinson, & Vigliocco, 2009). Kousta et al. thus merged ANEW with an additional set of randomly selected words, producing a total of 1446 words, including more neutral words than the prior studies. They also employed more sophisticated regression methods for detecting nonlinear relationships. Unlike Estes and Adelman, Kousta et al. found no effect of arousal on lexical decision latencies when controlling for valence. Critically, they also found that after controlling for several other lexical, semantic, and emotional factors (see Table 1), negative and positive words *both* elicited faster lexical decisions than neutral words, and the difference between negative and positive words was nonsignificant. That is, Kousta et al. found a nonlinear, inverted-U effect of valence on lexical decision times. They did not test for an interaction between arousal and valence. These findings based on the large-scale behavioral data set collected in US universities and available from the English Lexicon Project (*ELP*; Balota et al., 2007) have recently been replicated by Vinson, Ponari, and Vigliocco (2013) with the British Lexicon Project (Keuleers, Lacey, Rastle, & Brysbaert, 2012), a mega-study that reported lexical decision latencies to over 28,000 words collected at UK universities. Vinson et al. (2013) observed an inverted-U effect of valence on lexical decision times, and they found no evidence of a valence × arousal interaction.

Emotion × Frequency Interactions. Word frequency is among the most important factors of word recognition. To begin with, in most studies it explains a relatively large amount of the variance in word recognition latencies and accuracies (Balota et al., 2004; Brysbaert & New, 2009; Yap & Balota, 2009): Frequent words are recognized more quickly and accurately than infrequent words. More critically for the present study, frequency also tends to modulate the effects of other factors on word recognition. For instance, although both imageability and age of acquisition influence word recognition (Balota et al., 2004; Brysbaert & Cortese, 2011; Cortese & Khanna, 2007; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), both of those effects are significantly larger among low frequency words than among high frequency words (e.g., Cortese & Schock, 2013; Gerhand & Barry, 1999a, 1999b). Two plausible explanations of such interactions with frequency can be differentiated. One general explanation is purely statistical and relies on a base-rate effect, namely, that the magnitude of word recognition latencies is positively correlated with the magnitude of lexical effects on the speed of word recognition. The same relative effect size (say, a 25% difference in RTs between words of high and low imageability) leads to a larger absolute effect in words with longer mean latencies (e.g., 150 ms in words with a mean RT of 600 ms) than in words with shorter mean latencies (e.g., 100 ms in words with a mean RT of 400 ms). Since lower-frequency words take longer to recognize, all lexical effects may appear larger in those words (Butler & Hains, 1979; Faust, Balota, Spieler, & Ferraro, 1999; Kuperman & Van Dyke, 2013; Yap, Balota, Sibley, & Ratcliff, 2012). A second explanation is that because low frequency words take longer to recognize, there is more time for higher-level semantic factors (e.g., imageability) to affect responding. In contrast, because high frequency words are recognized relatively quickly, semantic factors exert little or no effect on word recognition (Cortese & Schock, 2013). Thus, the former explanation is purely mathematical, whereas the latter is cognitive.

Word frequency also appears to modulate emotional effects on word recognition, but the nature of this modulation is currently unclear. In the emotional Stroop task, valence influenced responses to low frequency words, such that negative words elicited slower color naming than positive words. Among high frequency words, however, valence had no effect (Kahan & Hely, 2008). This finding is analogous to the results described above, in that high frequency tends to reduce or eliminate effects of other factors (e.g., imageability, age of acquisition, valence). In lexical decisions, however, some evidence suggests an opposite effect. Scott, O'Donnell, Leuthold, and Sereno (2009) reported that among low frequency words, negative and positive words elicited equally slow responses, but that among high frequency words, negative words elicited slower lexical decisions than positive words. Further research with eye movements during sentence reading confirmed the interaction of valence and frequency. Whereas fixation durations did not differ between low frequency words of negative and positive valence, fixations on high frequency words were significantly longer for negative words than for positive words (Scott, O'Donnell, & Sereno, 2012). Furthermore, Sheikh and Titone (in press) observed speed benefits to both positive and negative words, as compared to neutral ones, but only when words were of low frequency and relatively concrete. Thus, despite empirical ambiguity in the direction of the effect, it is now clear that word frequency often modulates emotional effects on word recognition. Unfortunately, none of the recent controlled studies of emotional effects on word recognition (i.e., Estes & Adelman, 2008a, 2008b; Kousta et al., 2009; Larsen et al., 2008) controlled or tested for interactions with word frequency.

The Present Study

Empirical Contribution. Prior studies have demonstrated that emotion influences word recognition, but the precise nature of this relationship is unclear. Two main theoretical issues have arisen. First, there is disagreement about the functional form of the effect of valence on word recognition. Specifically, it is unclear whether the effect of valence on word recognition is monotonic but has a step-function form (Estes & Adelman, 2008a, 2008b), monotonic with a linear form (Larsen et al., 2008) or nonmonotonic with an inverted-U form (Kousta et al., 2009; Vinson et al., 2013). Second, it is unclear whether arousal and valence have independent effects on word recognition. Some researchers have found that both arousal and valence influence word recognition (Estes & Adelman, 2008a), whereas others have found effects of valence but not arousal (Kousta et al., 2009). Moreover, some have found that arousal and valence have an interactive effect on word recognition (Larsen et al., 2008), but others have argued against the validity of such an interaction (Estes & Adelman, 2008b; Vinson et al., 2013). Thus, the present study compared statistical models that varied in whether they treated arousal and valence as linear or nonlinear and independent or interactive.

The prior studies have also exhibited some potentially critical empirical limitations. To begin with, although they included substantially larger samples of stimuli than the pre-2006 experiments in this area of research, those regression studies each sampled little more than a thousand words (see Table 1). By the current standards of research on word recognition (e.g., Brysbaert & New, 2009; Yarkoni, Balota, & Yap, 2008; for review see Adelman, 2012), those are small samples. Moreover, although Kousta et al. (2009) and Vinson et al. (2013) added a few hundred neutral words, the prior studies nonetheless contain a paucity of neutral words, thus undermining their representativeness. Furthermore, the various studies have included different sets of control factors (see Table 1), making it difficult to compare results from one study to the next. For instance, the discrepant results between Estes and Adelman (2008a) and Kousta et al. could simply be due to the fact that Estes and Adelman did not control for age-of-acquisition, or that Kousta et al. did not control contextual diversity. Perhaps most importantly, those prior studies did not test for emotion × frequency interactions, which are known to occur in word recognition (Kahan & Hely, 2008; Scott et al., 2009, 2012; Sheikh & Titone, in press). The present study aimed to address these limitations by (1) sampling a substantially larger set of words that (2) were not sampled for their emotionality and thus are more representative of natural language, (3) including many more lexical and semantic control factors than any prior study, and (4) testing for interactions of valence and arousal with word frequency.

Thus, the present study used a sample (12,658 words) from the dataset of affective norms (psychological valence, arousal and dominance) collected by Warriner et al. (2013), which is about 9-13 times larger than the prior studies. The analyses also included about twice as many lexical and semantic control factors, to critically test multiple models of emotional word recognition. Our study has been made possible by the recent emergence of psycholinguistic mega-studies (for review see Adelman, 2012; Balota, Yap, Hutchinson, & Cortese, 2012), whereby massive datasets compiling the lexical (Brysbaert & New, 2009; Kuperman et al., 2012), semantic (Brysbaert, New, & Keuleers, 2012), and emotional characteristics (Warriner et al., 2013) of many thousands of words can be merged with behavioral data such as lexical decision and naming latencies and accuracies for those same words, namely the English Lexicon Project (Balota et al., 2007). The present research employs this mega-study approach to determine the precise nature of the effects of arousal and valence on lexical decision and naming latencies.

Theoretical Contribution. We anticipate two important theoretical contributions from this research. First, this research is most informative for models of automatic vigilance. The effect of valence on word recognition times has been a primary source of evidence for automatic vigilance (Algom et al., 2004; Estes & Adelman, 2008a, 2008b; Larsen et al., 2008; Pratto & John, 1991; Wentura et al., 2000; Williams et al., 1996), but several different relationships have been hypothesized. The simplest model of automatic vigilance supposes that humans immediately judge stimuli as either aversive (i.e., negative stimuli to be avoided) or appetitive (i.e., positive stimuli to be approached) in a binary, categorical manner (Estes & Adelman, 2008a, 2008b). Such simple evaluative judgments would be behaviorally adaptive in that they would facilitate rapid decisions and actions. Deliberating about whether a stimulus is extremely dangerous or only moderately dangerous could in fact be fatal, whereas over-reacting with an extreme response to a moderately dangerous stimulus is merely disruptive rather than fatal. Of course, humans are capable of differentiating extreme from moderate stimuli, but the implication is that such fine discriminations occur via a slower, more deliberative process than the one that influences word recognition times. By this categorical model of vigilance, the relation between valence and recognition times should be a step function, with slower responses to negative words than to positive words (Estes & Adelman, 2008a, 2008b).

A different model arises if humans do make use of the fine discrimination of negative, neutral and positive valence, such that it differentially affects either how fast a stimulus activates its lexical or semantic representation (slower for negative words) or how long it engages attention (longer for negative words) or both. By this account, a gradient effect of automatic vigilance is expected. The gradient model predicts a linear negative effect of valence on behavioral latencies, with slower responses to more negative words and a speed-up with an increase in valence. Somewhat surprisingly, none of the recent controlled studies supported such a gradient model of automatic vigilance.

In contrast to prior evidence that negativity *slows down* word recognition (Algom et al., 2004; Estes & Adelman, 2008a, 2008b; Pratto & John, 1991; Wentura et al., 2000; Williams et al., 1996), Kousta et al. (2009) found that valence, whether negative or positive, *sped* word recognition. Such an inverted-U relation between valence and recognition times would entail a double rejection of automatic vigilance: Responses are claimed to be (1) faster to negative words than to neutral words, and (2) equally fast to negative and to positive words. Kousta et al. instead explain their result in terms of motivational relevance: Because negative and positive stimuli respectively activate the avoidance and approach behavioral systems, both valences are "motivationally relevant," and motivationally relevant stimuli are preferentially processed (Lang, Bradley, & Cuthbert, 1990).

Alternatively, an interaction of valence and arousal (Larsen et al., 2008) would imply yet a different model of vigilance. In fact, such an interaction effect on word recognition times would corroborate some prior research on evaluative judgments. Robinson et al. (2004) presented images and words that varied in arousal and valence, and had participants indicate whether the stimulus was negative or positive. They found a similar interaction as that observed by Larsen et al.: Negative words tended to elicit faster responses when they were highly arousing than when they were calming, whereas positive words elicited faster responses when they were calming than when they were arousing. According to Robinson et al., high arousal facilitates responding to negative stimuli because this combination of arousal and negativity is characteristic of dangerous stimuli, and rapid responding to dangerous stimuli is adaptive. Thus, by examining the precise nature of the effects of arousal and valence on word recognition latencies, the present research provides a critical test of various models of automatic vigilance.

Secondly, this research may also inform models of word recognition. Lexical and semantic factors such as word frequency and age of acquisition have long been known to influence the speed with which words are recognized, and decades of research have identified a substantial list of factors that each explain some significant amount of variance in word recognition times. For instance, Adelman et al. (2013) recently assembled a regression model that included a comprehensive list of such factors, and the regression model outperformed all current cognitive models of reading. In so doing, however, Adelman et al. highlighted how little of the potentially explainable (i.e., non-noise) variance is actually explained by the current knowledge in the field. Essentially, Adelman et al. announced a call for the field to search for additional factors or alternative models that can more fully explain word recognition. One class of likely predictors of word recognition missing from current models is emotional factors, which could influence the early activation of lexico-semantic representations and/or the late decisional-response stage of word processing (Yap & Seow, 2013). Thus, by testing for effects of valence and arousal on word recognition, the present research contributes generally to models of word recognition.

Methods

Data

We compiled a set of 12,658 words for which all of the following variables were available.

Emotion variables. Mean *valence* and *arousal* ratings, retrieved from Warriner et al. (2013), served as our predictor variables of primary interest.

Behavioral variables. Mean *lexical decision and naming latencies*, retrieved from the ELP (Balota et al., 2007), served as our criterion variables.

Lexical control variables. *Word length* was controlled via several measures: Orthographic length in characters and morphemes, and phonological length in phonemes and syllables. *Lexical density* was also controlled via several measures: Orthographic, phonological and phonographic neighborhoods (these are the number of words that can be formed from a given word by replacing respectively one letter, one phoneme, or one letter corresponding to one phoneme, with another in its place), and orthographic and phonological Levenshtein distance (OLD and PLD; these are defined as the mean Levenshtein distance between a target word and its 20 closest neighbors, where Levenshtein distance is the minimum number of letter/phoneme insertions, deletions, or substitutions needed to transform the target word into another word). All these values were retrieved from the ELP. *Word frequencies* were retrieved from the 51 million-token SUBTLEX-US corpus of subtitles to the US films and media (Brysbaert & New, 2009), the 130 million-token HAL corpus of electronic communication (Burgess & Livesay, 1998), and the 8 million-token TASA12 corpus of educational materials for 12th graders (Zeno et al., 1995). *Contextual diversity* was also retrieved from SUBTLEX-US, and *age-of-acquisition* (AoA) was retrieved from the norms of Kuperman et al. (2012). We further included the word's *initial phoneme* and its *part-of-speech* (i.e., dominant PoS tag in Brysbaert, New, & Keuleers, 2012).

Statistical Analyses

As demonstrated by Larsen et al. (2008), arousal and valence may enter into interactions that form complex surfaces in the three-dimensional space with arousal, valence and behavioral latency as axes. Recent reports by Kahan and Hely (2008), Scott et al. (2012) and Sheikh and Titone (in press) additionally suggest the possibility of emotion × frequency interactions. These observations necessitate the use of a statistical technique that enables flexible modeling of complex surfaces, without imposing the planar functional form on interactions. Generalized additive mixed-effects (GAM) regression modeling (see e.g., Hastie & Tibshirani, 1990; Wood, 2006) – as implemented in the `mgcv` package (Wood, 2006, 2011) of the R statistical computing software (R Core Team, 2012) – affords the required flexibility and hence is the regression technique of choice here.[5]

[5] For detailed description and worked examples of the use of GAM models in psycholinguistics see Baayen, Kuperman, and Bertram (2010), Tremblay and Baayen (2010), Matuschek, Kliegl, and Holschneider (2012), Kryuchkova et al. (2012), and Balling and Baayen (2012), and for applications in linguistic studies see Wieling et al. (2011) and Koesling et al. (2012).

The distributions of raw lexical decision and naming latencies showed the typical skew (i.e., a heavy right tail), which biases estimates of the mean. A common solution is to transform the distribution such that it closely resembles the Gaussian, and to apply statistical methods that assume an underlying Gaussian distribution of the data (see e.g., Baayen & Milin, 2010; Kliegl et al., 2010). In keeping with this approach, we log-transformed the latencies, as indicated by the Box-Cox transformation test (Box & Cox, 1964). Regression models were thus fitted to log-transformed RTs with Gaussian as the underlying family of distributions and identity as a link function. The results reported below were also obtained with both untransformed RTs and inverse-transformed RTs, so our conclusions are not particular to the transformation itself.

We frame our discussion of the functional form of emotion effects in terms of (non)monotonicity rather than (non)linearity, because a linear effect of a predictor on a log-transformed dependent variable only guarantees a monotonic, not necessarily linear, effect. Moreover, the ratings of valence and arousal are ordinal variables, whereas claims of a linear relationship require variables that are at least interval. Therefore, our research question is better thought of as addressing the question whether the emotion effects on word recognition are monotonic with a (near-)constant rate of change across the entire range, or have a specialized form, such as the step-function, indicating a fast change over a limited part of the continuum and a lesser change in the remainder.[6]

Multicollinearity of predictors in a regression model may inflate standard errors and distort regression coefficients (Mason & Perreault Jr., 1991). In the present set of variables, strong correlations typically exist both within and between measures gauging the rate and time-course of word use (frequency of occurrence, contextual diversity, AoA) and measures

[6] We thank Stephen Lupker for this suggestion.

gauging formal lexical properties (e.g., length in characters, morphemes, phonemes and syllables; orthographic, phonological and phonographic neighborhood sizes, as well as PLD and OLD). Unsurprisingly then, the condition number test calculated for the entire set of continuous variables under consideration (frequency-related measures, length-related measures, valence and arousal) indicated substantial multicollinearity, $\kappa = 87.13$.

Several steps were taken to reduce multicollinearity. First, we applied principal components (*PC*) analysis to the nine variables representing formal lexical properties. Three principal components each explained over 5% of the variance in those formal lexical variables, and taken together accounted for over 90% of the variance. These principal components (labeled *PC1*, *PC2*, and *PC3*) were thus incorporated into our models as statistical estimators of formal lexical properties. (Variables that loaded most strongly on PC1 were length in characters, and orthographic and phonological density; on PC2 – orthographic, phonological and phonographic neighborhoods; and on PC3 – length in morphemes.) Second, the effect of word frequency (from SUBTLEX, log transformed) was partialled out from AoA and log contextual diversity estimates. The residual values (labeled rAoA and rCD) were thus de-correlated from the estimates of frequency and were used in further modeling. Finally, we centered all numerical predictors. The resulting set of PC1, PC2, PC3, rAoA, rCD, word frequency, valence, and arousal variables showed only a mild, acceptable level of multicollinearity, $\kappa = 16.85$.

The set of continuous predictors listed above, as well as factors reflecting the first phoneme and part-of-speech, were entered into GAM models with by-item average RTs as the dependent variable. All continuous predictors were first explored for nonlinear effects, implemented as restricted cubic splines. Predictors that showed no support for a nonlinear functional form were re-entered into final models as linear. We also modeled interactions (implemented as tensor product splines) for predictors that were shown or hypothesized to interact in prior research (i.e., valence × arousal, frequency × valence, and frequency × arousal). Because the dependent variables were by-item average RTs, there were no random effects in any models fitted to the item-level data.

Results

Our analyses addressed a progressive series of research questions, reported in turn.

What is the functional relation between word frequency and emotional factors?

Various corpora have been used for estimating word frequencies in prior studies. However, they differ in potentially relevant ways (e.g., content and size), and indeed they are not equally good at predicting word processing times (Brysbaert & Cortese, 2011; Brysbaert & New, 2009). We therefore first examine whether the various corpora yield systematically different patterns of word frequency estimates across the ranges of valence and arousal. Figure 1 demonstrates the functional relationship of valence and arousal with word frequency estimates from TASA12, SUBTLEX, and HAL. The figure is based on 12,092 words overlapping between the three corpora. The vertical separation among the lines simply reflects the differing sizes, and hence the differing absolute word frequencies, of the various corpora: TASA12 and HAL respectively are the smallest and largest of the three corpora, so they respectively yield the lowest and highest frequency counts. Valence and arousal are binned into twenty quantiles, each accounting for 5% of the respective distribution, and the mean log frequency is reported for each bin.

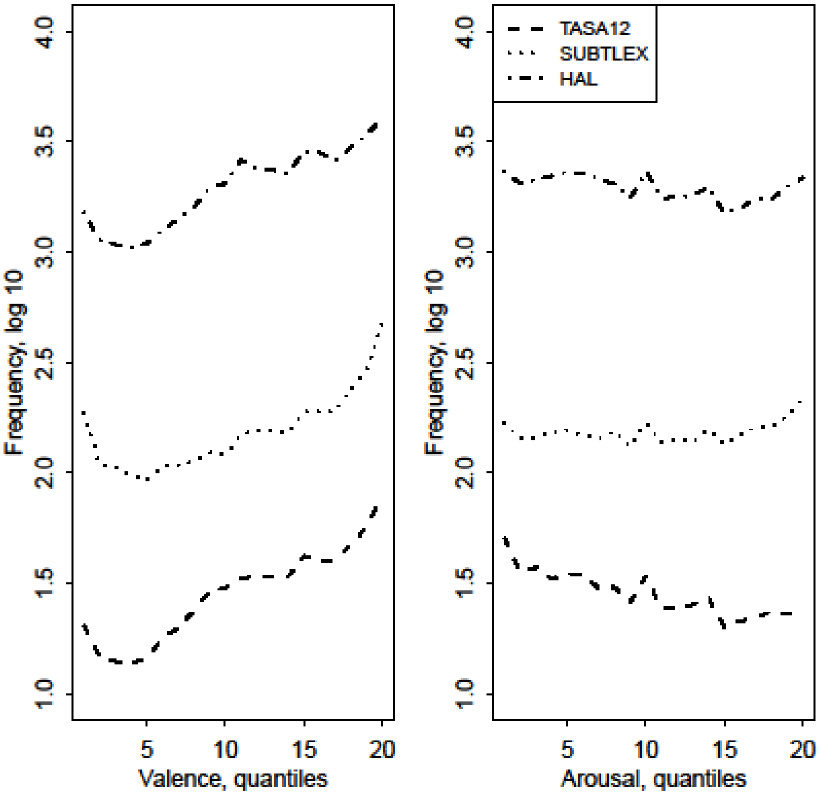


Figure 1. Functional relation between valence and word frequency (left) and arousal and word frequency (right). Log (10)-transformed word frequencies are estimated for the SUBTLEX corpus based on subtitles to US films and media, the TASA12 corpus based on reading materials for American 12th graders, and the HAL corpus based on internet communications. Valence and arousal are binned into twenty 5% quantiles and the mean log frequency is shown for each quantile.

Frequency distributions across the valence range (left panel) are similar across corpora. While there is an overall trend for more positive words to be more common within each of the three corpora (i.e., all three lines peak on the right end of the scale), very negative words are more frequent than moderately negative words. The observed spike in frequency of very negative words will become important in our comparison of prior and present findings. The functional relationships of frequency and arousal (right panel) differ substantially. TASA12 contains mostly low-arousal words, with highly arousing words being relatively rare, as indicated by a frequency curve that decreases sharply across the arousal range. Put simply, educational texts (TASA12) contain boring words, possibly due to editorial requirements to what counts as appropriate content for school-level reading. SUBTLEX, in contrast, shows a relatively flat pattern across the arousal range, with an increase in frequency in very arousing words. Film and media subtitles (SUBTLEX) thus unsurprisingly contain more exciting words, as befits their purpose of attracting and maintaining viewers' attention. Finally, HAL exhibits an essentially flat distribution of frequency over the arousal range. That is, electronic communication (HAL) contains an approximately equal number of boring, neutral and exciting words. In what follows we only consider SUBTLEX and HAL frequency estimates, as these two corpora are larger and show a stronger convergence than the TASA frequency counts which are based on a (6 to 16 times) smaller sample of edited educational materials.

What is the relation between emotional factors and word recognition when the emotion × frequency interaction is *not* taken into account?

Several recent studies have indicated that emotion may interact with word frequency in affecting word processing (Kahan & Hely, 2008; Scott et al., 2009, 2012; Sheikh & Titone, in press), but the prior regression studies did not include emotion × frequency interactions. For comparison with those prior regression studies, we thus examined such non-interactive relationships between emotional factors and behavioral latencies in our larger and more representative dataset. We plotted valence and arousal against lexical decision and naming response times: as shown in Figure 2, we replicated the inverted-U effect of valence on response times, as originally shown by Kousta et al. (2009). Importantly, the inverted-U shape of the valence effect was retained after statistically accounting for all of the control variables listed in the Methods (plot not shown). These control variables included word frequency (SUBTLEX) but not its interactions with valence and arousal. Unlike Kousta et al., however, our analysis also revealed an inverted-U effect of arousal on response times. Thus, when the hypothesized interactions of word frequency with valence and with arousal were omitted from the analyses (as in prior studies), the inverted U-shaped relationship between emotional factors and behavioral latencies was replicated.

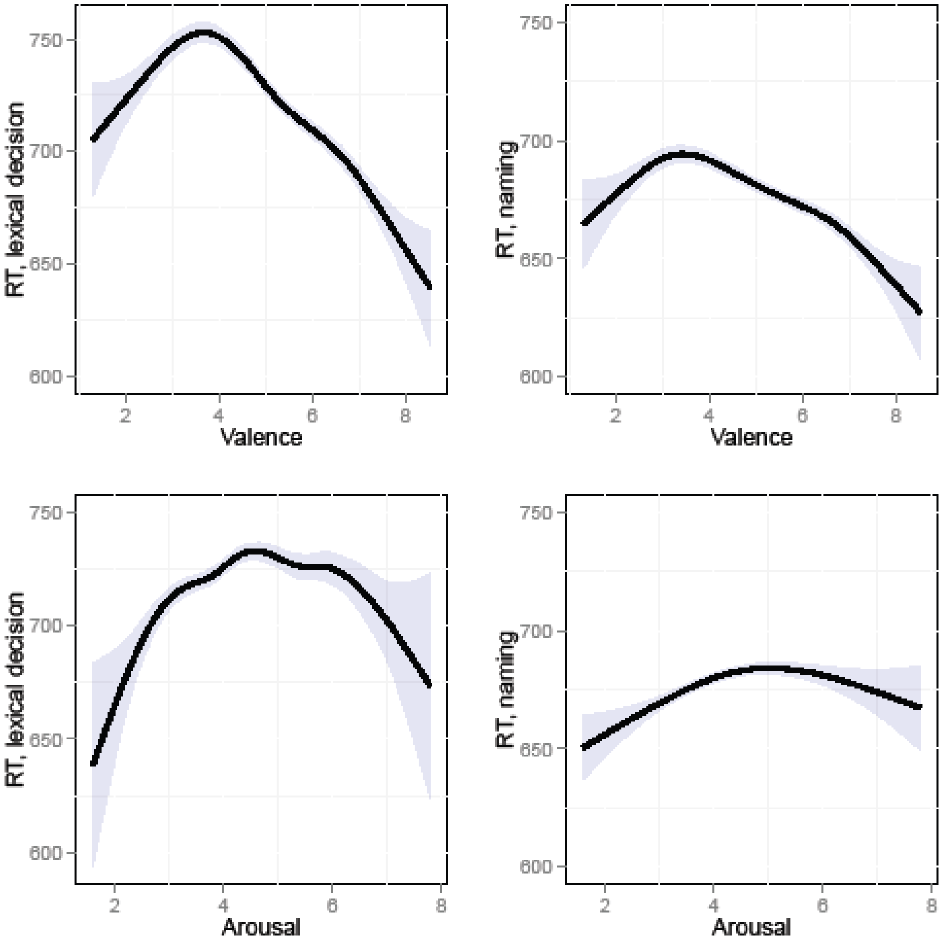


Figure 2. Functional relationships of valence (top row) and arousal (bottom row) with lexical decision latencies (left column) and naming latencies (right column) across all frequency levels (i.e. emotion × frequency interactions are unaccounted for). The shape of valence and arousal effects was evaluated using cubic splines. Each curve is reported with the 95% confidence interval (the gray area).

What is the relation between emotional factors and word recognition when the emotion × frequency interaction *is* taken into account?

Figure 3 summarizes the effects of valence (top row) and arousal (bottom row) on lexical decision (left column) and naming (right column) response times, plotted as a function of word frequency (SUBTLEX). Each panel displays a series of five trend lines estimated using the cubic spline function for words falling into respective quintiles of lexical frequency (from a solid line for the lowest frequency words to a dotted line for the highest frequency words). The top panels reveal that the effect of valence on behavioral latencies is negative, and the magnitude of the effect is attenuated as frequency increases (i.e., the slope is steep among the high lines but is flat in the lowest line). To illustrate, the magnitude of the effect of valence on lexical decision times (top left panel) was about 55 ms among the lowest frequency words, but among the highest frequency words valence had little or no effect. The bottom panels of Figure 3 reveal that the effect of arousal on behavioral latencies is instead positive, and again the magnitude of the effect is attenuated as frequency increases. In the extreme, the magnitude of the effect of arousal on lexical decision times (bottom left panel) was about 55 ms among the lowest frequency words, but among the highest frequency words arousal had little effect.

The patterns in Figure 3 are based on raw data, and are fully confirmed by the regression model that includes emotion × frequency interactions (see models below, plots not shown). The consistent near-linear trends observed in all frequency bands (Figure 3) reveal that the inverted-U shape (Figure 2), which is only observed when emotion × frequency interactions are unaccounted for, substantially mischaracterizes the effect of emotion on word recognition behavior. Finally, the patterns in Figures 2 and 3 are based on SUBTLEX frequencies, but those same patterns are also observed when HAL frequency counts are used instead (plots not shown). Thus, despite being independent corpora based on different genres of text (i.e., film and media subtitles; internet communications), SUBTLEX and HAL frequencies yielded strikingly similar emotion × frequency interactions. Full results of GAM regression models are reported next, separately for lexical decision and naming.

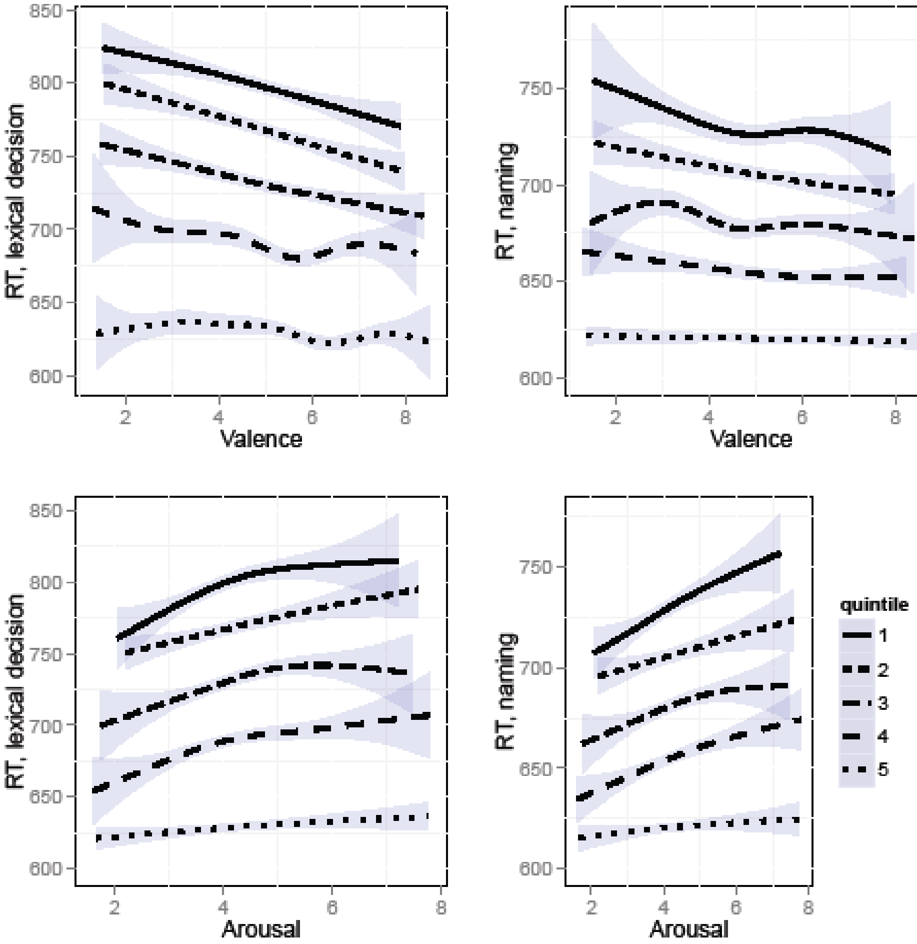


Figure 3. Functional relationships of valence (top row) and arousal (bottom row) with lexical decision latencies (left column) and naming latencies (right column), displayed by quintiles of word frequency (SUBTLEX). The highest-frequency words are the 5th quintile. The shape of valence and arousal effects was evaluated using cubic splines. Each curve is reported with the 95% confidence interval (the gray area).

Lexical Decision. A model fitted to lexical decision RTs identified a number of outliers (1.53% of the data points) that were further than 2.5 standard deviations from the model's fitted values (Baayen & Milin, 2010). These outliers were removed and the model refitted. Table 2 reports the model's outcome. Part A of the table lists the linear effects of continuous predictors. For brevity, the effects of factorial predictors with multiple levels – namely, part-of-speech and first phoneme – were omitted from the table. However, both of these control factors were significant, and the full model's output is available upon request. Part B lists the nonlinear effects (i.e., smooth terms, for which the assumption of nonlinearity was warranted, $p < 0.001$ and effective degrees of freedom $edf > 1$) and the emotion × frequency interactions (i.e., tensor products)[7]. The model explained 60.15% of the variance in latencies.

[7] The output of the generalized additive models differs from outputs of most regression or ANOVA models in that the estimates and inferential statistics for tensor products are reported for the entire hyperbolic surface, without separating it into more customary separate representations of main effects and interactions. The main effect of frequency is not omitted, but rather is fully accounted for when frequency is entered as one of terms in the tensor product with valence, arousal or any other variable.

Table 2: Generalized mixed additive model fitted to log-transformed lexical decision latencies. Linear effects (Part A) include linear predictors, whereas smooth terms (Part B) include nonlinear predictors and interactions.

A. Linear effects	Estimate	SE	*t*	*p*
Intercept	6.5702	0.0047	1390.7666	< 0.0001
PC2	-0.0016	0.0021	-0.7743	0.4388
B. Smooth terms	*edf*	*Ref.df*	*F*	*p*
PC1	6.5178	7.6971	244.5655	< 0.0001
PC3	5.0416	6.1925	38.5429	< 0.0001
Age of acquisition (residual)	4.8625	6.0297	128.1196	< 0.0001
Contextual diversity (residual)	5.2017	6.3745	52.2712	< 0.0001
Frequency × valence (tensor product)	12.0771	14.4453	39.9977	< 0.0001
Frequency × arousal (tensor product)	5.0539	20.0000	1.3966	< 0.0001

Note. *Part of speech* and *First phoneme* were both categorical predictors with multiple levels (4 and 32 respectively). For brevity we omit their inferential estimates from the model's output. *edf* = estimated degrees of freedom, *Ref.df* = reference degrees of freedom.

F-test model comparisons were conducted to establish whether the valence × frequency tensor product, and separately the arousal × frequency tensor product, significantly improved the performance of the baseline model with nonlinear non-interacting effects of frequency, valence, and arousal. Both tensor products were indeed warranted as terms in the best-performing model (Table 2), with all *ps* < 0.001 in model comparison tests. Furthermore, the tensor product of frequency and valence was preferred by the model comparison test over the independent non-linear effect of valence (which was significantly negative, *p* < 0.001). Likewise, the tensor product of frequency and arousal was preferred over the nonlinear effect of arousal (which was nonsignificant, *p* = 0.15). In short, adding the frequency × valence and frequency × arousal interactions significantly improved the fit of the models. The tensor product of valence and arousal did not reach significance in any of the models, suggesting that these affective properties have independent effects.

Critically, including these emotion × frequency interactions revealed effects (Figure 3) that are strikingly different from those observed when the interactions are excluded from the models (Figure 2): Namely, what previously appeared as inverted U-shaped effects of valence and arousal on response times are now revealed to actually be monotonic, essentially linear effects. In none of the frequency bands did the effect of valence on response times exhibit an inverted U-shape. The effect of arousal on response times was also monotonic and near-linear, rather than inverted U-shaped.

There was no straightforward way to estimate the unique variance explained by either valence or arousal, as their impact was modulated by frequency. As an approximate estimate, we compared the amounts of variance explained by (a) the nonlinear effect of frequency, (b) the tensor product of frequency and valence, and (c) the tensor product of frequency and arousal. Models with predictors outlined in (a)-(c) were fitted to RTs from which effects of all other predictors (principal components PC1, PC2, and PC3, AoA, contextual diversity, first phoneme, and dominant part-of-speech) were partialled out. Frequency alone (a) explained 24.4% of the variance, including the frequency × valence interaction (b) explained 26.3%, and including the frequency × arousal interaction (c) explained 24.5% (including both interactions together explained 26.4%). We conclude that the contribution of valence to explained variance (the difference between (a) and (b)) is on the order of 2%, while the contribution of arousal (the difference between (a) and (c)) is much smaller (0.1%).

Naming. The modeling procedure was repeated with naming latencies. Table 3 reports the model fitted to log-transformed (base e) naming latencies after removing outliers (1.82% of the data points). Part A of the table again lists the linear effects of continuous predictors, whereas Part B lists the nonlinear effects and the emotion × frequency interactions. Again, for brevity, the effects of part-of-speech (nonsignificant) and first phoneme (significant) were omitted from Table 3, but the full model's output is available upon request. The model explained 58.01% of the variance in latencies. As with lexical decisions, F-test model comparisons indicated that both tensor products (frequency × valence and frequency × arousal) significantly improve the model's performance as compared to a set of non-interacting, nonlinear effects of frequency, valence and arousal (all *ps* < 0.01). Once again, the interaction of valence and arousal was nonsignificant (*p* = 0.3), pointing to the independent nature of these effects.

Table 3: Generalized mixed additive model fitted to log-transformed naming latencies. Linear effects (Part A) include linear predictors, whereas smooth terms (Part B) include nonlinear predictors and interactions.

A. Linear effects	Estimate	SE	*t*	*p*
Intercept	6.4860	0.0039	1657.4851	< 0.0001
B. Smooth terms	*edf*	*Ref.df*	*F*	*p*
PC1	5.6380	6.8167	251.5300	< 0.0001
PC2	6.5550	7.7254	2.9642	0.0030
PC3	6.6463	7.7718	43.1921	< 0.0001
Age of acquisition (residual)	5.9141	7.1296	143.4290	< 0.0001
Contextual diversity (residual)	3.1576	4.0278	18.9447	< 0.0001
Frequency × valence (tensor product)	7.4234	8.1773	25.6161	< 0.0001
Frequency × arousal (tensor product)	3.8260	20.0000	0.9284	0.0001

Note. *Part of speech* and *First phoneme* were both categorical predictors with multiple levels (4 and 32 respectively). For brevity we omit their inferential estimates from the model's output. *edf* = estimated degrees of freedom, *Ref.df* = reference degrees of freedom.

Amounts of variance in naming latencies explained by valence and arousal, with all other effects partialled out, were as follows. Frequency alone (a) explained 11.1% of the variance, including the frequency × valence interaction (b) explained 11.3%, and including the frequency × arousal interaction (c) explained 11.2% (including both interactions together explained 11.5%). Thus, in naming the contribution of valence to explained variance is a small but significant 0.2%, while the contribution of arousal is an even smaller but still significant 0.1%.

Are emotion effects robust across individual trials?

The preceding analyses, and indeed all prior studies, examined emotion effects at the level of words (or "item means"): Each word has a mean response time, and among the set of words, we test whether the words' valence and arousal ratings tend to predict their mean response times. This analysis provides a general, averaged view of emotion effects on word recognition. Here we additionally examine emotion effects at the level of individual trials: Each trial of the lexical decision and naming studies produces a single response latency, and among all those individual trials, we test whether the given word's valence and arousal ratings tend to predict the individual response times that the given word elicited by each participant. To illustrate, suppose a hundred words are presented to each of a hundred participants in a lexical decision study. In the standard word-level analysis (a.k.a. "item analysis", "by-items analysis", or "F2"), there would be 100 rows of data (one per item). But in the trial-level analysis, there is a row for each trial of each participant, so there would be 10,000 rows of data (100 x 100). Clearly, this trial-level analysis is far more statistically powerful, though it must be noted that individual response latencies are also far more variable (due to random "noise" that is averaged out of word-level analyses).

As in previous analyses, only correct responses were considered, and we excluded outliers identified in the ELP data (Balota et al., 2007) as trials with latencies more than 3 standard deviations from the word's mean latency. The resulting data sets contained 384,113 and 329,871 data points for lexical decision and naming respectively. Our models (not shown) had the same configuration of predictors as outlined above, with an addition of such predictors as the latency and correctness of the previous response, and the position of the word in the participant's experimental list. The maximal random effects structure was implemented in the models, with by-subject and by-word intercepts, as well as by-subject slopes for valence, arousal, and frequency and their interactions (Barr et al., 2013). The results were very similar to the ones observed in average latencies. Namely, both lexical decision and naming latencies monotonically decreased with increasing valence, while the valence effect was at its strongest in the lower-frequency words and gradually diminished in magnitude as word frequency increased. The same attenuation of effect with increasing frequency was observed for the positive correlation of arousal with lexical decision and naming latencies. Finally, the valence × arousal interaction did not reach significance in the trial-level data, over and above the frequency × valence and frequency × arousal interactions. Thus, the trial-level analysis replicated the emotion effects observed in the word-level analysis reported above.

Are emotion effects on word recognition independent of semantic variables?

Our preceding analyses included a large number of lexical control factors, but recently several measures of additional semantic factors have emerged. Most pertinently, there is growing interest in "semantic richness", which is essentially the amount or diversity of information that a given word evokes. For instance, "dog" tends to evoke a rich array of sensory and encyclopedic information, whereas "twig" tends to evoke less information. It could reasonably be argued that emotion is merely one facet of semantic richness, and thus the question arises whether emotion effects on word recognition are really just another demonstration of semantic richness effects. We therefore examined the correlations of valence and arousal with a battery of semantic measures, and we tested whether these emotional factors explained any unique variance in word recognition times after statistically accounting for those semantic variables. For this analysis we identified a set of 1083 monosyllabic words for which all of the following measures were available: valence and arousal ratings (Warriner et al., 2013), SUBTLEX frequency of occurrence (Brysbaert & New, 2009), age-of-acquisition ratings (Kuperman et al., 2012), imageability ratings (Cortese & Fugett, 2004; Schock et al, 2012), sensory experience ratings (Juhasz, Yap, Dicke, Taylor, & Gullick, 2011; Juhasz & Yap, 2013), body-object interaction ratings (Tillotson, Siakaluk, & Pexman, 2008), semantic diversity measures (Hoffman, Ralph, & Rogers, 2012; see also Jones, Johns, & Recchia, 2012) and the word's number of senses from Wordnet (Miller, 1995).

Table 4: Spearman's correlations of valence and arousal with semantic richness measures.

Measure	Valence	Arousal
Body-object interaction	.15*	-0.15*
Imageability	.19*	-.09*
Number of senses	.11*	.00
Semantic diversity	.07*	.00
Sensory experience	.07*	.19*

* *p* < .05.

Table 4 demonstrates that although all correlations were weak in magnitude (|ρ| < 0.2), all were significant (p < .05) except those of arousal with semantic diversity and with the number of senses. Based on these correlations as well as ones reported in Warriner et al. (2013, Table 5) we observe that positive words are consistently associated with higher semantic richness: they are more concrete, imageable, sensorily acute, prone to be used in body-object interactions, etc. To evaluate the amount of variance explained by each of these affective and semantic variables, we calculated the difference in multiple R^2 between a model with non-linear functions of word length, log frequency and age-of-acquisition and a model which included those same predictors plus a non-linear function of one of the variables under comparison. All models were fitted to log-transformed lexical decision latencies. Inclusion of arousal explained an extra 1.4% of the variance (54.4% vs 53%), and valence explained an extra 1.1%. These increments were significant (p < 0.01) and stronger than those associated with most other semantic variables: Sensory experience ratings 0.3%, semantic diversity 0.2%, number of senses 0.1%, imageability 0.6%. The amount of variance explained by body-object interactions (1.1%) was on par with that of valence, and smaller than that of arousal. Finally, we observed a significant increment of R^2 when valence was added to form a tensor product with word frequency in the model that additionally had as predictors nonlinear functions of word length, AoA, and all semantic variables listed above. The amount of unique variance associated with valence, calculated over and above the influence of all semantic predictors, was 1.1% (56.1% vs 55%). The comparable quantity for arousal was 1.2%.

We conclude that the independent impacts of valence and arousal cannot be ascribed to their correlations with a large range of semantic variables (these correlations were weak). Nor can those emotion effects be attributed to the variance the affective measures share with the semantic richness measures: The contributions of both valence and arousal are independent of and stronger than those of the semantic variables, and are numerically the same regardless of whether they are estimated over and above the other semantic variables.[8]

Discussion

Converging empirical patterns observed in word-level and trial-level data from lexical decision and naming RTs in American English yield the following conclusions.

1. Valence has a monotonic effect on word response times, such that negative words (e.g., *coffin*) tend to be responded to more slowly than neutral words (e.g., *cotton*), which tend to be responded to more slowly than positive words (e.g., *kitten*). Specifically, the underlying functional form of this relation between valence ratings and log-transformed RTs was strictly linear in the regression analyses we ran; using a curvilinear form for valence failed to improve the fit of the model to the data. Note however that, because the precise statistical properties of the valence scale are currently unknown and because the RTs were log transformed, the linear nature of this effect must be interpreted with caution. What can be concluded with more confidence is that the effect is monotonic and thus constant in polarity across the entire range: Greater negativity generally slows lexical decision and naming RTs.

2. Arousal has a monotonic effect on word response time, such that calming words (e.g., *sleep*) tend to be responded to more quickly than arousing words (e.g., *sex*). That is, arousal slows word processing. As with valence, the relation between arousal and RT was

[8] A slightly more prominent predictive role of arousal, as compared to valence, in the subset of 1083 monosyllabic words is intriguing given arousal's negligible role in the entire data set of over 12,000 mono- and polysyllabic words. We link this inflation in the predictivity of arousal in the smaller dataset to the fact that monosyllabic words, as compared to the full word set, are significantly shorter in length (4.36 vs 7.21 characters), higher in (log 10) frequency (2.70 vs 1.99), higher in valence (5.16 vs 5.08) and lower in arousal (4.05 vs 4.20), among other differences (all *p*s < 0.01). The discrepancy serves as another argument against selecting data samples that differ in relevant ways from the language's lexicon as found "in the wild". In this case, a consideration of an exclusively or even predominantly monosyllabic data set would lead to a perception of arousal as a stronger predictor than it proves to be in a more exhaustive analysis.

strictly linear in our analyses, but again due to potential nonlinearities in the valence scale and/or the log-transformed RTs, we conclude only that the effect is monotonic.

3. Valence has a stronger effect on word processing than does arousal. Valence explains about 2% of the variance in lexical decision times and 0.2% in naming times, whereas the effect of arousal in both tasks is limited to 0.1% in the analysis of the full dataset.

4. The effects of valence and arousal on word response times are independent, not interactive. Adding an arousal × valence interaction term to the model failed to improve its fit, even when the interaction was flexibly modeled as a hyperbolic surface.

5. Valence and arousal both interact with word frequency, such that valence and arousal exert larger effects among low-frequency words than among high-frequency words.

6. Valence and arousal have stronger effects on lexical decisions than on naming. Valence and arousal together explained more than 2% of the variance in lexical decision latencies, whereas their effects on naming latencies were less than .5%.

Empirical Integration

Our results support many prior findings. Specifically, results 3, 5, and 6 corroborated prior studies showing respectively that valence is more powerful than arousal (see Table 1; see also Adelman & Estes, 2013), that both interact with frequency (Kahan & Hely, 2008; Scott et al., 2009, 2012; Sheikh & Titone, in press), and that they affect lexical decisions more than naming (Estes & Adelman, 2008a; Larsen et al., 2008). On the other hand, our findings 1, 2, and 4 are novel and inconsistent with some prior results. We consider each of these empirical discrepancies in turn.

The observed functional form of the valence effect is novel and contradicts prior claims that this effect is either a step function or an inverted-U function (Estes & Adelman, 2008a; Kousta et al., 2009; Vinson et al., 2013). Our additional analyses indicate that this discrepancy is likely due to a combination of factors. First, the present dataset is much (9-13 times) larger than the ones used in previous studies. This advantage yields a more natural representation of the ranges of frequency, arousal and valence; a more precise account of nonlinear functional relations between frequency, valence and arousal; and a higher accuracy of estimated curves and hyperbolic surfaces that characterize the effects of emotional variables over and above frequency and other statistical controls. One aspect that a larger dataset may have remedied is an over-representation of extremely negative words in prior studies (Estes & Adelman, 2008a; Kousta et al., 2009; Vinson et al., 2013). Those studies were based on an original or slightly extended ANEW data set, which was specifically developed to include a preponderance of emotional words. To illustrate, whereas the extremely negative words (i.e., those with a mean rating of less than 2 on a 1-to-9 scale) constitute 4.8% of the ANEW sample, they constitute only 0.7% of the Warriner et al. (2013) sample. That is, the relative frequency of extremely negative words is about 7 times higher in ANEW than in Warriner et al.'s randomly sampled word set that we use here. Yet very negative words come with a spike in frequency in all of the three corpora considered (Figure 1): for instance, the bottom 5% bin of the valence distribution (valence: 1.34-2.76) has a higher mean log frequency than any single bin between 5 and 35% of the valence distribution (valence: 2.77-4.74). The over-representation of relatively frequent words in the narrow very negative subrange of valence may have led to the attribution of the response speed-up in negative words to the valence effect, whereas it is in fact due to the effect of frequency.

Second, ours is the first study to consider interactions of frequency and emotion in lexical decision and naming. We show that the inverted-U shape of the valence and arousal effects is only observed when emotion × frequency interactions are not accounted for in the analysis (Figure 2). When considered in specific frequency bands, valence and arousal show monotonic near-linear effects, and never the inverted U-shaped effects (see Figure 3). The same monotonic effects are also observed when word frequencies are estimated from HAL instead of SUBTLEX. This suggests, again, that the inverted-U shape may be an artifact of skewed distributions of frequency across the valence range, with higher frequency associated both with very negative and very positive words. The interactions in which strong effects of emotion are observed in low-frequency bands (negative for valence, and positive for arousal) and attenuating effects are observed in words of increasing frequency dovetails perfectly with earlier findings that effects of imageability, age-of-acquisition and other lexical variables are the strongest in lowest-frequency words (e.g., Cortese & Schock, 2013; Gerhand & Barry, 1999a, 1999b).

The monotonic positive effect of arousal is also novel: Kousta et al. (2009) found no effect of arousal, and although Estes and Adelman (2008a) did obtain significant effects of arousal, those effects were in the opposite direction to the effect observed here. The fact that Kousta et al. (2009) found no effect of arousal is unsurprising, considering the extremely small magnitude of the effect that we observed here. The fact that Estes and Adelman (2008a) found a negative effect of arousal can be explained by differences in corpora used to estimate word frequencies. Figure 1 shows that highly arousing words are relatively more frequent in the SUBTLEX and HAL corpora than the TASA12 corpus (i.e., films and websites are more exciting than textbooks). This underestimation of the frequency of high arousal words in TASA12 as compared to SUBTLEX or HAL corpora leads the statistical models to misattribute the facilitative effect of their frequency to a facilitative effect of arousal instead. Because the frequency underestimation is at the high end of the arousal range, this produces an erroneously negative effect of arousal on word recognition. However, when the relatively high frequency of high arousal words is fully accounted for (via SUBTLEX or HAL frequencies), the relation between arousal and word recognition is shown to be positive rather than negative (see Figure 3).

Finally, the independent nature of the valence and arousal effects is novel and fails to replicate the interaction reported by Larsen et al. (2008) in lexical decisions, though it is in line with Vinson et al.'s (2013) findings. While we cannot identify the exact source of discrepancy, it is may stem from our more accurate estimation of effects and interactions due to a larger dataset, the use of hyperbolic surfaces rather than planes in the three-dimensional space to approximate interactive terms, and finally, from our consideration of emotion × frequency interactions, which could have absorbed the variance otherwise attributable to valence × arousal interactions.

One may reasonably wonder, then, why our results should be preferred over prior studies. First, it must be noted that the three preceding studies in Table 1 were *not* independent analyses. Larsen et al. (2008) analyzed the same dataset as Estes and Adelman (2008a), and Kousta et al. (2009) also analyzed a largely overlapping dataset with about 70% of the same stimulus words. So even in cases where our result differs from all three prior studies – as in the arousal effect – this should not be counted as three observations weighed against one observation, because those three observations were based effectively on a single dataset that was analyzed in three ways. Second, whereas the stimuli in prior studies were sampled for their emotionality, the stimuli in the present study represent all words rated as known by at least 70% of raters in the norming study of Kuperman et al. (2012), and without regard for their emotionality. Thus, our sample of stimuli presumably is more representative of natural language. Third, our stimulus sample is about 10 times larger than the previous studies. So again, our stimuli presumably are more representative. Fourth, our analyses included about twice as many lexical and semantic control factors as the prior studies, including multiple sources of word frequency estimates, and including the emotion × frequency interactions that are so important in word recognition. This greater stimulus control results in stronger internal validity for our study than for prior studies. Thus, overall, our results are more likely to be both internally and externally valid than prior results.

Theoretical Implications

The results also necessitate a new explanation of the affective effects in word processing. Previously, the automatic vigilance model was used to describe the origin of a valence effect that was thought to be categorical (Estes & Adelman, 2008a, 2008b), an inverted-U (Kousta et al., 2009), or interactive with arousal (Larsen et al., 2008). The present analyses revealed instead (1) that increasing valence speeds up lexical decisions, (2) that the effect is present across the entire range going from negative, over neutral, to positive words, (3) that the effect interacts with word frequency, and (4) that it does not interact with arousal (which itself has a small positive effect). The finding that the effect of valence is present across the entire continuum is a problem, for instance, for a view which attaches special status to negative (threatening) words, as this would predict a considerable difference between negative and neutral words but not between neutral and positive words. In fact, these results are problematic for all three of the prior models of automatic vigilance, as the effect of valence on RTs was neither categorical, an inverted-U, nor interactive with arousal. The present results instead suggest a gradient model of automatic vigilance, whereby a stimulus elicits a heightened effect in proportion to its negativity, and fine discriminations between negative, neutral and positive stimuli occur fast enough to influence the lexical decision or naming process.

Our results also reveal, for the first time, that arousal has a detrimental effect on word recognition times. More exciting words elicited slower responses. Among infrequent words this effect was about 40 ms in both lexical decision and naming, and again this effect was halved to about 20 ms among frequent words. The challenges are to explain why the effect (1) is detrimental, (2) is observed across the entire range, (3) interacts with word frequency, but (4) does not interact with valence. At the same time, it should be kept in mind that the contribution of arousal to lexical decision times is very small (.1% for the full dataset), so that it may not be warranted (yet) to come up with very strong theoretical proposals.

Factors influencing lexical decisions and naming can affect two processing stages: (1) the activation of word representations in the lexico-semantic system, and (2) the use of this information to execute a response (see Yap & Seow, 2013). Our correlational results do not allow us to pin down the sources of the effects, but plausible hypotheses do emerge from existing models of word recognition (e.g., Grainger & Jacobs, 1996; Norris, 2006) and affective priming (the finding that positive targets are processed faster after positive primes and negative targets faster after negative primes; e.g., Schmitz & Wentura, 2012; Spruyt, De Houwer, Hermans, & Eelen, 2007; Topolinski & Deutsch, 2013). These possible sources of the emotional effects on word processing are considered in detail below.

Lexico-semantic explanations of automatic vigilance. As Schmitz and Wentura (2012) report, there is a long-standing debate about the representation of valence in semantic memory. Bower (1991) suggested there were nodes for positive and negative valence in the semantic network with which valence-laden concepts were associated. In this way, the valence of concepts was not only known, but concepts (and hence words) could prime concepts of similar valence as well (i.e., affective priming). An alternative view was proposed by Masson (1995) and McRae, de Sa, and Seidenberg (1997). In their distributed models, valence was coded in a series of units (roughly representing semantic features) and shared units between concepts made it easier to activate one concept on the basis of another. Topolinski and Deutsch (2013) showed that participants' affect changes briefly (for around 1 s) when stimuli with a strong positive or negative valence are presented, and critically for our purposes here, the degree of semantic priming is larger after positive affect inductions than after negative affect inductions. Thus, positive words may briefly lift the affect of the participants, increasing the affective or semantic priming of subsequent positive words. Negative words, in contrast, would temporarily induce negative affect and therefore prime responses to negative words, but crucially this negative affective priming would be smaller than positive affective priming.

Another possibility is that there are more positive word types than negative. A small but significant positivity bias is indeed observed in the rating study of Warriner et al. (2013), as 55.6% of about 14 thousand words were rated above the midpoint of the valence scale (5): positivity biases of a similar magnitude were also observed in multiple other corpora, see Kloumann et al. (2012) and references therein. Given that there are more positive words than negative words, more affective priming could occur for positive words than for negative words. Thus, positive words may elicit greater priming than neutral and negative words because (a) positive words are slightly more common (Warriner et al., 2013), and/or (b) positive words induce larger priming effects (Topolinski & Deutsch, 2013). That is, automatic vigilance could be due to affective priming, as positive words could produce more frequent or larger priming effects than negative words.

A lexico-semantic origin of the valence effect would also offer a parsimonious explanation of why the effect interacts with word frequency (see Kahan & Hely, 2008; Scott et al., 2009, 2012; Sheikh & Titone, in press for similar results in other tasks). Among less frequent words, the size of the valence effect was estimated by the regression model to be about 50 ms in lexical decisions and about 35 ms in naming (Figure 3). Among more frequent words, however, the effect of valence was reduced to about half that magnitude. This modulation by word frequency is common among lexico-semantic factors affecting word recognition. For instance, age of acquisition, letter-sound consistency, and imageability effects are also larger among low frequency words than among high frequency words (Cortese & Schock, 2013; Gerhand & Barry, 1999; Strain, Patterson, & Seidenberg, 1995). Typically, when two factors exert an interactive effect on word recognition, those factors are assumed to arise at the same stage of processing: If the two factors operated at different processing stages, it is unclear how they could interact. So given that frequency effects arise at the lexico-semantic stage of processing, and that frequency interacts with emotional factors, those emotional effects presumably also arise at the lexico-semantic stage.

Decision-response explanations of automatic vigilance. A second locus of the valence effect could be response execution (Yap & Seow, 2013). For instance, automatic vigilance could arise from task-specific processes. Much research in this respect has been done about the decision stage of the lexical decision task (see Kinoshita & Lupker, 2003, for context effects in naming). Two findings are particularly important: (1) lexical decisions are not always made on a full processing of the stimulus materials, and (2) any difference between word and nonword trials speeds up the decision process. Grainger and Jacobs (1996) convincingly showed that “yes”-responses to words are partly based on the overall activation in the lexico-semantic system induced by the stimulus. That is, a yes-decision can be based on the fact that the stimulus activates many resembling word representations rather than on the identification of the stimulus itself. This explains why nonwords with many word neighbors elicit more erroneous responses than nonwords with few word neighbors, and why reaction times to words are faster when the nonwords do not resemble words than when they do (because then the overall activity elicited by the stimulus makes it possible to come to a correct decision). Within this view, positive words could result in faster responses because they have a lower response threshold, perhaps because positive stimuli are less life-threatening than negative stimuli and/or because humans in general seem to show a positivity bias in information processing (Walker, Skowronski, & Thompson, 2003). This would also explain why the valence effect is smaller (or even reversed) in participants with depression (Sharot, 2011) and when participants are brought into a situation that questions unrealistic optimism (Shepperd, Ouelette, & Fernandez, 1996).

Finally, it is simply possible that nonwords in general are perceived as slightly negative because they are unfamiliar: Warriner et al. (2013) show that lower-frequency words tend to be rated with lower valence. If this is the case, the valence of the stimulus will provide information about its “wordness” and will speed up the acceptance of positive words (e.g., Keuleers & Brysbaert, 2011). Thus, the automatic vigilance hypothesis – that negative stimuli engage attention longer than other stimuli – can be translated into “require more word-specific activation” or a “higher level of activation” to exceed the response threshold in a lexical decision task.

An explanation in terms of decision factors makes sense of seemingly contradictory results. Because negative stimuli in general require faster responses, they tend to be detected more rapidly. For instance, Nasrallah and colleagues (2009) subliminally presented negative, neutral, and positive words in an emotion detection task, and they found that negative words were identified more accurately than positive words. This finding suggests that negative stimuli are identified *faster*, or earlier, than other stimuli. However, the automatic vigilance hypothesis was developed to account for the observation that these same words in other tasks elicit *slower* responding (see also Pratto & John, 1991; Williams et al., 1996). A simple solution is that negative stimuli hold attention longer than other stimuli (Fox et al., 2001), and this sustained attention to negativity delays responding on other tasks such as color naming. After all, if the adaptive significance of automatic vigilance is to facilitate avoidance of dangerous stimuli, then negativity should speed rather than slow responding. Estes and Verges (2008) tested this hypothesis directly by having participants make either lexical decisions or valence judgments to the same set of negative words and positive words. Whereas the negative words slowed lexical decisions (as in the present study), they elicited faster valence judgments than positive words. Thus, automatic vigilance does not work by generally slowing responses to negative stimuli. Rather, by this account, negativity slows lexical decisions and color naming because valence is irrelevant to those judgments and therefore must be ignored or disengaged (cf. Fox et al., 2001; Kuperman, 2013).

An explanation in terms of decision factors also readily accounts for the finding that valence has a smaller effect on naming than on lexical decision, because the naming task is less susceptible to decision processes (but see Kinoshita & Lupker, 2003, for evidence that it is not completely insusceptible to decisional factors). Whereas valence and arousal collectively explained 2% of the variance in lexical decision latencies, they explained only 0.3% of the variance in naming latencies.

Lexical processing. Finally, this research also contributes to our understanding of which variables affect performance in word processing tasks. Adelman et al. (2013) demonstrated that even after removing the random noise in word recognition times, the currently best-performing models and sets of word features leave unexplained a relatively large percentage of the variance in word recognition times. Similarly, although Rey and

Courrieu (2010) noticed that there is 85% systematic variance in megastudy lexical decision data, current models do not go beyond 65% (e.g., Kuperman et al., 2012). Therefore Adelman et al. (2013) issued a general call to the field to search for additional factors that affect word recognition, and the present research does just that. Given the broad influence of emotion on cognitive tasks, it is rather surprising that current psycholinguistic models of word recognition entirely neglect the effects of emotion. Although valence and arousal exerted very modest effects on naming times (see also Adelman et al., 2013), we found that valence and arousal collectively explained a reasonably substantial amount (about 2%) of the unique variance in lexical decision times, with most of that effect arising from valence rather than arousal. Although this is a modest effect, it is a further step towards our understanding of which variables do and do not matter in language processing. For instance, it appears that valence may be a more important variable than many of the semantic richness variables recently proposed as relevant for characterizing word recognition.

Acknowledgements

VK’s contribution was supported in part by the SSHRC Insight Development grant 430-2012-0488, the NSERC Discovery grant 402395-2012, and the NIH R01 HD 073288 (PI Julie A. Van Dyke). MB's contribution was made possible by an Odysseus Grant from the Government of Flanders. ABW’s contribution was supported by the Ontario Graduate Scholarship. Thanks are due to James Adelman, Steve Lupker and two anonymous reviewers for their valuable comments on earlier drafts.

References

Adelman, J. S. (Ed.) (2012). *Visual word recognition, Volume 1: Models and methods, orthography and phonology*. Hove, England: Psychology Press.

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word naming and lexical decision times. *Psychological Science, 17*(9), 814–823.

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

Adelman, J. S., Marquis, S. J., Sabatos-DeVito, M. G., & Estes, Z. (2013). The unexplained nature of reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 1037-1053.

Algom, D., Chajut, E., & Lev, S. (2004). A rational look at the emotional Stroop phenomenon: A generic slowdown, not a Stroop effect. *Journal of Experimental Psychology: General, 133*(3)*,* 323-338.

Baayen, R. H. and Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research, 3*, 12–28.

Balling, L. and Baayen, R. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. Cognition, 125, 80–106.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words, *Journal of Experimental Psychology: General*, *133*(2), 283-316.

Balota, D. A., & Lorch, R. F. (1986). Depth of automatic spreading activation: Mediated priming effects in pronunciation but not in lexical decision. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 12*(3)*,* 336-345.

Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed.) *Visual word recognition volume 1: Models and methods, orthography and phonology*. Hove, England: Psychology Press.

Balota, D.A., Yap, M.J., Hutchinson, K.A., Cortese, M.J., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*(3), 445-459.

Barr, D., Levy, R., Scheepers, C., and Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255-278.

Bower, G. H. (1991). Mood congruity of social judgments. In J. P. Forgas (Ed.), *Emotion and social judgments* (pp. 31–53). Elmsford, NY: Pergamon.

Box, G. and Cox, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological), 26*(2), 211–252.

Bradley, M. M., & Lang, P. J. (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.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A.M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology, 58*, 412-424.

Brysbaert, M. & Cortese, M. J. (2011). Do the effects of subjective frequency and age of acquisition survive better word frequency norms? *Quarterly Journal of Experimental Psychology, 64(3),* 545-559.

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), 977-990.

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding Part-of-Speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods, 44*, 991-997.

Butler, B. & Hains, S. (1979). Individual differences in word recognition latency. *Memory & Cognition, 7*(2), 68-76.

Burgess, C. & Livesay, K. (1998). The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kucera and Francis. *Behavior Research Methods, 30*(2), 272-277.

Cortese, M. & Fugett, A. (2004). Imageability ratings for 3,000 monosyllabic words. *Behavior Research Methods, 36*(3):384–387.

Cortese, M. J., & Khanna, M. M. (2007). Age of acquisition predicts naming and lexical decision performance above and beyond 22 other predictor variables: An analysis of 2342 words. *Quarterly Journal of Experimental Psychology, 60(8),* 1072-1082.

Cortese, M. J., & Schock, J. (2013). Imageability and age of acquisition effects in disyllabic word recognition. *Quarterly Journal of Experimental Psychology.* Advance online publication. doi: 10.1080/17470218.2012.722660

Erdelyi, M. H. (1974). A new look at the New Look: Perceptual defense and vigilance. *Psychological Review, 81*(1), 1-25.

Estes, Z., & Adelman, J. S. (2008a). Automatic vigilance for negative words in lexical decision and naming: Comment on Larsen, Mercer, and Balota (2006). *Emotion, 8*, 441–444.

Estes, Z., & Adelman, J. S. (2008b). Automatic vigilance for negative words is categorical and general. *Emotion, 8*, 453–457.

Estes, Z., Jones, L. L., & Golonka, S. (2012). Emotion affects similarity via social projection. *Social Cognition, 30,* 582-607.

Estes, Z. & Verges, M. (2008). Freeze or flee? Negative stimuli elicit selective responding. *Cognition, 108,* 557-565.

Faust, M.E., Balota, D.A., Spieler, D.H., & Ferraro, F.R. (1999). Individual differences in information processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin, 125(6),* 777-799.

Forgas, J. P. (1995). Mood and judgment: The affect infusion model (AIM). *Psychological Bulletin, 117*(1)*,* 39-66.

Fox, E., Russo, B., Bowles, R., & Dutton, K. (2001). Do threatening stimuli draw or hold visual attention in subclinical anxiety? *Journal of Experimental Psychology: General, 130*(4)*,* 681-700.

Gerhand, S., & Barry, C. (1999a). Age of acquisition and frequency effects in speeded word naming. *Cognition, 73*(2)*,* B27-B36.

Gerhand, S., & Barry, C. (1999b). Age of acquisition, word frequency, and the role of phonology in the lexical decision task. *Memory & Cognition, 27*(4)*,* 592-602.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: a multiple read-out model. *Psychological review, 103(3)*, 518-565.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models.* London: Chapman & Hall.

Hoffman, P., Ralph, M. A. L., & Rogers, T. T. (2012). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior research methods, 45*, 718-730.

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology, 66,* 115-124.

Juhasz, B. J., & Yap, M. J. (2013). Sensory experience ratings for over 5,000 mono- and disyllabic words. *Behavioral Research Methods, 45*, 160–168.

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

Kahan, T. A., & Hely, C. D. (2008). The role of valence and frequency in the emotional Stroop task. *Psychonomic Bulletin & Review, 15*(5), 956-960.

Kensinger, E. A., & Corkin, S. (2004). Two routes to emotional memory: Distinct neural processes for valence and arousal. *Proceedings of the National Academy of Sciences, 101*(9)*,* 3310-3315.

Keuleers, E., & Brysbaert, M. (2011). Detecting inherent bias in lexical decision experiments with the LD1NN algorithm. *The Mental Lexicon, 6(1),* 34-52.

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(1)*, 287-304.

Kinoshita, S., & Lupker, S. J. (2003). Priming and attentional control of lexical and sublexical pathways in naming: A reevaluation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29(3),* 405-415.

Kliegl, R., Masson, M., & Richter, E. (2010). A linear mixed model analysis of masked repetition priming. *Visual Cognition, 18*(5), 655–681.

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.

Koesling, K., Kunter, G., Baayen, R. H., & Plag, I. (2013). Prominence in triconstituent compounds: Pitch contours and linguistic theory. *Language and Speech.* Advance online publication. doi:10.1177/0023830913478914

Kousta, S-T., Vinson, D. P., & Vigliocco, G. (2009). Emotion words, regardless of polarity, have a processing advantage over neutral words. *Cognition, 112*(3), 473-481.

Kryuchkova, T., Tucker, B. V., Wurm, L., & Baayen, R. H. (2012). Danger and usefulness in auditory lexical processing: evidence from electroencephalography. *Brain and Language, 122*, 81–91

Kuperman, V. (2013). Accentuate the positive: Semantic access in English compounds. *Frontiers in Language Sciences, 4:203*, 1-10.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*, 978-990.

Kuperman, V., & Van Dyke, J. A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance, 39(3), 802 – 823.*

LaBar, K. S., & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience, 7*(1)*,* 54-64.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1990). Emotion, attention, and the startle reflex. *Psychological Review, 97,* 377-395.

Larsen, R. J., Mercer, K. A., & Balota, D. A. (2006). Lexical characteristics of words used in emotional Stroop experiments. *Emotion*, *6*(1), 62–72.

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*(4), 445-452.

Mason, C., & Perreault Jr, W. (1991). Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research, 28*(3), 268–280.

Masson, M. E. J. (1995). A distributed memory model of semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 3–23.

McRae, K., de Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General, 126*, 99–130.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38(11)*, 39-41.

Nasrallah, M., Carmel, D., & Lavie, N. (2009). "Murder she wrote": Enhanced sensitivity to negative word valence. *Emotion, 9*(5), 609-618.

Norris, D. (2006). The Bayesian Reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review, 113(2)*, 327-357.

Ohman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review, 108*(3), 483-522.

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

Pratto, F., & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of Personality and Social Psychology, 61*(3), 380-391.

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rey, A., & Courrieu, P. (2010). Accounting for item variance in large-scale databases. *Frontiers in Psychology, 1*. doi: 10.3389/fpsyg.2010.00200

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*(11), 1472-1484.

Rowe, G., Hirsh, J. B., & Anderson, A. K. (2007). Positive affect increases the breadth of attentional selection. *Proceedings of the National Academy of Sciences, 104,* 383-388.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110*(1)*,* 145-172.

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology, 76*(5), 805-819.

Schmitz, M., & Wentura, D. (2012). Evaluative Priming of Naming and Semantic Categorization Responses Revisited: A Mutual Facilitation Explanation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 984-1000.

Schock, J., Cortese, M., & Khanna, M. (2012). Imageability estimates for 3,000 disyllabic words. *Behavior Research Methods, 44*(2), 374-379.

Scott G. G., O'Donnell P. J., Leuthold H., & Sereno S. C. (2009). Early emotion word processing: Evidence from event-related potentials. *Biological Psychology, 80*(1), *95-104.*

Scott, G., O'Donnell, P., & Sereno, S. (2012). Emotion words affect eye fixations during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(3), 783-792.

Sheikh, N. A., & Titone, D. A. (in press). Sensorimotor and linguistic information attenuate emotional word processing benefits: An eye movement study. *Emotion*.

Sharot, T. (2011). The optimism bias. *Current Biology, 21(23),* R941-R945.

Shepperd, J. A., Ouellette, J. A., & Fernandez, J. K. (1996). Abandoning unrealistic optimism: Performance estimates and the temporal proximity of self-relevant feedback. *Journal of Personality and Social Psychology, 70(4),* 844-855.

Spruyt, A., De Houwer, J., Hermans, D., & Eelen, P. (2007). Affective priming of nonaffective semantic categorization responses. *Experimental Psychology, 54(1)*, 44-53.

Strain, E., Patterson, K., & Seidenberg, M. S. (1995). Semantic effects in single-word naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21(5),* 1140-1154.

Tillotson, S., Siakaluk, P., & Pexman, P. (2008). Body-object interaction ratings for 1,618 monosyllabic nouns. *Behavioral Research Methods, 40*, 1075–1078.

Topolinski, S., & Deutsch, R. (2013). Phasic affective modulations of semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 414-436.

Tremblay, A., & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In Wood, D., (Ed.), *Perspectives on Formulaic Language: Acquisition and communication* (pp. 151–173). London: The Continuum International Publishing Group.

van Kleef, G. A. (2009). How emotions regulate social life: The emotions as social information (EASI) model. *Current Directions in Psychological Science, 18,* 184-188.

Vinson, D., Ponari, M., & Vigliocco, G. (2013). How does emotional content affect lexical processing? Proceedings of the Cognitive Science Society Conference, Berlin.

Walker, W. R., Skowronski, J. J., & Thompson, C. P. (2003). Life is pleasant--and memory helps to keep it that way!. *Review of General Psychology, 7(2),* 203-210.

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

Wieling, M., Nerbonne, J., & Baayen, R. H. (2011). Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE, 6*(9):e23613.

Wentura, D., Rothermund, K., & Bak, P. (2000). Automatic vigilance: The attention-grabbing power of approach- and avoidance-related social information. *Journal of Personality and Social Psychology, 78*(6)*,* 1024-1037.

Williams, J. M., Mathews, A., & MacLeod, C. (1996). The emotional Stroop task and psychopathology. *Psychological Bulletin, 120*(1)*,* 3-24.

Wood, S. (2006). Generalized Additive Models. Chapman & Hall/CRC, New York.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73*, 3–36.

Yap, M.J., & Balota, D.A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language, 60(4),* 502-529.

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Pereception and Performance, 38*(1), 53-79.

Yap, M. J., & Seow, C. S. (2013). The influence of emotion on lexical processing: Insights from RT distributional analysis. *Psychonomic Bulletin & Review,* advance online publication.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's *N*: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971–979.

Zeno, S., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide*. Touchstone Applied Science Associates.

Chapter 5: Feeling close or far: Graded approach and avoidance of affective stimuli.

Warriner, A.B., Shore, D.I., Schmidt, L.A., & Kuperman, V. (in preparation)

Abstract

Previous research has shown a congruency effect between approach and avoidance movements and valence – people are quicker to take action that will ultimately move them closer to a positive stimulus and further away from a negative stimulus. This, however, dichotomizes valence, a continuous variable, and does not take arousal into account. In this paper, a new method is presented for proportionally measuring approach and avoidance in relation to both valence and arousal. Participants were shown a word and asked to move a manikin representing themselves closer to or farther away from the word. The manikin's distance from the word reflected that word's valence in a very strong, graded and linear fashion. Moreover, gender differences in both valence and arousal ratings to words were commensurate with gender differences in distances to those words. Finally, individual differences were investigated with shyness resulting in greater average distances and sociability resulting in lesser average distances. The correspondence of this graded effect of word valence on approach-avoidance behavior with recent research on the graded nature of automatic vigilance in lexical decision and naming times is discussed along with possible extensions for this new method.

Keywords: valence, arousal, approach, avoidance, individual differences

Introduction

People automatically assess the valence of everything they encounter, unconsciously determining if it is positive or negative (e.g. Carretie, Hinojosa, Martin-Loeches, Mercado, & Tapia, 2004; Duckworth, Bargh, Garcia, & Chaiken, 2002; Giner-Sorolla, Garcia, & Bargh, 1999). The proposed purpose of this automatic assessment is to trigger a preparation for action (Osgood, 1953; Lang, Bradley, & Cuthbert, 1990): a positive evaluation triggers approach, while a negative evaluation triggers avoidance. These two responses are viewed as stemming from, and being subserved by, fundamental —appetitive and aversive— motivational systems (Carver & White, 1994; Lang, 1995). Evidence supporting the link between valence and behaviour derives from multiple demonstrations of a congruency effect between affectively-motivated movement and physical movement required by the experimental task using: levers (Chen & Bargh, 1999), joysticks (Fishbach & Shah, 2006; Rinck & Becker, 2007), manikins (De Houwer, Crombez, Baeyens, & Hermans, 2001), steps (Stins, Roelofs, Villan, Kooijman, Hagenaars, & Beek, 2011) and facial expressions (Neumann, Hess, Schulz, & Alpers, 2005). Participants are faster to make approach-like responses to positive stimuli than to negative ones, and avoidance-like responses to negative stimuli than to positive ones. The implicit nature of these measurements provides an advantage over explicit (i.e., conscious) decisions to approach or avoid. However, these methods have several disadvantages. First, both valence and direction are dichotomized, which also dichotomizes any interaction between the two. Second, this approach largely overlooks other hypothesized dimensions of affect, including arousal, which—according to some theorists (e.g., Lang, 1994; Schachter & Singer, 1962), represents the intensity of motivational system engagement via automatic assessment of stimulus valence. Here we present a new method for quantifying the relation between affect (including both valence and arousal) and the degree to which a person approaches or avoids a stimulus.

Early reports of the congruency effect (Chen & Bargh, 1999) demonstrated that approach and avoidance behaviours are automatically invoked. Participants were given congruent instructions (i.e., to pull a lever when the word presented was good and to push it when it was bad) for half of the blocks, and incongruent instructions for the other half of the blocks. As such, participants were not consciously choosing to approach or avoid stimuli, but their automatic inclination to do so was revealed by shorter reaction times on congruent trials. The authors then made the movement arbitrary: in one block, participants pulled the lever regardless of the stimulus valence and in the other, they pushed it. A congruency effect was still found showing that conscious evaluation is not necessary for the activation of approach or avoidance. A later study showed a congruency effect even when using masked stimuli that were not consciously perceived (Alexopoulos & Ric, 2007).

Early studies (Cacioppo, Priester, & Berntson, 1993; Chen & Bargh, 1999; Wentura, Rothermund, & Bak, 2000) tended to focus on the specific motor act being executed rather than the framing provided by the instructions. Later research, which framed the task in relative terms, supported the approach–avoidance aspect of the congruency. For example, pulling a joystick could be seen as bringing the stimuli closer or as pulling one's hand away (Krieglmeyer & Deutsch, 2010). Reframing the task to focus the observer on the self or the object (Siebt, Neumann, Nussinson, & Strack, 2008) highlighted the importance of relative distance: compatibility effects depended on change in distance between the object and the self rather than the absolute direction of the action. Indeed, it was the *outcome* of the action, and not the specific movement that determined the direction of the compatibility effect (van Dantzig, Pecher, & Zwaan, 2008). Even actions that initially moved positive stimuli away from the self but ended with them closer produced faster reaction times than those that initially brought the stimulus closer, but ended up further away (Krieglmeyer, De Houwer, & Deutsch, 2011). One shortcoming of all of these studies is their dichotomous nature—observers make ballistic responses and reaction time is taken as the dependent measure.

Valence (and arousal) are not dichotomous variables; they fall along a continuum of responses, which can be effectively indicated by participants on a scale (Warriner, Kuperman, & Brysbaert, 2013). Further, these gradient ratings enter into a linear relation with response times in lexical decision and word naming tasks leading to a “gradient automatic vigilance” hypothesis (Kuperman, Estes, Brysbaert, & Warriner, 2014). The present series of experiments examined if approach and avoidance are similarly graded.

Previous work on this question (Puca, Rinkenauer, & Breidenstein, 2006) examined the force applied to a joystick in addition to the RT of the response. Valence did *not* significantly predict the force applied to the joystick. In another attempt to track the continuous nature of valance, participants were asked how many steps they would make towards or away from a person with a particular emotional expression (Seidel, Habel, Kirschner, Gur, & Dernti, 2010). While participants approached sad faces in an implicit joystick task, they avoided them in this explicit rating task. This contrast highlights the importance of comparing implicit and explicit approaches, even when the results are difficult to interpret.

Within this domain, individual and group differences are also an important factor but have rarely been considered. Thus, no substantial gender differences in approach and avoidance motivation have been reported (e.g. Dickson & MacLeod, 2004; Elliot, Gables, & Mape, 2006; Maio & Esses, 2001). It is possible that the two genders pattern similarly in their approach-avoidance behavior overall, but show specific differences in relation to those words which were rated differently in terms of valence and arousal.

Other individual differences may also influence these effects, specifically, personality variables such as behavioral approach or behavioural inhibition tendencies (Carver & White, 1994). Previous studies have hinted at individual differences in approach–avoidance behaviours. Participants with a high avoidance temperament did not show an RT advantage for positive over negative words in a joystick task whereas those with low avoidance temperament did (Puca, Rinkenauer, & Breidenstein, 2006). Participants with high social anxiety were much faster to push a joystick in response to both smiling and angry faces, while non-anxious controls showed no difference despite the fact that valence ratings of those same faces did not differ (Heuer, Rinck, & Becker, 2007).

The present paper introduces a paradigm in which participants make conscious decisions about the relative distance between a manikin representing themselves and a word. Critically, the words chosen represent the entire range of valence and arousal thereby allowing us to evaluate the relation between these dimensions and approach–avoidance behaviour. Our paradigm also allows an examination of the combinatory effects of both valence and arousal which has not been examined in this domain previously.

We conducted three experiments. In Experiment 1, we introduce the paradigm and establish its validity as a measure of the relation between affect (i.e., valance and arousal), and approach–avoidance behavior by collecting responses to a balanced set of words representing the entire range of both dimensions. Participants moved a manikin – a stick figure of a person – towards or away from a word presented at the top of the screen. The distance between the word and the manikin served as the independent variable of approach–avoidance. We predict that participants will place the schematic figure closer to highly positive items, and further from negative items. In Experiment 2, we test whether the well-documented gender differences in affective ratings to words translate into gender differences in approach and avoidance distances obtained with this new paradigm. We predicated that people will choose to move the manikin closer to those words which were rated more positively by their gender. In Experiment 3, we introduced four personality variables (Alexithymia, Behavioral Approach and Inhibition, Affective Style, and Shyness/Sociability) to test whether individual differences played a role in the relation between affect and approach–avoidance distance. We hypothesized that people who focus on punishments, adopt a social avoidance strategy, or are highly shy might adopt a hesitant stance and tend to stay further away from all stimuli while those who focus on rewards, adopt a social approach strategy, or are highly sociable may adopt an exploratory stance and get closer to stimuli. People who are variable in their strategy use or who weight emotional information highly in their decision making may use a fuller range of the distance scale than those who stick with a particular strategy or who focus on sensorimotor information to the exclusion of emotional information.

As arousal has rarely been considered in this domain, we do not have a firm prediction for how it will affect distance choices.

Experiment 1

This first experiment was designed as an initial test of the slider method's ability to detect any systematic relations between distance and other variables. We used a balanced set of words in which valence and arousal were uncorrelated.

Method

Participants

Forty-six students at McMaster University participated in this experiment in exchange for partial course credit. The data from three participants were removed for not making a response on more than 25% of trials. The data from two participants were removed for not being native English speakers. Of the remaining 41 participants, 34 were female, 7 male (age range = 17 to 25, M = 19.10, SD = 2.00).

Stimuli

We selected the 13,763 words from Warriner et al. (2013) that had frequency information available from SUBTLEX, a 51 million-token corpus based on subtitles to US films and TV programs (Brysbaert & New, 2009). These words were divided into 25 bins, by crossing quintiles of valence and arousal. We selected 10 words from each bin for a total of 250. In this subset, valence and arousal were uncorrelated (Spearman's r = -0.019, p = n.s.). All words were monosyllabic, and the mean length was 4.4 characters (range [3, 6]). Mean natural-log SUBTLEX frequency was 6.3 (range [3.1, 10.7]).

Procedure

Participants were tested in groups of up to ten at a time in a computer lab. Each was seated in front of a monitor with a resolution of 1024x768 pixels and responses were made with a mouse. The experiment was programmed using the Experiment Builder software (SR Research, Kanata, ON, Canada).

After answering demographic questions about age, gender, handedness, and education level, participants were instructed as follows:

> *Each trial will start with a fixation symbol (+) near the center of the screen. When you are ready, click at its centre and a new screen will appear. On this screen, you will see a word*

> *either at the top or the bottom with a vertical line below or above it. There will be a person in the centre of that line. The person represents you. Your job is to assess how close you would like to be to the word and communicate that by clicking a point on the line to position the person (you). For example, if the word was DISASTER, you'd probably want to be far away and would click somewhere on the line far away from the word. But if the word was TRIUMPH you might want to be close and would somewhere on the line really close to the word. You are to position the person as QUICKLY AND AS TRUTHFULLY AS POSSIBLE and then click the 'Continue' box to move to the next word.*

The manikin was initially centered at the 400th pixel (see Figure 1 for a sample screenshot). After five practice words, participants were asked if they had any questions before proceeding with the remaining stimuli. Order of word presentation was randomized for every participant. Altogether, this experiment took approximately 30 minutes and was counterbalanced in its order of presentation with another 30 minute experiment which utilized the same words in a digit parity paradigm.

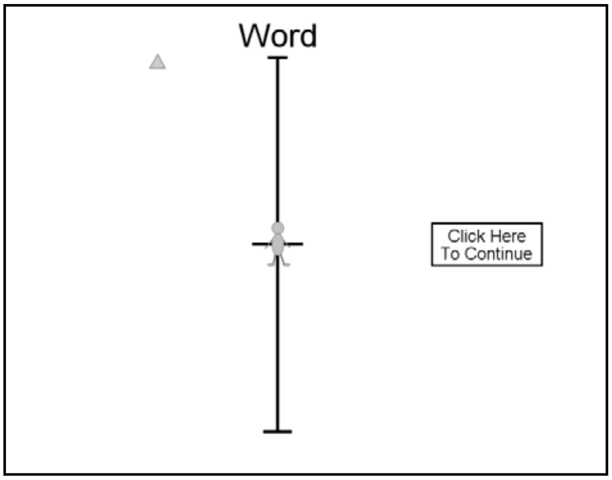


Figure 1. A screenshot showing how the slider scale, figure and word (in top position) appeared at the beginning of each trial.

Variables

The dependent variable of interest consisted of the distance (in pixels) from the centre of the anthropomorphic manikin in its final position to a line just below the word. The distance occupied a range of 600 pixels, from 100 pixels (closest to the word) to 700 pixels (farthest from the word). Participants could move the manikin as many times as they choose by clicking on different points on the line. The only variable of interest was the final position when the participant clicked the 'continue' button. Manikin positions after each click, the number of clicks, and response time for each click were all recorded, but did not shed any additional light on the emotional effects on the avoid-approach decisions and are not reported here.

Critical independent variables were valence and arousal ratings for each word stimulus (Warriner et al., 2013). Valence norms were obtained using a 1 (unhappy, annoyed, melancholic) to 9 (happy, pleasant, satisfied) scale, while the scale used for arousal norms ranged from 1 (calm, sluggish, dull, relaxed) to 9 (excited, aroused, frenzied). As the ease of word recognition is a plausible modulator of any behavioral response to a word, we also included word frequency from the 51 million-token SUBTLEX corpus and word length as statistical controls. Word length did not affect performance in the task and is not reported further.

Statistical Analyses

We used linear mixed-effects multiple regression models with participants and words as crossed random effects (cf., Baayen, Davidson, & Bates, 2008; Pinheiro & Bates, 2000), as implemented in package lme4 version 0.999999-2 (Bates Maechler, Bolker & Walker, 2013) for R version 3.0.1 (R Core Development Team, 2013). This method enables a simultaneous exploration of multiple factors and covariates, while accounting for between-participants and between-items variance. Each model was initially fitted with a maximal random-effects structure (Barr, Levy, Scheepers, & Tily, 2013) and trimmed down to only contain the random effects that significantly improve the model's performance, as indicated by a series of likelihood ratio tests that compared a model with a given random effect and a model without this random effect. Using the same test in the backwards elimination procedure, we removed from the models all fixed effects that did not improve the model's performance. No model reached a harmful level of collinearity (condition index < 13). To reduce the influence of outliers, the frequency estimates were (natural) log-transformed, as indicated by the Box-Cox power transformation test.

Results and Discussion

Although we instructed participants to click on the slider on every trial, even if they wanted to leave the anthropomorphic manikin in the centre, some did not and simply clicked 'Continue'. Participants who failed to register their response more than 25% of the time were removed (see 'Participants'). We removed any trials that were more than 2.5 SD from the participant's mean RT. Doing so removed 2.1% of the data. We further trimmed the data set as a whole by removing remaining trials with exceedingly long and short RT (1% of trials from both extremes of the first click RT distribution). The resulting data pool contained 9,859 trials.

Distance of the manikin from the word was negatively and near-linearly related to the valance of the word (Pearson's r = -0.62, 95% CI [-0.63,-0.61], see Figure 2).

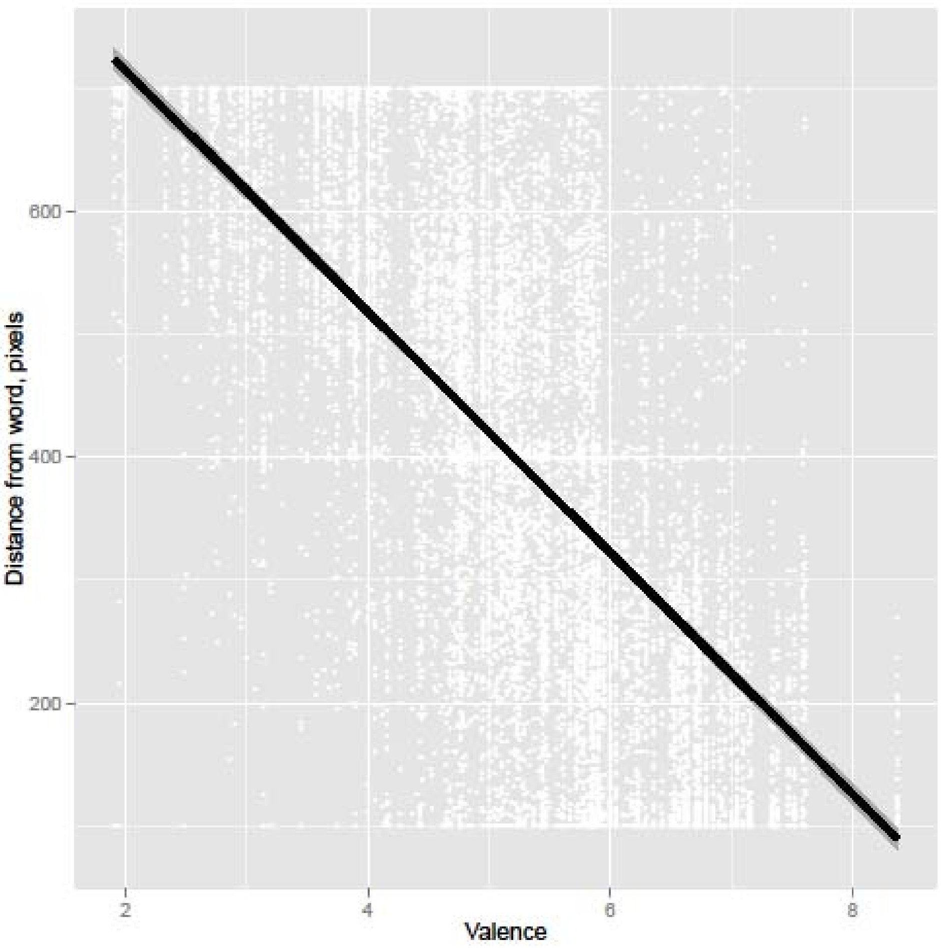


Figure 2: Scatterplot of the manikin's distance from the word as a function of the word's valence. The individual data points are shown in white and the trend line in black, with the 95% confidence interval presented as a gray area.

Responders showed a very strong preference for approaching positive words and withdrawing from negative ones. Each point on the 1-9 valence scale corresponded to about 90 pixels, or 15% of the available distance range. This tendency was confirmed in the linear mixed-effects multiple regression model which estimated the effect of valence on the distance over and above other predictors and individual variability in the overall distance and the strength of valence effect [b = -92.7, SE = 4.8, t = -19.2], see Table 1a for the summary of fixed effects on the manikin's distance from the word.

Arousal did not influence the manikin's distance from the word, nor did it interact with valence to influence that distance (all |t|values < 1.5 in the regression models, not shown). Frequently occurring words were approached more readily than uncommon words (Pearson's r = -0.33, 95% CI[-0.34:-0.31]), even when valence was controlled for in the regression model [b = -14.7, SE = 2.6, t = -5.6].

Table 1a: Fixed effects of the multiple regression model fitted to the distance of the manikin from the word. R^2 of the model is 0.55, and the standard deviation of residual is 133.14.

Predictor	Estimate	SE	t-value
Intercept	965.278	28.534	34.11
Valence	-92.746	4.836	-19.18
log frequency	-14.688	2.628	-5.59

Negative correlations between by-participant adjustments to intercepts and slopes in the random effects structure (Column 3 in Table 1b) pointed to individual variability in behavioral responses. Participants who tended to keep a larger distance from the word overall were more sensitive to the effects of both valence and frequency of the word. For these participants, an increase in one unit of valence (or frequency) translated into a bigger reduction of the distance from the word. The same increase in valence had a weaker effect on participants who maintain a shorter distance from words overall. This observation is consistent with a well-established base-rate effect, whereby a larger magnitude of a response in one condition (i.e., a longer latency, larger amplitude, longer duration) tends to come with a stronger effect (e.g., a larger amount of change) associated with a critical predictor (see Butler & Hains, 1979; Faust, Balota, Spieler, & Ferraro, 1999).

Table 1b: Random effects of the multiple regression model fitted to the distance of the manikin from the word, including random intercepts for participants and items, as well as by-participants random slopes for valence and frequency, and full random correlations.

Random effect	Standard deviation	Correlations between by-participant slopes and intercepts	Correlations between by-participant slopes
by-word intercept	62.063		
by-participant intercept	133.797		
by-participant valence slope	22.220	-0.95	
by-participant frequency slope	3.841	-0.67	0.49

Results of Experiment 1 reveal a strong relation between the word's valence and approach-avoidance behavior measured by slider position. This relation was particularly noteworthy given that our participants responded behaviorally to stimuli that other individuals (those tested in Warriner et al., 2013) evaluated as happy or unpleasant. Thus, the observed patterns validates the slider methodology and supports the generalizability of Warriner et al.'s (2013) ratings over a different population and different task.

Experiment 2

One of the questions we posed was whether the slider task was sufficiently sensitive to group differences in valence ratings. Given gender differences in emotional responses to words attested, among others, by Warriner et al. (2013), we set out to test whether those differences would be mirrored in the approach-avoidance behavior of male versus female participants in the slider task.

Method

Participants

Eighty-seven students at McMaster University participated in this experiment in exchange for partial course credit. None took part in any other experiment. The data from fifteen participants were removed for not making a deliberate response on more than 25% of trials (see Experiment 1). Data from 8 participants were removed for not being native English speakers. Of the remaining 64 participants, 35 were female, 29 male (age range = 17 to 23, M = 18.97, SD = 1.39).

Stimuli

Warriner et al.'s (2013) dataset reports ratings of valence and arousal averaged by gender. For each of the 13,763 words that had frequency information available in the SUBTLEX corpus, we calculated the difference between average male and female ratings for both valence and arousal. We selected 50 words with the most extreme positive and 50 words with the most extreme negative difference scores for valence, i.e. 50 words associated with higher valence ratings from males than females (*beer, gun, topless, hotshot*) and 50 vice versa (*flower, caterer, faith, parent*). We also selected 30 words associated with higher arousal ratings from males than females (*panties, hunting, scuffle, velvet*) and 30 words vice versa (*nerd, limo, skinny, toddler*). To make sure that not all stimuli represent extreme gender differences in valence and arousal, we prepared fillers with no or little gender difference in ratings. For this purpose, the remaining dataset was divided into 25 bins (crossing quintiles of valence and arousal), and 5 words were randomly chosen from each bin for a total of 125. Altogether, the difference and filler words resulted in a set of 285 words. One word was subsequently lost due to a programmatic error. In the final stimulus set, valence and arousal were uncorrelated (ρ= -0.101, p = n.s.) and difference scores for both were approximately normally distributed, as indicated by the Shapiro-Wilk normalcy test. Mean affective ratings for the entire stimulus list and for each subset of words are reported in Table 2.

Table 2. Average ratings across stimuli subsets showing gender differences for both valence and arousal. M = male; F = female; V = valence; A = arousal

	N	Male V	Female V	Male A	Female A	Avg V Diff	Avg A Diff	Avg Log Freq	Avg Length
M happier	50	6.05	3.28	5.43	4.74	2.77	0.69	4.76	7.30
F happier	50	4.85	7.47	4.06	3.93	-2.62	0.13	6.50	6.26
M more aroused	29	5.63	5.00	6.59	3.23	0.62	3.36	4.68	7.90
F more aroused	30	5.47	5.48	2.64	5.30	-0.01	-2.66	5.06	6.43
Remaining	125	5.17	4.92	4.40	4.13	0.24	0.26	4.88	7.72
All words	284	5.34	5.15	4.56	4.24	0.20	0.32	5.14	7.27

Procedure

The procedure for this study was the same as for Experiment 1 with the following differences. First, the words were counterbalanced to appear at the top or the bottom of the slider: once selected, the position for each word was constant across participants. The order of the words was then randomized for each participant. Second, the first 49 participants completed the task on a monitor with a 1024x768 pixel resolution while the last 38 completed it on a monitor with a 1600x900 pixel resolution. There was no difference in responses based on screen resolution, t(607) = -0.18, 95% CI [-17.37, 14.43], Cohen's d = 0.016.

Third, the participants in Experiment 2 rated all words for both valence and arousal via a web-based form, a task counterbalanced with the slider task. The two sets of mean ratings (those in the Warriner et al.'s (2013) norming study and those collected during the experiment) showed a strong correlation (Experiment 2: Valence r = .817, 95% CI [0.77,0.85]; Arousal = .514, 95% CI [0.42,0.59]) and produced a nearly identical pattern of effects. These are comparable to the inter-group correlations (ie. old vs. young, male vs. female, high vs. low education) reported in Warriner at al. (2013) in which Pearson's correlation coefficients for valence ranged from .789 to .831 and for arousal from .467 to .516. For comparability across experiments, we only report the analyses made with the mean ratings from Warriner et al. as independent variables.

Variables

Dependent variables included the distance from the word as chosen by male and female participants. For each word, we also averaged the distance for male and female participants separately, and considered the gender difference between average distances as another dependent variable. The difference ranged from -249 pixels (males closer to the word than females) to 258 (females closer to the words than males)

Independent variables were gender-specific mean ratings of valence and arousal from Warriner et al. (2013). Additionally, we considered the difference in valence and arousal ratings per word as a predictor: positive when a rating given by males was higher (showing a happier, or more excited response) than that given by female raters.

Data Analyses

We trimmed the data in a similar manner to Experiment 1. Participants who did not move the anthropomorphic manikin more than 25% of the time were removed (see 'Participants'). One additional female participant was removed for having an average first click RT greater than 2.5 SD's than the mean of rest of the participants. We removed any trials that were more than 2.5 SD from the mean as calculated by participant. Doing so removed 2.6% of the data. We then trimmed the data set as a whole by removing 1% of trials from both ends of the first click RT distribution. The resulting dataset contained 17,068 trials, with 34 male and 28 female participants.

Results and Discussion

A linear mixed-effects multiple regression model fitted to the manikin's distance from the word as a dependent variable and gender-specific affective ratings as critical predictors replicated all findings of Experiment 1 (model not shown). Higher valence ratings given by either male or female raters in Warriner et al.'s (2013) dataset were related to the tendency of both male and female participants to move the manikin closer to the word. There was no main effect of gender ($t < 0.5$). Similarly, higher frequency words were approached closer regardless of the participant's gender. Additionally, the distance from the word was slightly larger (by 14 pixels) if the word was positioned on the top and not the bottom of the slider. Also, male and female responders were found to both show stronger effects of valence on distance (i.e. changed the manikin position by more pixels in response to the same change in valence) if their distance from the word was overall larger.

The central point of this experiment was to test if the relation between (a) gender differences in valence ratings to words varied along with (b) gender differences in the average manikin's distance to those words. Figure 3 and Table 3 summarize the outcome of the linear multiple regression model (a mixed-effects model was not used as only one value of the dependent variable was associated with each word).

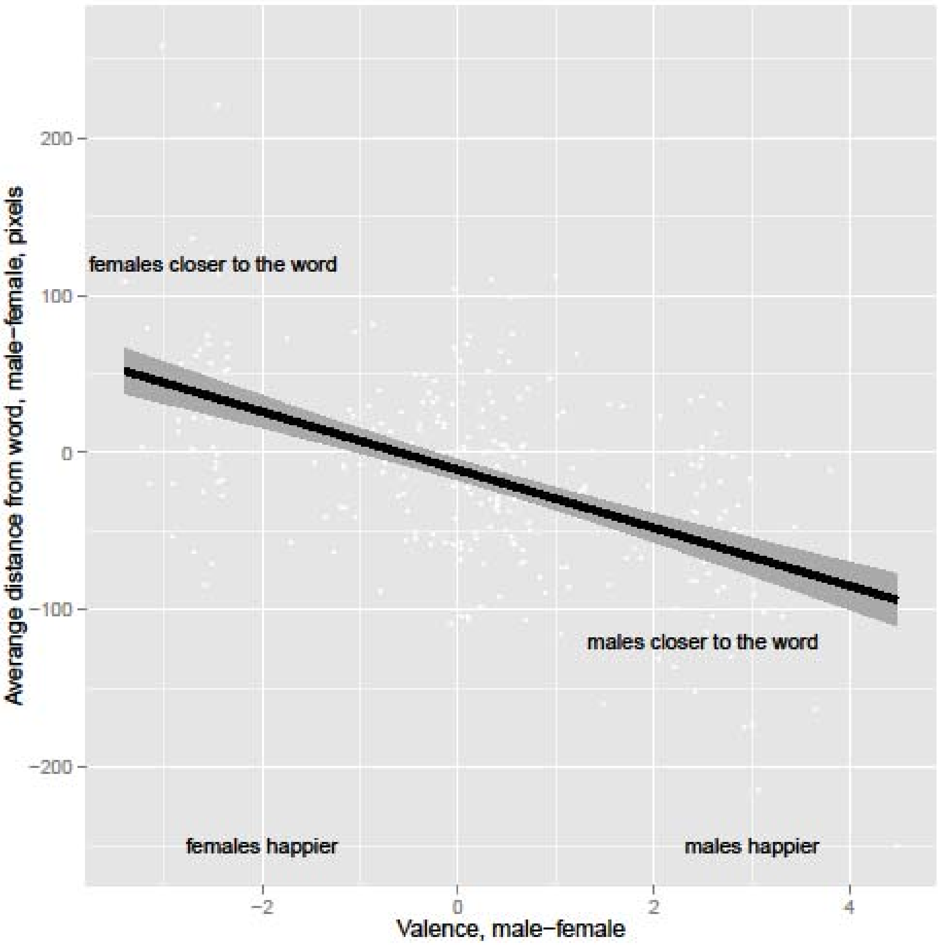


Figure 3. Difference between male and female distance choices as a function of the difference between male and female valence ratings. The trend line is in black with the 95% confidence interval presented as a gray area.

Table 3: Summary of the regression model fitted to the male-female difference in valence ratings with male-female differences in valence and arousal ratings as predictors. $R^2 = 0.29$.

	Estimate	SE	t-value
Intercept	-14.416	3.307	-4.359
Valence difference	-18.098	1.864	-9.710
Arousal difference	-6.230	2.014	-3.093

Figure 3 points to a tendency for participants of one gender to preferentially approach words rated as more pleasant or arousing by raters of the same gender [valence: b = -18.1, SE = 1.9, t = -9.7; arousal: b = -6.2, SE = 2.0, t = -3.1]. On average, women moved the manikin about 18 pixels (or 3% of the 600 pixel range) closer to a word whose valence ratings given by female raters was 1 point higher than that given by male raters (e.g. *adoring, drink, manuscript*). Between the extremes of the gender difference in valence ratings (-3.40 *mommy* to 4.48 *threesome*, where positive numbers indicate higher male ratings), the gender difference in locations reached a substantial magnitude of 139 pixels (or 23% of the available position range).

Table 3 additionally indicates a similar, though weaker, tendency to move the manikin closer to the words judged as more arousing by the same gender. A gender difference in 1 point of arousal ratings to a word came with a difference of about 6 pixels in the distance of the manikin from the word. Between the extremes of the gender difference in arousal (-3.30 *seafood* to 4.25 *musket*), the magnitude of the distance difference was 39 pixels (or 6.5% of the available position range).

The symmetrical nature of the relation in Figure 3 is further confirmed by the value of the intercept: for completely neutral words, i.e. with no gender difference in either valence or arousal, the gender difference in the distance to the word is minimal (only 14 pixels or 2% of the available range). Thus, as suggested by a weak effect of gender in the linear mixed effects model, there is no overall difference between genders in how approaching or avoidant they are in response to emotional stimuli. Both genders reduce distance to pleasant and arousing words, yet – crucially – more so when the words are judged as particularly pleasant or arousing by that gender. We conclude that the slider task is sensitive to group, specifically gender, differences in emotional responses to stimuli.

Experiment 3

We have observed that the position of the manikin relative to a word reflects emotionality of this word, and even gender differences in word emotionality. Our next step was to test the sensitivity of the scale to even more subtle individual differences in personality traits that may influence the participants' biases towards approaching and avoiding (un)pleasant or (non)arousing phenomena. In view of gender differences reported in Experiment 2, we restricted participation to female students only in Experiment 3.

Methods

Participants

Thirty-nine female McMaster University students participated in this experiment in exchange for partial course credit. None of them took part in any other experiment. The data from 4 participants were removed for not making a deliberate response on more than 25% of trials. The data from an initial 4 participants were removed for not being native English speakers. Thirty-one participants remained (age range: 18 to 21, M = 19.03, SD = 1.02).

Stimuli

Stimuli were the same as in Experiment 1.The stimuli from Experiment 2 were chosen with a specific gender skew while the stimuli from Experiment 1 were a balanced representation of the entire affective space. By reusing them in Experiment 3, we could also show a replication of the results from Experiment 1 without any concern that stimuli choice would cause differences.

Procedure & Measures

The procedure for this study was the same as for Experiment 2 except that four additional personality questionnaires were added. One, the Behavioral Approach/Behavioral Inhibition Scale (BAS/BIS; Carver & White, 1994), identifies the degree to which people focus on avoiding punishment, fear, sadness, etc. versus focusing on acquiring rewards and achieving goals while another, the Alexithymia Scale (TAS-20; Bagby, Parker, & Taylor, 1994), measures how strongly people weight sensorimotor information over emotional information (high score) and vice versa (low score). The first is measured on a 4 point Likert scale from “very true for me” to “very false for me” while the second is measured on a 5 point Likert scale from “strongly disagree” to “strongly agree”. The Affective Style Questionnaire (ASQ; Hofmann & Kashdan, 2010) measures people’s tendencies to use three different strategies to handle emotional reactions and the Cheek and Buss Shyness and Sociability Scale (SSS; Cheek, 1983; Cheek & Buss, 1981) includes the five highest loading shyness items (Bruch, Gorsky, Collins, and Berger, 1989) from the original Cheek and Buss (1981) shyness measure and the 5 item sociability scale from the Cheek and Buss (1981) measure. Both were measured on a 5 point Likert scale from “not true of me at all” to “extremely true of me”.

Twenty-nine participants completed all four scales. All subsequent analyses use data from only these participants. Mean scores and ranges for these questionnaires are reported in Table 4. All participants completed the slider portion of the experiment on a monitor with a 1600x900 pixel resolution. As in Experiment 2, we additionally collected affective ratings from participants using a web-based form. The ratings were strongly correlated with the ones in Warriner et al. (2013) and elicited a highly similar pattern of effects. For comparability between experiments, we report the model with Warriner et al.'s ratings as predictors. Altogether the experiment took approximately 1 hour.

Table 4. Mean, maximum, and minimum scores for each of the scales and their subscales.

	Mean	Min	Max
BAS Drive	10.45	7	15
BAS Fun	10.93	6	15
BAS Reward	17.90	15	20
BAS Approach (D+F+R)	39.28	32	47
BAS Inhibit	23.93	15	28
Alexithymia: Feeling	17.76	8	28
Alexithymia: Describing	15.24	10	20
Alexithymia: External	18.00	13	29
Alexithymia: TOTAL	51.00	36	71
ASQ: Concealing	24.72	13	38
ASQ: Adjusting	22.10	10	34
ASQ: Tolerating	16.76	9	22
ASQ: TOTAL	63.59	43	93
Shyness	8.79	1	20
Sociability	12.93	4	20

Data Analyses

We recorded the same information as in the previous two experiments. We removed any trials that were more than 2.5 SD from the mean as calculated by participant. Doing so removed 2.4% of the data. Then we trimmed the data set as a whole by removing 1% of trials from both ends of the first click RT distribution. The remaining data pool contained 7,560 trials.

Results and Discussion

The linear mixed-effects regression model fitted to the manikin's distance from the word replicated effects observed in Experiments 1 and 2 (see Tables 5a, b). Participants moved the manikin closer to relatively positive and more frequent words and increased the distance the more negative or rare the words were. Similarly, the base-rate effect is replicated in the random effect structure of the model (see the negative correlation between individual intercepts and slopes in Table 5b). Participants who tended to maintain a larger rather than a shorter distance from the target word showed a larger amplitude in avoidant or approaching behavior as a function of valence or frequency.

Table 5a: Fixed effects of the multiple regression model fitted to the distance of the manikin from the word. R^2 of the model is 0.58, and the standard deviation of residual is 122.13.

Predictor	Estimate	SE	t-value
Intercept	961.642	31.405	30.621
shyness	2.271	1.239	1.833
valence	-94.862	5.262	-18.029
log frequency	-14.221	2.604	-5.462

Table 5b: Random effects of the multiple regression model fitted to the distance of the manikin from the word, including random intercepts for participants and items, as well as by-participants random slopes for valence and frequency, and full random correlations.

Random effect	Standard deviation	Correlations between the by-participant slope and intercept
by-word intercept	58.587	
by-participant intercept	112.187	
by-participant valence slope	21.648	-0.952

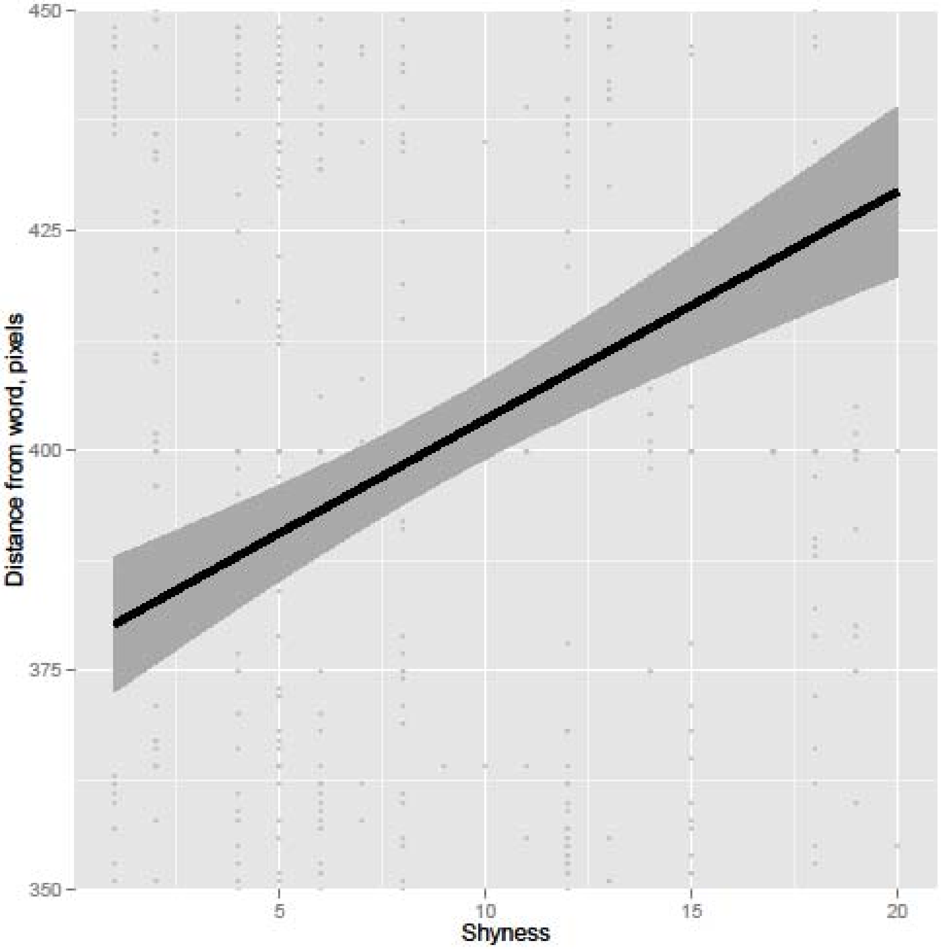


Figure 4. Scatterplot of the manikin's distance from the word as a function of shyness scores. The individual data points are shown in white and the trend line in black, with the 95% confidence interval presented as a gray area.

The battery of personality measures that we conducted revealed main effects of sociability and shyness. As predicted, on average, participants with higher sociability scores tended to move the manikin closer to all words, while those with higher shyness scores placed the manikin at a larger distance [b = 2.3, SE = 1.2, t = 1.8] (Figure 4). Thus, in an average response, a person with the highest shyness score (20) would place the anthropomorphic manikin some 50 pixels farther away from the word than the person with the lowest shyness score (1): the effect range was similar for sociability. Shyness and sociability showed a moderate negative correlation (r = -0.49, 95% CI [-0.51,-0.47]). Due to this degree of collinearity, neither measure showed a reliable effect (i.e. the standard error was similar in magnitude to the regression coefficient) in the model that included both shyness and sociability. Thus, we cannot establish, based on the present sample, whether effects of shyness and sociability are separable and independent. Other tests of individual differences (BAS, ASQ, alexithymia) did not affect the participants' performance in a consistent way. We conclude that the slider task is sensitive to individual differences across selected personality traits, with avoidant behavior more prevalent in individuals who also self-report a preference to socially avoidant behavior.

General Discussion

In all three experiments, we showed that the greater the valence of the stimulus, the smaller the distance people choose from that stimulus. When the stimulus is positive, people approach it and when the stimulus is negative, people withdraw from it, with distance being proportional to the degree of emotionality. While previous studies have shown a congruency effect with faster responses occurring when the movement and stimulus valence were compatible, none have shown that people make graded determinations of approach and avoidance in relation to a stimuli's emotional content (e.g. Chen & Bargh, 1999; De Houwer, Crombez, Baeyens, & Hermans, 2001; Rinck & Becker, 2007).

Using this method, we were also able to identify both group and individual level variability. In Experiment 1, we found that participants who chose to move farther away from a word on average were more responsive to valence. When a stimuli was positive or familiar (higher in frequency), they were willing to move closer to a greater degree than those who tended to move closer to words on average, regardless of their emotionality. A similar individual effect was observed in both Experiment 2 and 3. In Experiment 3, we demonstrated that these differences may be attributable to personality differences. Participants who scored high on shyness tended to stay further away from stimuli while those who scored high in sociability tended to move closer to stimuli.

In Experiment 2, we showed that participants' distance choices paralleled the affective ratings given by their respective gender in Warriner et al. (2013). Females moved closer to words that female participants in a different study had rated more positively and further from words that females had rated less positively. Males moved closer to words that males from a different study had rated more positively and further from words that males had rated less positively. There was a similar, albeit smaller tendency to move closer to words that a participant's same gender had rated as more arousing. There was no average difference between gender meaning that females didn't consistently approach words more than males or vice versa. The difference was specific to those words that were rated differently and in magnitude only, not direction.

Although we argued that one of the advantages of this method was the ability to consider multiple emotional dimensions and lexical features, we did not observe any effects of arousal on distance. The only impact of arousal was in Experiment 2 where each gender chose to move closer to words that their respective gender found more arousing. This effect was small and in the opposite direction of what might be expected if high arousal signifies threat. It is possible that arousal would show a stronger effect in reaction times in this task. Arousal may attract attention and marshal resources (Fernandes, Koji, Dixon, & Aquino, 2011; Vogt, De Houwer, Koster, Van Damme, & Crombez, 2008) without actually changing the ultimate distance. While our instructions encouraged people to answer quickly, our allowance of multiple adjustments to the location of the manikin impeded a precise measure of speed.

We did, however, find an effect of word frequency. In all three experiments, participants moved closer to higher frequency words than to lower frequency ones. This effect was found even after valence was controlled. Perhaps exposure to a word in natural language, indexed by word frequency, lessens the magnitude of one's desire to withdraw from a negative stimulus and increases one's desire to approach a positive one (e.g. Exposure to snakes might make them less scary while exposure to money might make it more desirable).

One advantage of congruency methods is that they are able to measure unconscious associations between valence and approach and avoidance. When participants are instructed to pull a joystick lever towards themselves regardless of the valence of the word displayed, they are not consciously pulling faster in response to positive stimuli and slower in response to negative stimuli. Our method asks participants to make a conscious decision about distance. One could question how they are interpreting this task and whether we are actually measuring approach and avoidance. In addition, the task of choosing a distance from a word becomes less natural the more abstract that word is. For instance, how does one judge how close or far they would like to be to the word *dole* or the word *literacy*? Were participants substituting *close* for virtuous or desirable and *far* for maladaptive or undesirable? If yes, one could argue that we were simply measuring valence and thus, of course distance was correlated. However, if distance and valence are equivalent, we should have found an inverse-U relationship between distance and arousal that is reported in numerous studies (Bradley & Lang, 1999; Warriner et al., 2013) but we didn't. Valence should also have explained far more of the variance in distance responses. As such, distance is separable from but related to valence. It will be important for future studies to explore different ways of operationalizing distance such as actual movement rather than a slider, to consider possible differences between parts of speech, and to perhaps question participants about their interpretations.

Importantly, participants were able to respond and did so in a consistent manner both across studies and within subgroups. Our results show that they do choose to approach or avoid proportionally to valence and not all or nothing. This relates to past research showing that valence in related to other behavioural measures such as lexical decision and word naming in a graded manner (Kuperman et al., 2014). Our finding that distance choices vary in relationship to gender and individual differences also opens up new areas of research. For example, emotional words have been used in Stroop tasks with clinical populations (e.g. People with particular phobias are more hindered in their identification of the color of phobia related words than controls) (Williams, Mathews, & MacLeod, 1996). This slider method may be another way of using emotional words to identify the presence of particular pathologies such as extreme shyness.

Acknowledgments

This work was supported by the Margo Wilson and Martin Daly Ontario Graduate Scholarship to the first author. The last author's contribution was supported in part by the SSHRC Insight Development grant 430-2012-0488, the NSERC Discovery grant 402395-2012, and the NIH R01 HD 073288 (PI Julie A. Van Dyke).

References

Alexopoulos, T. & Ric, F. (2007). The evaluation-behavior link: Direct and beyond valence. *Journal of Experimental Social Psychology, 43*, 1010-1016.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390-412.

Bagby, R. M., Parker, J. D., & Taylor, G. J. (1994). The twenty-item Toronto Alexithymia Scale—I. Item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research*, *38*(1), 23-32.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.0-4. Accessed online: December 2013.

Bruch, M. A., Gorsky, J. M., Collins, T. M., & Berger, P. A. (1989). Shyness and sociability reexamined: A multicomponent analysis. *Journal of Personality and Social Psychology*, *57*(5), 904.

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.

Butler, B., & Hains, S. (1979). Individual differences in word recognition latency. *Memory & Cognition, 7*(2), 68-76.

Cacioppo, J. T., Priester, J. R., & Berntson, G. G. (1993). Rudimentary determinants of attitudes: II. Arm flexion and extension have differential effects on attitudes. *Journal of Personality and Social Psychology*, *65*(1), 5-17.

Carretié, L., Hinojosa, J. A., Martín-Loeches, M., Mercado, F., & Tapia, M. (2004). Automatic attention to emotional stimuli: neural correlates. *Human Brain Mapping*, *22*(4), 290-299.

Carver, C. S. & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology*, *67*(2), 319-333.

Cheek, J. M., & Buss, A. H. (1981). Shyness and sociability. *Journal of Personality and Social Psychology*, *41*(2), 330-339.

Cheek, J. M. (1983). The revised Cheek and Buss shyness scale. *Unpublished manuscript,* Wellesley College, Wellesley, MA.

Chen, M. & Bargh, J.A. (1999). Consequences of automatic evaluation: Immediate behavioural predispositions to approach or avoid the stimulus. *Personality and Social Psychology Bulletin, 25*(2), 214-224.

De Houwer, J., Crombez, G., Baeyens, F., & Hermans, D. (2001). On the generality of the affective Simon effect. *Cognition and Emotion, 15*, 189-206.

Dickson, J. M., & MacLeod, A. K. (2004). Approach and avoidance goals and plans: Their relationship to anxiety and depression. *Cognitive Therapy and Research*, *28*(3), 415-432.

Duckworth, K. L., Bargh, J. A., Garcia, M., & Chaiken, S. (2002). The automatic evaluation of novel stimuli. *Psychological Science*, *13*(6), 513-519.

Elliot, A. J., Gable, S. L., & Mapes, R. R. (2006). Approach and avoidance motivation in the social domain. *Personality and Social Psychology Bulletin*, *32*(3), 378-391.

Faust, M., Balota, D., Spieler, D., & Ferraro, F. (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin, 125*(6), 777-799.

Fernandes, M. A., Koji, S., Dixon, M. J., & Aquino, J. M. (2011). Changing the focus of attention: The interacting effect of valence and arousal. *Visual Cognition*, *19*(9), 1191-1211.

Fishbach, A., & Shah, J. Y. (2006). Self-control in action: Implicit dispositions toward goals and away from temptations. *Journal of Personality and Social Psychology, 90*, 820-832.

Giner-Sorolla, R., García, M. T., & Bargh, J. A. (1999). The automatic evaluation of pictures. *Social Cognition*, *17*(1), 76-96.

Heuer, K., Rinck, M., & Becker, E.S. (2007). Avoidance of emotional facial expressions in social anxiety: The approach-avoidance task. *Behaviour Research and Therapy, 45*, 2990-3001.

Hofmann, S. G., & Kashdan, T. B. (2010). The affective style questionnaire: development and psychometric properties. *Journal of Psychopathology and Behavioral Assessment*, *32*(2), 255-263.

Krieglmeyer, R., De Houwer, J., & Deutsch, R. (2011). How farsighted are behavioural tendencies of approach and avoidance? The effect of stimulus valence on immediate versus ultimate distance change. *Journal of Experimental Social Psychology, 47*(3), 622-627.

Krieglmeyer, R. & Deutsch, R. (2010). Comparing measures of approach-avoidance behaviour: The manikin task vs. two versions of the joystick task. *Cognition & Emotion, 24*, 810-828.

Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. B. (2014). Emotion and Language: Valence and Arousal Affect Word Recognition. *Journal of Experimental Psychology: General, 143*(3), 1065-1081.

Lang, P.J. (1994). The motivational organization of emotion: Affect-reflex connections. In S.H.M. van Goozen, I.E. Van de Poll, & Seargeant, J.A. (Eds.). *Emotions: Essays on Emotion Theory* (pp. 61-96). New York: Psychology Press.

Lang, P. J. (1995). The emotion probe: Studies of motivation and attention. *American Psychologist*, *50*(5), 372-385.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1990). Emotion, attention, and the startle reflex. *Psychological Review*, *97*(3), 377.

Maio, G. R., & Esses, V. M. (2001). The need for affect: Individual differences in the motivation to approach or avoid emotions. *Journal of Personality*, *69*(4), 583-614.

Neumann, R., Hess, M., Schulz, S.M., & Alpers, G.M. (2005). Automatic behavioural responses to valence: Evidence that facial action is facilitated by evaluative processing. *Cognition and Emotion, 19*(4), 499-513.

Osgood, C. E. (1953). *Method and theory in experimental psychology*. New York: Oxford University Press.

Pinheiro, J. C. & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer-Verlag.

Puca, R.M., Rinkenauer, G., & Breidenstein, C. (2006). Individual differences in approach and avoidance movements: How the avoidance motive influences response force. *Journal of Personality, 74*(4), 979-1014.

Rinck, M. & Becker, E. S. (2007). Approach and avoidance in fear of spiders. *Journal of Behavior Therapy and Experimental Psychiatry, 38*, 105-120.

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, *69*(5), 379-399.

Seidel, E., Habel, U., Kirschner, K., Gur, R.C., & Dernti, B. (2010). The impact of facial emotional expressions on behavioral tendencies in females and males. *Journal of Experimental Psychology: Human Perception and Performance, 36*(2), 500-507.

Siebt, B., Neumann, R., Nussinson, R. & Strack, F. (2008). Movement direction or change in distance? Self and object related approach-avoidance motions. *Journal of Experimental Social Psychology, 44*(3), 713-720.

Stins, J.F., Roelofs, K., Villan, J., Kooijman, K., Hagenaars, M.A., & Beek, P.J. (2011). Walk to me when I smile, step back when I'm angry: Emotional faces modulate whole-body approach-avoidance behaviours. *Experimental Brain Research, 212*, 603-611.

van Danztzig, S., Pecher, D. & Zwaan, R.A. (2008). Approach and avoidance as action effects. *The Quarterly Journal of Experimental Psychology, 61*(9), 1298-1306.

Vogt, J., De Houwer, J., Koster, E. H., Van Damme, S., & Crombez, G. (2008). Allocation of spatial attention to emotional stimuli depends upon arousal and not valence. *Emotion*, *8*(6), 880-885.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191-1207.

Wentura, D., Rothermund, K., & Bak, P. (2000). Automatic vigilance: The attention-grabbing power of approach-and avoidance-related social information. *Journal of Personality and Social Psychology*, *78*(6), 1024-1037.

Williams, J. M. G., Mathews, A., & MacLeod, C. (1996). The emotional Stroop task and psychopathology. *Psychological Bulletin*, *120*(1), 3-24.

CHAPTER 6: Emotion is like a boomerang:
Revisiting affective biases in the English language

Warriner, A.B. and Kuperman, V. (in revision). *Cognition and Emotion.*

Abstract

A long-standing observation about the interface between emotion and language is that positive words are used more frequently than negative ones, leading to the Pollyanna hypothesis which alleges a predominantly optimistic outlook in humans. This paper uses the largest available collection of affective ratings as well as insights from linguistics to revisit the Pollyanna hypothesis as it relates to two dimensions of emotion: valence (pleasantness) and arousal (intensity). We identified systematic patterns in the distribution of words over a bi-dimensional affective space, which (a) run counter to and supersede most prior accounts, and (b) differ drastically between word types (unique, distinct words in the lexicon) and word tokens (number of occurrences of available words in the lexicon). We argue for two factors that shape affect in language and society: a pro-social benevolent communication strategy with its emphasis on useful and dangerous phenomena, and the structure of human subjective perception of affect.

Keywords: valence, arousal, lexicon, cognitive bias, subjective experience

Introduction

Researchers have long explored the emotional structure of the lexicon and the emotionality of word choices in individuals as the basis for identifying affective characteristics of individual language users and language communities. For example, older people are shown to use more positive, future tense, and complex words than younger people (Pennebaker & Stone, 2003) and females use more tag questions and intensifiers (Lakoff, 1975) while males use more judgmental adjectives and directives (Mulac, Braden & Gibbons, 2001). Word choice can also reveal personality - the frequency of use of negative words increases with neuroticism while that of positive words increases with extraversion (Pennebaker & King, 1999); emotion – anxiety comes with the use of explainers and negations while anger is related to a lack of qualifiers (Weintraub, 1981, 1989); and affective disorders - depressed individuals use more first person singular pronouns (Bucci & Freedman, 1981; Rude, Gortner, & Pennebaker, 2004). The affect of word choices can even reveal what people wish to hide in that liars can be detected by their increased use of negative words (Newman, Pennebaker, Berry, & Richards, 2002; Vrij, 2000) among other factors. With increased computing power has come the ability to examine the link between linguistic behavior, emotional dynamics and cultural values of large social groups through extensive collections of language samples, both written and spoken. Examples include new research fields such as opinion mining and sentiment analysis: large-scale data-mining studies conducted to measure attitudes towards products and events, through mostly online texts such as websites, forums, and twitter feeds (for a review, see Liu, 2012). Another instance of an emerging relevant field is culturomics which analyzes digital texts, observing cultural trends and their change over time through distributions of word frequencies (Michel et al., 2011).

The present paper takes as its point of onset an observation about language and its cultural and emotional characteristics that informs all research disciplines mentioned above, namely that positive words (those with higher psychological valence) occur more frequently than negative words (those with lower valence). Boucher and Osgood (1969) are credited as being the first to report this observation about the English language: they dubbed the phenomenon the "Pollyanna hypothesis" after the title character of a series of children's novel known for her invariably optimistic outlook. Multiple further studies (see the review below) have confirmed the Pollyanna or the positivity bias across languages and genres. Importantly, however, the use of the term positivity bias has come to conflate two theoretically and empirically distinct phenomena: (a) positive words are used more frequently than negative words, and (b) language has more distinct positive words than distinct negative words: see Kloumann, Danforth, Harris, Bliss, and Dodds (2012), and references below. To rephrase the two observations in linguistic terms, the positivity bias may express itself as (a) a correlation of valence with token frequency, or the number of instances of a given word observed in a sample of text; (b) its correlation with type frequency, or the number of distinct, unique words (regardless how often each of them is used) in a sample of text; or as correlations with both (a) and (b). Below, we cast our review of the prior literature in terms of token and type frequency and their link with positivity. We follow the review with a discussion of the empirical and theoretical implications of this distinction.

Several contemporaries of Boucher and Osgood (1969) confirmed their observation of a correlation between the positivity of a word and its token frequency across languages. For example, a year earlier, Zajonc's (1968) paper on the mere exposure effect noted that affective connotation was related to word frequency. The possibility that positivity affects word type frequency and leads to a prevalence of positive words in language was not considered by Bouchard and Osgood (1969), however, just a few years earlier, Johnson, Thomson and Frincke (1960) reported that the ratio of positive to negative words in their sample was 2:1. In recent years, several researchers have set out to confirm that a positivity bias continues to exist now, forty years later. Augustine, Mehl, and Larsen (2011) used words that had been rated for both valence (negative vs. positive) and arousal (calm vs. excited) in the Affective Norms for English (ANEW) study (Bradley & Lang, 1999). They correlated these ratings with both the small-sample Kucera and Francis (1967; 1.014 million word-token corpus) and a much larger set of frequency norms from the Hyperspace Analogue to Language (HAL; Lund & Burgess, 1996; 160 million word-token corpus). Regardless of which corpus was used, valence was strongly related to token frequency but arousal was not. Thus, despite the fact that the percentages of distinct positive and negative words in the sample were nearly equal (i.e. the type frequencies of positive and negative words were similar), on average positive words appeared far more often (i.e. had more tokens) in the two corpora than negative words. Augustine et al. (2011) extended past findings by showing that the correlation between valence and token frequency was significant across different parts of speech (nouns, verbs, and adjectives), and within a sample of spoken speech. It is worth noting that the results of Augustine et al.'s study, as well as any study using the ANEW ratings of emotionality, may be confounded by the fact that lexical stimuli of the ANEW study were specifically chosen to equally represent the entire affective space, including its extremes, (Bradley & Lang, 1999), rather than represent the distributions of affective words as observed in natural language. Datasets that select stimuli solely based on their frequency of occurrence (e.g. Kloumann et al., 2012 and Warriner, Kuperman & Brysbaert,2013) do not introduce this distributional bias. Rozin, Berman, and Royzman (2010) took a different approach and examined a small set of adjectives and nouns denoting emotional states (disgust, sympathy) by interviewing native speakers of 20 different languages to identify patterns in their use. They determined that there was a distinct advantage for positive words in all studied languages: positive words are more typically unmarked ('not bad' is more common than 'not good') and more likely to be reversed to form their opposite ('unhappy' is more common than 'unsad'). Negated positive adjectives ('not pretty') are viewed as negative while negated negative adjectives ('not ugly') are viewed as neutral. Also when paired, positive adjectives are typically mentioned first (e.g. "good and bad", "pros and contras"). This advantage associated with positive words forms a corollary to the fact that they are observed more frequently; there are more tokens of each type. Notably, Rozin et al. (2010) also report a word-type negative bias: negative emotion labels are more diversified and lexicalized in a larger number of languages as compared to their positive antonyms. That the specific lexical space of words that label emotions (e.g. *pleasure* and *disgust*) contains more negative than positive types of emotion labels is further confirmed in corpus studies and in free-listing experiments (Schrauf & Sanchez, 2004; Semin & Fiedler, 1992; Russell, 1991)[9]. The reported negativity bias in emotion labels appears to run counter to the widely reported positivity bias in the lexicon at large and requires an explanation.

A correlation between positivity and word token frequency was also replicated in written English by Unkelbach et al. (2010) and in Italian adjectives by Suitner and Maas (2008). Garcia, Garas, and Schweitzer (2012) similarly demonstrated a strong positive relationship between valence and token frequency in English, German, and Spanish words. Despite this accumulation of evidence, one significant limitation in all these studies has been the small set of emotional norms upon which the researchers have been able to draw. The

[9] We are indebted to an anonymous reviewer for raising this point.

largest of these studies are those that used the 1,034 English word norms in ANEW (Bradley & Lang, 1999).

Addressing this limitation, Kloumann et al. (2012) considered the top 5000 most frequently used words from each of four different corpora, collecting valence ratings for the resulting list of 10,222 unique word types. They confirmed a positivity bias in all four corpora – Twitter, Google Books, New York Times, and music lyrics – at the word type level. The mean of the positivity ratings of the top 5,000 word types in each of these datasets was greater than the valence scale mid-point, suggesting that more word types are positive than negative. However, being at odds with the majority of previously published work, Kloumann et al. disconfirmed the well-established correlation of positivity with word token frequency, as they only observed a weak relationship between word valence and token frequency in their Twitter and music lyrics sources and an even weaker correlation in Google Books and the New York Times. They concluded that the two dimensions of language use reflected by type- and token-frequency should be considered independent, and that the positive outlook reflects itself in a richer gamut of positive rather than negative experiences, and not in the fact that people tend to mention their positive experiences more often than negative ones. To sum up, what was strong cross-linguistic support for a relationship between emotionality and token frequency has become contested in Kloumann et al. (2012) (do we use a word more frequently if it is more positive?) and the possibility of a type-based positivity bias (more words are positive) received its first strong evidence.

As argued above, the distinction between type and token frequency is theoretically important, as token and type frequencies of occurrence reflect different mechanisms in language use (see examples in Bybee, 2010). If a language’s word-stock is metaphorically construed as a toolkit, with each word as a tool developed to satisfy a specific communicative need, then a higher *type-frequency* of positive words means that there is a need to have more diverse tools to express phenomena related to positive experiences. A higher *token-frequency* of positive words means, however, that the tools existing for relatively positively phenomena are more in demand than those for relatively negative phenomena. Logically, the number of unique tools and how often some of the tools are used might not be related, and so the correlations of positivity with type and token frequency of words need to be considered and interpreted independently in order to characterize the interplay of language and emotion.

Another reason for treating type and token frequency as empirically and theoretically distinct concepts goes beyond the need for factual accuracy in characterizing affective language use. Rather it derives from the fact that they give rise to different causes for the positivity bias. The preference for a larger number of positive words (i.e. the type-based bias) in language may reflect a broader diversity of positive than negative phenomena in cumulative human experience (Gable, Reiss, & Elliot, 2000; Rozin et al., 2010). Alternatively, the prevalence of positive word types may indicate a stronger communicative need to express fine-grained semantic aspects of positive phenomena than negative ones: the Semantic Growth model by Steyvers and Tenenbaum (2005) argues that semantic differentiation of word meanings in language primarily affects words that are in heavy use and is achieved via creation of new word types with more specific meanings. Finally, the type-based bias may be informed by both the emotional structure of human experience and the communicative needs of language users. Interestingly, while Boucher and Osgood (1969) and much subsequent work entertained the prevalence of positive phenomena as an explanation for the positive outlook, very few of these studies have actually explored the type frequency of positivity.

Conversely, the token-based positivity bias, i.e. a more frequent use of a typical positive word, does not reflect the spectrum of emotional possibilities but rather suggests a preferred selection of relatively positive meanings from the available spectrum. This optimistic bias does not only characterize human communication, where it is described as pro-social benevolent behavior (Augustine et al., 2011), but is also observed in the human tendency to make psychological and economic predictions that overestimate reality: e.g. the likelihood of divorce, employment prospects and children's success are often anticipated to be more favorable than is warranted by statistics (see Sharot, 2011 for biological and evolutionary underpinnings of this bias). Intriguingly, while Kloumann et al. (2012) resort to pro-social behavior as an explanation of the distributional patterns in their data, they only find a weak token-based positivity bias, i.e. the correlation that is an index of pro-social behavior.

Apart from revealing the emotional structure of language, the existence of the bias appears to have behavioral consequences. For instance, negative stimuli have an advantage when it comes to grabbing attention, being remembered, and appearing more potent (for reviews, see Rozin & Royzman, 2001; Baumeister, Bratslavsky, Finkenauer, & Vohs 2001). However, positive information tends to be classified, evaluated and responded to faster (Bargh, Chaiken, Govender, & Pratto, 1992; Unkelbach et al., 2010, Kuperman, Estes, Brysbaert, & Warriner, in press). The linguistic positivity bias with its possible and separable links between psychological valence and word type and token frequencies suggests a unifying explanation. Negative word types are rarer and hence more marked which makes them stand out (Clark & Clark, 1977) and leads to improved attention and memory performance for negative words. Positive words have a higher token-frequency and thus carry less information (Garcia, et al., 2012), are more densely clustered in the lexicon, and are therefore privileged by faster and greater spreading activation (Unkelbach, Fiedler, Bayer, Stegmuller, & Danner, 2008; Unkelbach, 2012). A potential explanatory power of the positivity bias requires understanding of the nature of this systematic relationship between language and emotion, and particularly the aspects of linguistic use that it affects: lexical diversity (roughly corresponding to type frequency), lexical entrenchment (reflected in word token frequency) or both. Our first goal is to utilize a recently collected, large set of emotional norms for 13,915 words (Warriner et al., 2013) to definitively characterize an affective bias in the English language and determine its relationship to word type and token frequency, and its implications for the emotional tenor of the society.

Emotion, however, is more than negativity vs. positivity. An influential view on the structure of emotion holds that emotion is dimensional and that the two primary dimensions are those of valence (a negative vs. positive emotional state) and arousal (a calm vs. excited state: for reviews, see Barrett and Russell (1999); Fontaine, Sherer, Roesch, and Ellsworth (2007); and Power (2006). Furthermore, while proposals vary widely as to the nature of the relationship between these dimensions, arousal is often (partly or wholly) associated with the intensity of pleasure or displeasure that one experiences in response to a stimulus (for a review of theories, see Kuppens, Tuerlinckx, Russell, & Barrett, 2013). Most findings in the realm of subjective experience – i.e. self-reports of affect elicited, among others, in the form of ratings – support a characteristic “boomerang”-shaped functional relationship between valence and arousal (Bradley & Lang, 2007, 2009). Namely, stimuli subjectively perceived as very negative or very positive tend to come with a high level of arousal, while mildly positive or negative stimuli are perceived as relatively calm. The boomerang- or the U-shaped relationship is found in aggregate ratings to a broad range of stimuli types, including emotionally laden pictures, affective experience in daily life, current and remembered affective experiences (Kuppens et al., 2013), and, importantly, words (cf. Bradley & Lang, 1999; Redondo, Fraga, Padrón, & Comesaña, 2007; Soares, Comesaña, Pinheiro, Simões, & Frade, 2012; Warriner et al., 2013). While opposing accounts exist, the psychobiological basis of this relationship is thought to be that subjective valence of a stimulus engages one of two motivational subsystems: an appetitive/positive one geared towards attaining objects beneficial for survival (e.g. sustenance, nurturance, and caregiving), and an aversive/negative one associated with response to threat and danger (Bradley, 2000; Bradley & Lang, 2000). Arousal then is a metric of how strongly these systems are activated and how much effort needs to be mobilized to respond to environmental demands (for early proposals see Duffy, 1951; Kahneman, 1973). The effort is maximal in extremely positive or negative cases and minimal when stimuli are neutral and the activation level of motivational systems is low (Kuppens et al., 2013).

The well-established systematic relationship between valence and arousal in indices that aggregate subjective experience of affect across multiple individuals naturally raises the following questions: Does language have a larger number of calm than exciting words, and are exciting words used more frequently than their calmer counterparts? Do distributional patterns of positivity vary by the level of arousal? Are bi-dimensional patterns different for word types and tokens? Our second goal then is to explore – using this same large set of norms – the possibility of a type-based and a token-based bias with regards to the arousal of words in English, and ultimately a compound bi-dimensional distributional bias influenced both by valence and arousal.

Method

Emotional Ratings

We used Warriner et al.'s (2013) collection of valence and arousal ratings for 13,915 lemmas, or vocabulary word forms. For instance, *sing* is a lemma for such inflected word forms as *sing*, *sang*, *sung* and *singing*: thus a lemma merges grammatical variants of the word, which are not expected to vary emotionally. These ratings were collected using the crowd-sourcing online Amazon Mechanical Turk platform (Schnoebelen & Kuperman, 2010) and validated via correlations with previously collected ratings. Participants were drawn from native English speaking residents of the U.S. Each was assigned to rate a sample of words on a 9 point scale ranging either from sad (1) to happy (9) or from calm (1) to excited (9). At least 18 participants rated the majority of words and overall means per word were calculated, one for valence and one for arousal. The words were selected on the basis of their familiarity to a typical English speaker. They were drawn from the list of 30,000 words that were indicated as known by at least 70% or more participants in the study by Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012). All lemmas represented content words (nouns, verbs, and adjectives) rather than function words ("the", "this", "from"). For overall analyses in the current paper, we used the 13,763 word types from Warriner et al. (2013) for which there was frequency information in the 51 million-token SUBTLEX-US corpus based on subtitles to US films and media (Brysbaert & New, 2009): altogether, the word types chosen accounted for 14.5 million tokens.

Other Corpora

We used multiple additional corpora in order to explore genre, regional and age variability, and thus validate the generalizability of our findings. Type and token frequency measures were drawn from these corpora and correlated with the emotional ratings from Warriner et al. (2013). We analyzed the following corpora: the Corpus of Contemporary American English (COCA; Davies, 2009) – a 450 million token balanced corpus including spoken, fiction, magazines, newspapers, and academic texts from 1990 to 2012; the British National Corpus (BNC; Oxford University Computing Services, 2007) – a 100 million token collection of the UK-based written and spoken language from the 1990s; the Touchstone Applied Science Associates Corpus (TASA; Zeno, Ivens, Millard, & Duvvuri, 1995) – 12 million tokens from textbooks, literature and novels separated by cumulative grade level from grade 3 to college; and the Hyperspace Analogue to Language (HAL; Lund & Burgess, 1996) frequency norms – roughly 160 million token gathered from Usenet groups in 1995. The number of words that overlapped between each corpus and the rating study are listed in Column N of Table 1.

Table 1: Summary of type and token frequencies for various genres. Column "Percent positive" reports the percent of words above the midpoint of the positivity scale. Columns V ρ and A ρ report Spearman's correlations between the emotional ratings from the Warriner et al. (2013) study and the word token frequency of that particular corpus. Column N indicates how many words in that corpus had emotional ratings available in Warriner et al. (2013). Overall refers to the SUBTLEX corpus.

Dataset	Percent positive	Median valence	Skewness valence	Vρ	Median arousal	Skewness arousal	Aρ	N
Overall	55.6	5.20	-0.29	0.180	4.11	0.51	0.039	13,763
Overall - Nouns	58.5	5.25	-0.42	0.172	4.05	0.57	0.044	8,202
Overall – Verbs	52.3	5.00	-0.01	0.236	4.27	0.35	-0.007	1,753
Overall - Adjs	49.1	5.10	-0.21	0.149	4.15	0.47	0.086	3,129
COCA	55.6	5.20	-0.29	0.247	4.11	0.51	-0.053	13,762
COCA – Fiction	71.5	5.67	-0.59	0.222	4.05	0.55	-0.109	3,158

COCA – Magazines	76.7	5.72	-0.72	0.167	4.00	0.61	-0.069	3,472
COCA – Spoken	69.6	5.59	-0.64	0.126	4.14	0.53	-0.043	3,424
COCA – News	73.3	5.68	-0.70	0.135	4.05	0.57	-0.046	3,383
COCA – Academic	74.4	5.61	-0.72	0.132	4.00	0.61	-0.032	3,302
BNC	56.4	5.21	-0.29	0.242	4.05	0.54	-0.068	7,823
TASA – Grade 3	60.3	5.32	-0.36	0.209	4.05	0.50	-0.093	8,301
TASA – Grade 6	57.9	5.25	-0.32	0.231	4.09	0.51	-0.101	11,185
TASA – Grade 9	57.3	5.24	-0.31	0.233	4.09	0.51	-0.098	11,693
TASA – Grade 12	56.8	5.23	-0.30	0.241	4.10	0.51	-0.105	12,135
TASA - College	56.4	5.21	-0.29	0.235	4.10	0.51	-0.104	12,344
HAL	55.9	5.20	-0.29	0.209	4.10	0.51	-0.037	12,952

Results and Discussion

For comparability with prior research, we first report the results of considering the valence and arousal biases independently. We then proceed to analyses of the compound valence-and-arousal bias.

Independent positivity and arousal biases: Type Frequency

In our overall dataset (words and frequency drawn from SUBTLEX-US), the distribution of valence ratings showed a longer left tail and a slight negative skew (-0.29). Both the mean (5.06) and the median (5.20) were above 5, which is the midpoint of the scale used. Overall, 55.6% of the words were positive, i.e. had valence ratings above the scale's midpoint, confirming the widely-reported positivity type bias, see the histogram in Figure 1. We also confirmed that the positivity bias holds when a subset of more extreme valence ratings is considered, rather than the entire dataset with numerous valence-neutral items: the number of words with valence above 6 was 10% larger than the number of words with valence below 4 (3106 vs 2869).

By contrast, the distribution of arousal ratings showed a strong positive skewness (0.51), indicating that the majority of the probability mass is concentrated on the lower end of the 1-9 arousal scale, see Figure 1. The mean and median ratings of arousal were 4.21 and 4.11 respectively. These patterns point to the preference towards calmer rather than more exciting words. We do not draw a comparison of arousal ratings against this scale's midpoint, since in the case of a unipolar psychological dimension like arousal, this point is not informative.

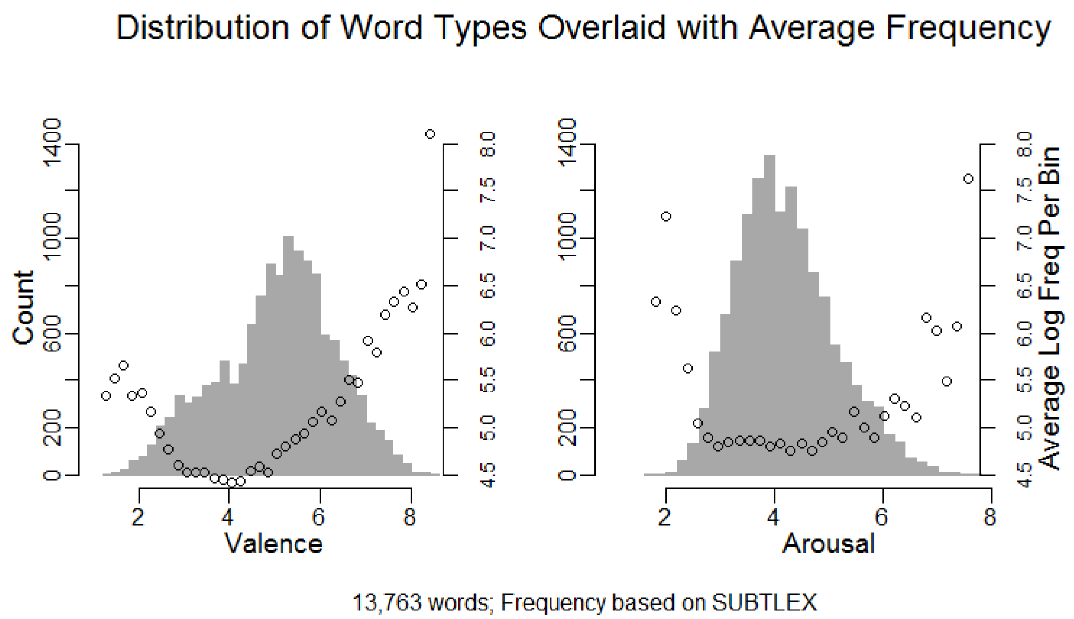


Figure 1. Histograms of valence (left) and arousal (right) for the 13,763 words from Warriner et al. (2013). The overlaid points represent the average log frequency for each histogram bin.

Skewness in similar directions was found in the other corpora examined (see Table 1 for a summary), which is not surprising given how many words overlapped across corpora. UK English (BNC) and US English (HAL and COCA) showed nearly equivalent type biases (i.e. median and level of skewness almost identical) for both valence and arousal. To test variability across genres, we further created subsets of 4000 words representing most frequent nouns, verbs and adjectives for each of the five genres of COCA. Arousal medians and skewness in each of the genres were nearly equivalent to that of COCA as a whole. Together these similarities suggest the robustness of the positivity and calmness type-based biases across dialects and genres. The biases were also robust developmentally, i.e. across grade levels of the TASA corpus: nearly identical skewness was found for both arousal and valence with a small numerical decrease in magnitude for valence as grade level increases.

We also examined how the type bias was manifested within the three parts of speech we selected from SUBTLEX-US. For nouns (N = 8,802), there was a significant positive type bias with 58.5% being positive (χ^2 = 254.38, p < .0001). For verbs (N = 1,753), the positive bias was only marginally significant with 52.3% being positive (χ^2 = 3.71, p = .054), while for adjectives (N = 3.129), there was no type bias in any direction with only 49.1% being positive (χ^2 = 1.01, p = .314). For all parts of speech, distributions of arousal were positively skewed, indicating a strong bias towards calmer word types. This skew was strongest in nouns (0.57) followed by adjectives (0.47) and then verbs (0.35).

To sum up, the percentage of positive words ranged across corpora from 55.6 to 76.7. Thus, we confirm across corpora the previously reported tendency (Kloumann et al., 2012) for English to have a larger diversity of word types with positive rather than negative connotations. We further determined that this type bias is more nuanced than reported earlier, in that it is present mostly within nouns, and somewhat within verbs. There is no positivity type bias for adjectives. Given the prototypical semantic roles of parts of speech in a language, this suggests that the positivity bias is largely driven by a stronger representation of positive rather than negative objects (nouns) in the word stock. There is however a roughly equivalent representation of positive and negative actions and states (verbs), and object qualities (adjectives). Some of our findings run counter to results reported by Kloumann et al.'s (2012), which offers a similar statistical power and diversity of written genres of English. We address methodological reasons for these discrepancies in the Appendix.

We report for the first time a strong and consistent tendency for English to have a greater variety of words for expressing *calm* things across all parts of speech, and to offer a larger toolkit for expressing calm-evoking objects, actions, and traits.

Independent positivity and arousal biases: Token Frequency

In Warriner et al.'s (2013) dataset, we found the Spearman's correlation between valence and log SUBTLEX-US frequency to be $\rho = 0.180$, i.e. stronger than the correlations reported in Kloumann et al. (2012), see Table 1 and Figure 1. As a combined result of the positivity bias observed both in word types and tokens, the prevalence of positive word tokens is decisive: there are 4.4 times more words with valence above the mid-point than those below the mid-point, and 1.7 times more words with valence above 6 than those with valence below 4. For completeness, we note a slight increase in token frequency in very negative words, along with a stronger and steeper increase in token frequency of relatively positive words, see Figure 1. Our examination of other corpora showed similar patterns of correlations (see Table 1 for a summary). The correlation between valence and token frequency ranged from $\rho = 0.126$ to $\rho = 0.247$.

We further observed a negligible Spearman's correlation between arousal and log SUBTLEX-US frequency (ρ = 0.039). Similarly, correlations between arousal and token frequency across corpora were weak, fluctuated in polarity and ranged between ρ = -0.109 to ρ = 0.034. This apparent null effect in line with the finding of Augustine et al. (2011). However, a closer look at the distribution of word tokens over the range of arousal revealed a quadratic functional relationship, with a steep increase in token frequency in very calm and very exciting words, see Figure 1. A similar pattern was confirmed across corpora. Given its symmetrical nature, the zero-order correlation yields a value that is close to zero: yet this apparent null effect is merely an artifact of imposing linearity on a nonlinear relationship between variables (see Kuppens et al., 2013 for elaboration of this methodological point).

To summarize, we observed consistent positive correlations between word token frequency and its emotional valence. The relationship between word token frequency and word arousal was also systematic and had an U quadratic shape.

Interactions Between Positivity and Frequency

It is a logical possibility that the type- and token-based biases towards positive and calm words are specific to certain frequency ranges and would not be observed if very uncommon words were taken out of consideration. We ruled out this possibility by supplementing the analyses of the entire (100%) word list of Warriner et al. (2013) with the analyses of words in the top 90%, 80%, 70% … 10% of the dataset's frequency range. For each subset, Table 2 reports skewness coefficients and medians for the type-based distributions of valence and arousal, as well as the percent of positive words, and Spearman's correlation coefficients for valence and log token frequency. Correlations between arousal and log token frequency are not reported, as they offer a poor fit to nonlinear relationship between the two variables.

Table 2. Type and token frequencies per cumulative frequency bin. Column "Percent positive" reports the percent of words above the midpoint of the positivity scale. Column V ρ reports Spearman's correlations between the emotional ratings from the Warriner et al. (2013) study and SUBTLEX word token frequency. Column N indicates how many words are in each bin.

Percent of dataset	Percent positive	Median valence	Skewness valence	Vρ	Median arousal	Skewness arousal	N
100	0.56	5.20	-0.29	0.18	4.11	0.51	13763
90	0.57	5.23	-0.32	0.18	4.14	0.50	12272
80	0.58	5.25	-0.33	0.18	4.14	0.50	10994
70	0.59	5.29	-0.36	0.18	4.14	0.49	9563
60	0.60	5.32	-0.38	0.19	4.14	0.49	8255
50	0.62	5.37	-0.43	0.18	4.14	0.49	6873
40	0.64	5.44	-0.49	0.16	4.13	0.49	5497
30	0.67	5.52	-0.54	0.16	4.13	0.49	4127
20	0.69	5.61	-0.61	0.16	4.14	0.51	2753
10	0.74	5.78	-0.74	0.11	4.15	0.49	1374

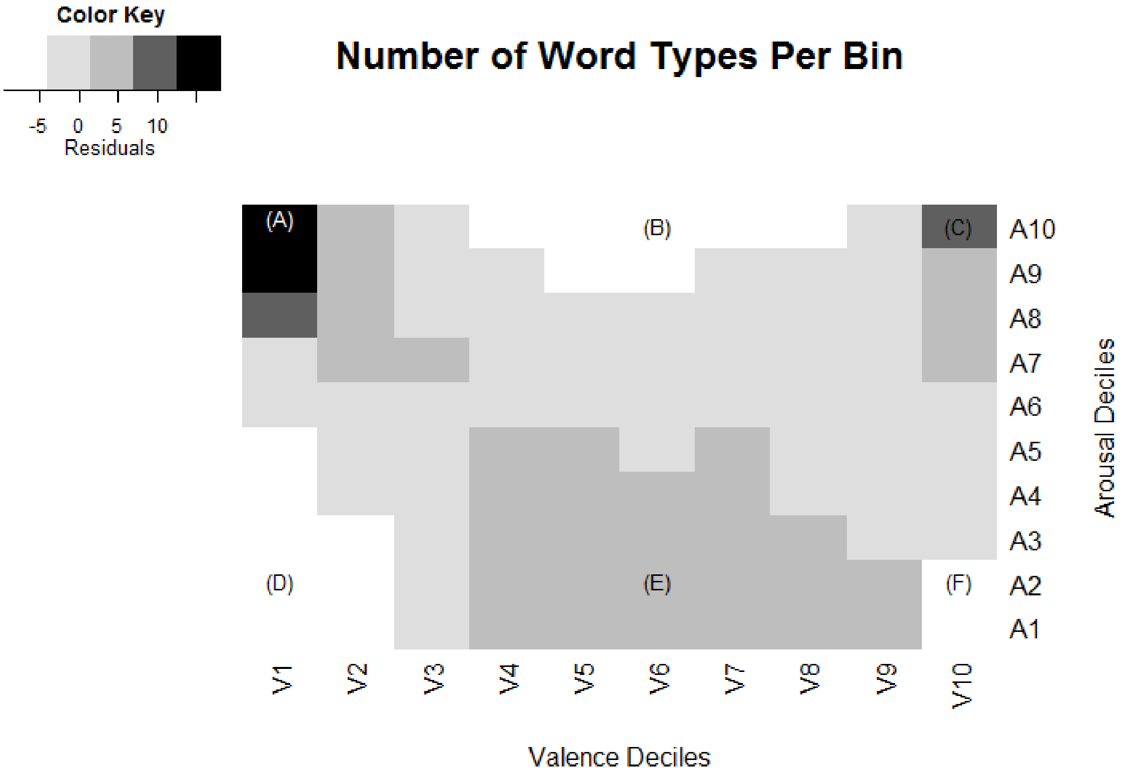
Table 2 confirms all observations made on the entire data set of Warriner et al. (2013). The number of positive words (valence above 5) steadily increases as word frequency increases, and so does the median and absolute value of the skewness of the valence distribution over types. Taken together, these indices suggest that word types become more positive (and the left tail of their distribution becomes longer) in the higher frequency ranges. The token-based positivity bias remains stable over the entire dataset, with only a minor non-significant decrease in the magnitude of the correlation between word valence and log frequency in the top decile of the frequency distribution. That is, the token-based positivity bias is not driven by exceedingly rare words or even words in the mid-range of frequency.

Finally, a strong type-based bias towards calmer words is confirmed by the virtually invariant estimates of the median and skewness of arousal rating.

Compound affective bias: types

As discussed in the Introduction, valence co-varies with arousal in a number of tasks that elicit self-reports of subjective experience, including affective ratings to words. The relationship forms a characteristic boomerang functional curve, with higher arousal accompanying extreme values of valence. It is plausible that the observed biases towards positivity and calmness only partially reflect an overarching bias that spans over both affective dimensions. To explore the distribution of unique words over a bi-dimensional affective space, we calculated the number of word types for each of 100 bins formed by the crossing of deciles of arousal and valence. Figure 2 is a heatmap of residuals of the chi-squared test associated with each bin.

Figure 2. A heat map showing chi square residuals associated with 100 bins formed by crossing valence with arousal deciles. Regions of particular interest were identified with letters. The table below provides examples of words pulled from each region.



(A)	(B)	(C)	(D)	(E)	(F)
abuse	airplane	affection	abandon	bathroom	comfortable
casket	classified	caffeine	constipation	cauliflower	daydream
gunfire	emotional	enthusiastic	drab	daycare	floral
homophobic	freeway	laugh	gurney	foam	grandma
incurable	hyperspace	money	lazy	intercom	honesty
lice	lizard	orgasm	nutcase	liquid	lake
messy	masculinity	pretzel	penalty	northern	meditation
obscene	news	speedboat	scab	penicillin	savings
pigheaded	premonition	thrill	telethon	recycle	vineyard
wasp	striptease	youthful	unromantic	technician	wish

Resulting patterns reveal that the distribution of word types over valence ratings is strongly modulated by how arousing those words are. Regions of the affective space that accumulate most of the English word-stock are extremely arousing and valenced (positive or negative) words. There is also a large lexical space of calmer words that are relatively neutral in their pleasantness. As such, the word-type distribution in Figure 2 faithfully replicates the boomerang functional curve observed in Bradley and Lang (1999) and confirmed since in multiple studies, including the present Warriner et al.' s set of affective ratings. As more valenced words tend to be more arousing due to a stronger activation of motivational systems, it is not surprising that arousing valenced word types are more prevalent in the language than calm valenced words. Similarly, the tendency for neutrally valenced words to only weakly engage motivational systems and thus elicit lower levels of arousal translates into a larger number of calm valence-neutral words than exciting valence-neutral words. The visual patterns in Figure 2 are corroborated by the prediction of a generalized additive model with word type frequency as a dependent variable and the tensor product of valence and arousal as a predictor. The model (not shown) fitted a complex hyperbolic surface to the observed data in the three-dimensional space with arousal, valence and word type frequency as axes and confirmed the statistical significance ($p < 0.0001$) of a surface that reaches maxima of type frequency in extremely arousing extremely valenced region of the affective space, as well as in the region characterized by low arousal and relatively neutral valence.

The replication of the boomerang shape in the distribution of word types (Figure 2) highlights the inadequacy of considering an inherently bi-dimensional affective phenomenon as a uni-dimensional one. Columns 1-3 in Table 3 illustrate this point by showing how (a) a well-established positivity bias in word types holds true for 60% of words with lower arousal (there are significantly more words with valence > 5 as indicated by the proportion test), (b) the bias is not observed in either direction in words falling into deciles 7 and 8 of arousal, and (c) the bias reverses in highly arousing words leading to an advantage to negative word types in this corner of the affective space.

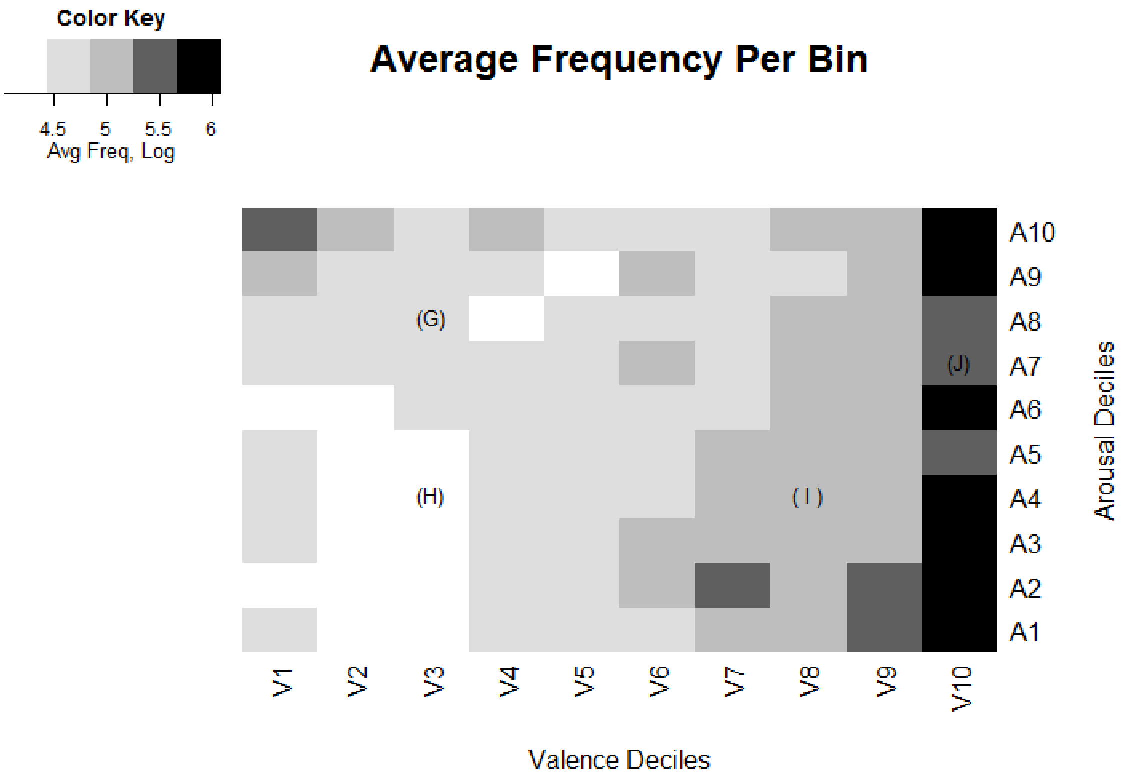
Table 3: Valence type bias per arousal decile and arousal type bias per valence decile. Columns 1-4 report sample size, proportion of positive word types (valence > 5) and Spearman's correlation of positivity with token frequency per decile of arousal. Values of Proportion positive column in italics represent deciles without a significant valence bias, while values in bold represent deciles with a negativity bias as indicated by the proportion test. Columns 5-7 report sample size, and Spearman's correlation of arousal with token frequency by decile of valence.

Arousal decile	N	Proportion positive	V ρ	Valence decile	N	A ρ
1	1337	0.69	0.22	1	1377	0.22
2	1404	0.64	0.27	2	1376	0.16
3	1370	0.64	0.23	3	1321	0.13
4	1364	0.60	0.24	4	1414	0.04
5	1381	0.58	0.22	5	1393	-0.00
6	1309	0.56	0.20	6	1357	-0.02
7	1435	*0.52*	0.19	7	1374	-0.06
8	1366	*0.49*	0.18	8	1367	-0.05
9	1405	**0.44**	0.12	9	1393	-0.05
10	1392	**0.41**	0.06	10	1391	-0.02

Compound affective bias: tokens

Equally nuanced is the token-based positivity bias. Figure 3 reports average log-transformed token frequency for the 100 bins formed by crossing deciles of valence and arousal.

Figure 3. A heat map showing average log frequency per 100 bins formed by crossing valence with arousal deciles. The average log frequency of the words in each bin was calculated and plotted according to the color key provided. Regions of particular interest, complementary to those in Figure 2, were identified with letters. The table below provides examples of words pulled from each region.



(G)	(H)	(I)	(J)
belch	absent	commodity	buddy
chili	collagen	duet	charm
drivel	diet	gown	deserve
eyewitness	electoral	informative	educate
firewall	flimsy	ketchup	luck
incorrigible	gutter	lumberjack	nacho
midterm	overcast	mammal	recipe
offense	renal	person	shiny
socialism	unlisted	route	waterfall
wrinkle	vertigo	waterproof	yummy

Figure 3 points to highly positive words (deciles 9 and 10) across all arousal levels as the area of the lexicon that is in the most active use in communication. This area is complemented by high token-frequency words associated with extremely low valence and extremely high arousal, i.e. danger words (see selection (A) in Figure 2). This spike in words reflecting danger may partly explain the paradoxical discrepancy between an overall positivity bias in word types of the entire language, and a negativity bias observed in words serving as emotion labels (Russel, 1991; Schrauf & Sanchez, 2004; Semin & Fiedler, 1992). If a larger number of emotion labels reflect higher arousal states, the preponderance of negative labels is compatible with our findings.

As discussed above, the bi-dimensional patterns further point to a limitation of any study that approaches characterization of affect in language through the lens of any one dimension. Columns 4 and 7 in Table 3 demonstrate a drastic change in the correlation between valence and token frequency estimated per arousal decile (invariably positive but substantially weakening in higher-arousal words) as well as the correlation between arousal and token frequency (negative and much stronger in negative words than in positive ones).

General Discussion

The present study explores two novel aspects of the relationship between emotion and language using the largest available dataset of affective ratings (Warriner et al., 2013). First, we revisit the widely reported positivity bias or Pollyanna hypothesis -- the prevalence of positive over negative words in language – using the linguistically and psychologically motivated distinction between token frequency (the number of occurrences of a specific word in a language sample) and type frequency (the number of different words in a language sample), see (Clark & Clark, 1977; Semin & Fiedler, 1992). Types can be thought of as different tools available for a certain purpose (in this case, for expressing meanings), while tokens can be thought of as how often a given tool is used (how often a certain meaning needs to be expressed). The number of tools and how often specific tools are used are independent metrics of language use, and so identifying a bias in one does not presuppose nor exclude a bias in the other. With these two ideas firmly separated, we report distributional patterns of affect over the word-stock of English using a broad multi-genre selection of corpora and a new comprehensive set of affective ratings (Warriner et al., 2013). The second novel aspect of our study is the expansion of the traditional uni-dimensional view of affective bias in language as a function of valence to the bi-dimensional view, in which both valence and arousal play a role. In the next two sections, we summarize our findings for distributional biases observed in affective dimensions of valence and arousal considered independently, and then we discuss the bi-dimensional affective bias as an overarching pattern.

Independent biases

A consideration of the word's valence as an independent sole affective dimension -- in line with prior literature – reveals that there are a larger number of positive words (above the mid-scale of valence) in the English language, and overall positive words tend to be in use more often, Figure 1. Thus, we confirm both the type-based positivity bias (in agreement with Kloumann et al., 2012) and the token-based positivity bias (contra Kloumann et al., 2012, and in agreement with Boucher and Osgood, 1969 and subsequent cross-linguistic reports): for treatment of discrepancy with Kloumann's et al.'s findings see Appendix. The type-based trend towards positivity is further qualified by the fact that the share of positive lexical items significantly exceeds 50% only in English nouns (58.2%), while positive and negative verbs and adjectives tend to be equal in number. If we consider the type-based bias as an indication of the variety of phenomena characterizing human experience, the prevalence of positive phenomena is due to words denoting objects (nouns), but not states, motions, or qualities (verbs or adjectives). Regardless of the part of speech, English speakers draw upon the positive words more often than the negative ones: while there is a spike in token frequency in very negative words, very happy words are still in more frequent use.

We found, for the first time, a strong type-based bias towards calm words in the English language, observed in all corpora and for all parts-of-speech, see Figure 1. As such, English speakers have far more calm words available to them than arousing ones. A correlational analysis suggested the absence of a noticeable token frequency biased associated with arousal, in line with Augustine et al. (2011). This is, however, only reflects an inability of linear zero-order correlations to approximate nonlinear curves (Kuppens et al., 2013). In fact, there is a symmetrical U shape of the relationship between token frequency and arousal, with very calm and very exciting words being similarly frequent.

Bi-dimensional affective bias: word types and tokens

Importantly, distributional biases along independent dimensions of affect are superseded by our identification of systematic patterns in the distribution of words over a bi-dimensional affective space formed by valence and arousal. The distribution of word types over valence and arousal axes follows a characteristic boomerang shape, with arousal increasing with extremity of valence (positive or negative). Most word types concentrated around the following regions of affect: high-valence high-arousal (exhilaration, sexual gratification), low-valence high-arousal (danger, threat), and mid-valence low-arousal (emotionally unmarked phenomena), see Figure 2. Regions that are under-represented in word types are those of low-valence low-arousal (depression, shame), high-valence low-arousal (serenity, comfort) and mid-valence high-arousal (see selection (B) in Figure 2). The tendency of valenced words to elicit higher arousal is consistent with a view that arousal is a measure of how strongly the perceived valence of the stimulus engages motivational aversive and appetitive systems and what level of energy the organism must mobilize to respond to the environmental need associated with the stimulus (Bradley, Codispoti, Cuthbert, & Lang, 2001; Duffy, 1951; Higgins, 2006; Lang, Bradley, & Cuthbert, 1990, among others). Furthermore, the boomerang shape of the functional relationship that we observe in our data is a hallmark of subjective judgments of affect and is robustly found in self-reports or aggregated ratings of, for instance, recent and distant emotional memories, current emotional states, emotionally laden and balanced pictures and, finally, words (Bradley & Lang, 1999; Kuppens et al., 2013).

Pitted against prior reports, our data patterns suggest that the word-stock of the English language provides more unique tools (i.e. word types) for the regions of the affective space that are favored in human subjective experience. Put differently, an increased communicative need to express certain experiences – the ones that link more extreme valence with higher arousal – appears to have prompted a more extensive creation and diversification of lexical items denoting those experiences. Remarkably, distributional patterns obtained from large corpora of English, which summarize the collective verbal behavior of millions of speakers, dovetail perfectly with results of laboratory studies eliciting highly constrained responses (typically, ratings) to a hand-picked set of stimuli obtained from a much smaller number of participants[10]. It is noteworthy that the convergence takes place even though our stimuli were selected without regard for emotionality and thus represent language in a

[10] Kuppens et al. (2013) point out that the U- or a V-shaped relationship observed in ratings aggregated over hundreds or thousands individuals co-exists with very large individual variability in the relationship between arousal and valence. We are not able to test this between-levels difference as individual frequency distributions are not available to us.

naturalistic way, in stark contrast with prior studies offering either a carefully balanced representation of affective language or an over-representation of its extremes.

The availability of tools for expressing affect in the English language does not dictate how these tools are used: i.e., the distribution of word types over the affective space does not overlap with how often the types are selected for communication. The preference in word tokens is for highly positive words (regardless of their level of arousal), as well as for extremely negative extremely arousing words, see Figure 3: the distribution is markedly different from the boomerang-like curve observed in word types, see Figure 2.

The body of findings presented above gives rise to a number of methodological and theoretical claims regarding two novel distinctions that we made at the outset of the paper: valence vs arousal, and types vs tokens. We discuss these in turn.

Valence vs arousal

The methodological need for a joint consideration of the two key dimensions of affect (Boucher & Osgood, 1969) stems from the fact that one dimension plays a critical modulating role in the strength and direction of distributional biases shown by the other dimension. As Table 3 demonstrates, the advantage in the percent of positive words is attenuated and even significantly reversed as words become more arousing. A similar reversal of sign is observed in correlations of arousal with token frequency calculated per decile of valence.

Thus, prior studies of the positivity bias essentially collapse a bi-dimensional distributional pattern onto a single dimension of valence. This is arguably more harmful for the boomerang-shaped functional curve observed in word types. When collapsed onto the valence dimension, many properties of the shape are not observable, including the presence of counter-directed slopes, the slopes' relative magnitudes (reflecting potential asymmetry between the polarity of affect and its strength), or the inflection point of the functional curve (as an index of true emotional neutrality), see Kuppens et al. (2013). A sole observable property of the boomerang shape is its very indirect characteristic, namely, the number of words formed by a dissection of the boomerang at the midpoint of the valence scale, aggregated over levels of arousal. Similarly, focusing on positivity in the distribution of word tokens over the affective space misses a theoretically significant increase in token frequency associated with danger (negative arousing) words. To sum up, an accurate depiction of the spectrum of affect, as represented by lexical statistics of the English language, requires a bi-dimensional (or perhaps a multi-dimensional) perspective on both the words and the phenomena they denote.

Types vs Tokens

Separation of word types and tokens in our analyses as linguistically and psychologically independent indices of language use has received strong empirical support in the present data. Indeed, distributional patterns characterizing the two types of linguistic units are highly dissimilar, from a boomerang shape in types to a concentration of tokens in the opposite bands and corners of the affective space. The type-token distinction also sheds light on, and enables a revision, of social factors proposed in the literature as causes of affective biases. These factors revolve around two statements about the emotional structure of the world and society: life contains more positive than negative events, concepts and objects (Augustine et al., 2010; Gable et al., 2000; Rozin et al., 2010) and humans consciously prefer to talk about the bright side of life to please the interlocutor and maintain more positive social interactions

(Augustine et al., 2010). The present data considerably qualify these claims of a positive outlook and pro-social benevolence in communication.

As we argue above, the word-stock of English is organized around subjective experience of affect. This implies that the bi-dimensional bias towards extremely valenced and arousing words and neutral calm words is not necessarily due to the prevalence of dangerous, exhilarating, or even mundane phenomena in daily life. It is there because humans tend to preferentially assign these affective values to the spectrum of phenomena they encounter in their life: the bias is in the structure of subjective perception rather than in the structure of the world.

Furthermore, the tendency of English speakers to preferentially draw on all and any positive words and dangerous words in their communication from the word-stock which is not even especially diversified in all these regions of affect is symptomatic. We consider this bias in light of the long-standing proposal (Boucher & Osgood, 1969) that communicative behavior is pro-social or benevolent in nature, where pro-sociality is broadly defined as a conscious choice of behavior aiming at benefitting the recipient of the message (an individual interlocutor or a group). Importantly, however, pro-sociality is often equated with a tendency to preferentially look at the bright side of life (cf. Boucher & Osgood, 1969; Augustine et al., 2011). The observed data patterns – and especially a spike in token frequency for low-valence high-arousal words – corroborate the notion of pro-sociality, but not of the bias towards all matters positive. It stands to reason that language users benefit from communicating intensely about sources of danger, and not only about sources of pleasure. Our data suggest then that conscious communicative behavior, gauged by the choice of words and topics, is organized in a way that benefits recipients of written and spoken messages by provindg a broader coverage of both phenomena that have a potential of a reward and that of a threat.

One explanation for the token-biased bias towards these phenomena may come from the appetitive motivation to seek rewards and the fearful motivation to avoid threats (Bradley, 2000; Bradley & Lang, 2000). As argued in Lang, Bradley and Cuthbert (1990), positive stimuli associated with usefulness for survival (including indices of sustenance, nurturance, and caregiving) or negative-arousing stimuli associated with danger have a privileged status in regulating hormonal control, engaging attention, and shaping cognitive processing (cf. Bradley, Codispoti, Cuthbert, & Lang, 2001; Wurm, 2007). Öhman and Mineka's (2001) studies of fear in humans and primates further show that the threat-detection system is physiologically grounded, automatically activated and is relatively impervious to cognitive control. We argue then that the statistical patterns observed in language use, i.e. word-token bias, faithfully demarcate the most salient regions of the affective space, the ones that show the most immediate and critical impact of emotion on physiological and cognitive processes. We add that the observed patterns do not allow for distinguishing between approach towards dangerous objects (with the purpose of attacking them) or avoidance of dangerous objects (with a purpose of escaping them) as a preferred behavioral strategy.

Conclusions

To summarize, a substantially large collection of emotional ratings has enabled us to identify and confirm distributional biases towards the usage of positive/negative and calm/arousing words in the English language. Statistical regularities of language use mirror both the emotional structure of the world and society and even more so the subjective emotional structure of a human being, with his or her primary motivations and behavior in cognitive tasks influenced by affect. We argue for two factors that shape the structure of affect in society – as revealed via language. One is a pro-social benevolent communication strategy: we talk more about phenomena that can benefit our interlocutors by contributing to their survival and well-being, i.e. highly pleasurable and highly dangerous things. Another is prevalence and a broader diversification of lexical items expressing affective states that are most common in human subjective experience, with more extreme valence associated with higher arousal. In this sense, cumulative linguistic behavior of vast collectives of language users replicates subtle experiential preferences observed in small groups of individuals. Finally, we argue that both factors are rooted in the fundamental motivational systems underlying emotional responses. As such, an accurate characterization of affective biases provides a fuller understanding of how language, society, emotion and thought interrelate.

Acknowledgments

This work was supported by a SSHRC Joseph-Armand Bombardier CGS Scholarship to the first author and the SSHRC Insight Development grant 430-2012-0488, the NSERC Discovery grant 402395-2012, and the NIH R01 HD 073288 (PI Julie A. Van Dyke) to the second author.

Appendix

Re-analysis of Kloumann et al's (2012) Data

Our observation of consistent correlations between word token frequency and word positivity across genres and language varieties (US vs UK), runs counter to a recent study by Kloumann et al. (2012) in which valence was found to only correlate very weakly with frequency in a few genres (Twitter and music lyrics) and not at all in others (Google Books and The New York Times). They concluded that valence and frequency were independent, a finding at odds not only with our results but also the reports of previously published work (see Introduction).

A close reading of Kloumann et al. (2012) methods reveals a few potential areas for concern. First, they did not employ any inclusion criteria with regards to the character strings they included in their sample, beyond the fact that the strings were among the top 5,000 from each of their corpora. This led to an inclusion of multiple spelling variants (bday, b-day, and birthday), words with special characters (#music, #tcot), foreign words not borrowed into English (cf. Dutch hij "he" and zijn "to be"), alphanumeric strings (a3 and #p2) and others. The meaningfulness of happiness ratings for items such as these may be limited as a reflection of the emotional representation of the item in a typical speaker of English. (For comparison, all words in the stimulus list of Warriner et al. were identified as known by at least 70% of participants in another mega-study, Kuperman et al., 2013). Another consequence of Kloumann et al.'s decision to consider all character strings occurring in corpora "as is" is that word forms of the same lemma (e.g., *walk*, *walks*, *walked*, and *walking* as word forms of lemma *walk*; or *table* and *tables* as word forms of lemma table) are represented as emotionally discernible, independent language events. While such representation is logically possible, it is bound to represent a psychologically unlikely situation in which word forms of a lemma are each grounded in their own emotional experience and associated with values of positivity or arousal that are independent of those in other word forms related to the same lemma. We also note that the inclusion of word forms, instead of lemmas, in a word list, leads to inflation in the number of word types, and a lower token count, relative to lemma, for each specific word form: both implications of treating word forms independently are predicted to affect the type- and token-bias estimates.

We compared our results with Kloumann et al's in two ways. First we calculated Spearman correlations between our emotional ratings and ranked frequency for the overlapping words in each of their four corpora. Their correlations and ours are both reported in Table 4. They found the strongest relationship between emotion and frequency rank with their Twitter corpus, a relationship which is nearly equivalent to what we found in the overlapping words. However, they found only very weak relationships in the remaining three corpora where we found correlations 2 to 16 times stronger than theirs. We also divided each set of overlapping words into 10 deciles of frequency and plotted the distribution of valence in each (see Figure 4). In the plots, the proportion of words falling above the midpoint of the scale increases as frequency decile increases. Inset are plots showing how this change does not occur when Kloumann et al.'s full word list are used for each corpus along with their average happiness ratings and frequency ranks.

Table 4. Comparison of data from Kloumann et al. (2012) and Warriner et al. (2013). Note that in Kloumann et al. (2012), frequency estimates were provided as a ranking with lower numbers representing the highest frequency. As such, the correlations they report are opposite in sign to the correlations reported in other parts of this paper where frequency is reported as a log-transformed frequency count. For comparability, in the Warriner et al. portion of the table, we report correlations between Warriner et al.'s ratings and a similarly frequency ranking based on the frequencies provided in SUBTLEX.

		Twitter	Google Books	New York Times	Music Lyrics
KLOUMANN ET AL.	Words	5,000	5,000	5,000	5,000
	% pos	72.0	78.8	78.4	64.1
	Median V	5.54	5.64	5.56	5.34
	Skewness V	-0.64	-0.89	-0.83	-0.44
	V ρ	-0.103	-0.013	-0.044	-0.081
WARRINER ET AL.	Words Overlap	2,443	2,704	2,354	2,458
	% pos	73.3	72.9	74.3	66.7
	Median V	5.73	5.59	5.68	5.57
	Skewness V	-0.71	-0.66	-0.76	-0.53
	V ρ	-0.101	-0.219	-0.187	-0.218
	Median A	4.14	3.95	4.04	4.16
	Skewness A	0.51	0.63	0.64	0.47
	A ρ	0.034	-0.054	-0.020	0.044

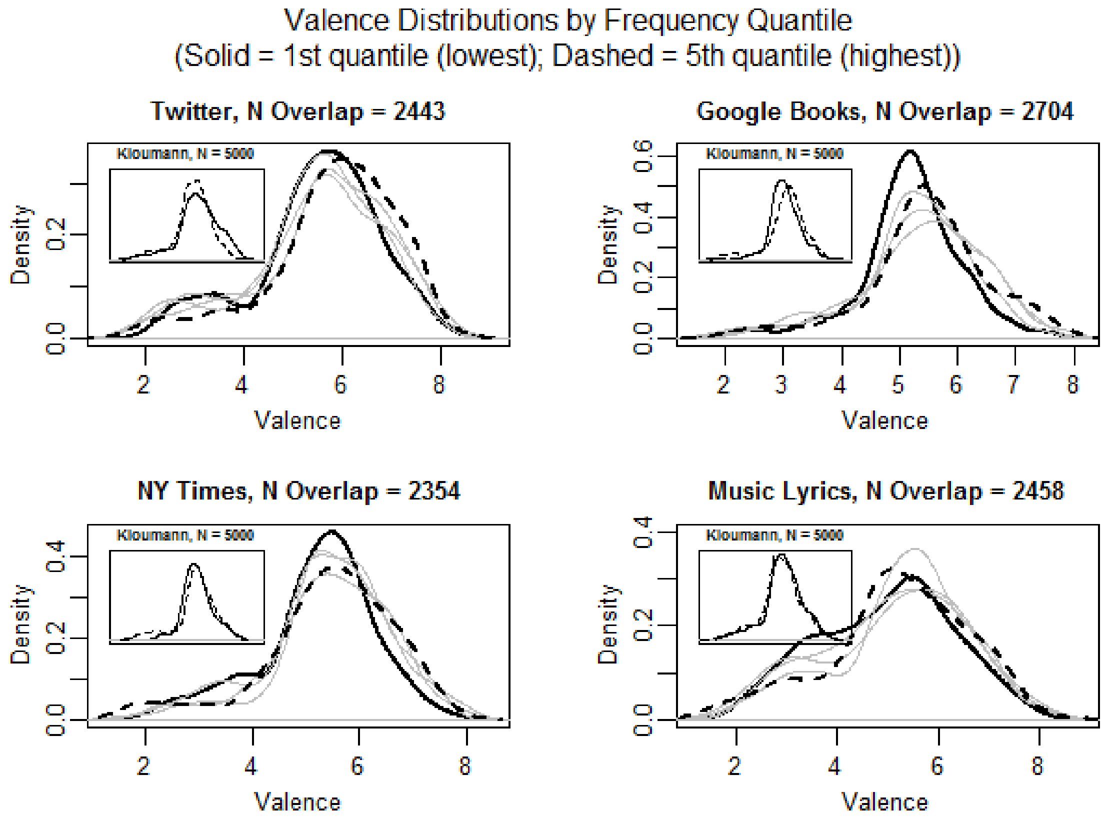


Figure 4. Density plots for each of the corpora in Kloumann et al. (2011). In the main area, the overlapping words between Kloumann et al. and Warriner et al. (2013) are plotted based on ratings from Warriner et al. These words were divided into quartiles based on log Frequency, each quartile then being plotted separately – the solid line represents the lowest quartile and the dashed line the highest quartile. The inset plots use the full dataset from Kloumann et al. along with their frequency ranks and happiness ratings.

References

Acerbi, A., Lampos, V., Garnett, P., & Bentley, R. A. (2013). The Expression of Emotions in 20th Century Books. *PloS one*, *8*(3), e59030.

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*, 508–515.

Bargh, J. A., Chaiken, S., Govender, R., & Pratto, F. (1992). The generality of the automatic evaluation effect. *Journal of Personality and Social Psychology, 62,* 893–912.

Barrett, L.F. & Russell, J.A. (1999). The structure of current affect: Controversies and emerging consensus. *Current Directions in Psychological Science, 8*(1), 10-14.

Baumeister, R.F., Bratslavsky, E., Finkenauer, C., & Vohs, K.D. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323-370.

Boucher, J. & Osgood, C.E. (1969). The Pollyanna hypothesis. *Journal of Verbal Learning and Verbal Behavior, 8*, 1-8.

Bradley, M. M., Codispoti, M., Cuthbert, B. N., & Lang, P. J. (2001). Emotion and motivation I: defensive and appetitive reactions in picture processing. *Emotion, 1(3)*, 276-298.

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings (Technical Report No. C-1).* Gainesville, FL: University of Florida, NIMH Center for Research in Psychophysiology.

Bradley, M. M. (2000). Emotion and motivation. In J. T. Cacioppo, L. G. Tassinary, & G. Berntson (Eds.), *Handbook of Psychophysiology* (pp. 602–642). New York: Cambridge University Press

Bradley, M. M., & Lang, P. J. (2000). Measuring emotion: Behavior, feeling, and physiology. In R. D. Lane & L. Nadel (Eds.), *Cognitive neuroscience of emotion*. New York: Oxford University Press.

Bradley, M. M., & Lang, P. J. (2007). The international affective picture system (IAPS) in the study of emotion and attention. In J. A. Coan & J. J. B. Allen (Eds.), *Handbook of emotion elicitation and assessment* (pp. 29–46). New York, NY: Oxford University Press.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera 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), 977-990.

Bucci, W. & Freedman, N. (1981). The language of depression. *Bull. Menninger Clin, 45*, 334–58.

Bybee, J. L. (2010). *Language, usage and cognition* (Vol. 98). Cambridge: Cambridge University Press.

Clark, H. H., & Clark, E. (1977). *Psychology and language: An introduction to psycholinguistics.* New York: Harcourt Brace Jovanovich.

Davies, M. (2009). The 385+ Million Word Corpus of Contemporary American English (1990-2008+): Design, Architecture, and Linguistic Insights. *International Journal of Corpus Linguistics, 14*, 159-90.

Davies, M. (2012). Expanding Horizons in Historical Linguistics with the 400 million word Corpus of Historical American English. *Corpora, 7*, 121-57.

Doerksen, S. & Shimamura, A.P. (2001). Source memory enhancement for emotional words. *Emotion, 1*(1), 5-11.

Dresler, T., Mériau, K., Heekeren, H. R., & Van der Meer, E. (2009). Emotional Stroop task: effect of word arousal and subject anxiety on emotional interference. *Psychological Research PRPF*, *73*(3), 364-371.

Duffy, E. (1951). The concept of energy mobilization. *Psychological Review, 58(1)*, 30-40.

Fontaine, J.R.J., Sherer, K.R., Roesch, E.B., & Ellsworth, P.C. (2007). The world of emotions is not two-dimensional. *Psychological Science. 18*(12), 1050-1057.

Gable, S., Reis, H. T., & Elliot, A. (2000). Behavioral activation and inhibition in everyday life. *Journal of Personality and Social Psychology, 78*, 1135–1149.

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

Higgins, E. T. (2006). Value from hedonic experience *and* engagement. *Psychological review, 113(3)*, 439-460.

Johnson, R.C., Thomson, C.W., & Frincke, G. (1960). Word values, word frequency, and visual duration thresholds. *Psychological Review, 67*(5), 332-342.

Kahneman, D. (1973). *Attention and effort*. Prentice-Hall Inc., Englewood Cliffs, New Jersey.

Kensinger, E.A. & Corkin, S. (2003). Memory enhancement for emotional words: Are emotional words more vividly remembered than neutral words? *Memory & Cognition, 31*(8), 1169-1180.

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

Kucera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A.B. (in press). Emotion and language: Valence and arousal affect word recognition. *Journal of Experimental Psychology: General.*

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*, 978–990. doi:10.3758/s13428-012-0210-4

Kuppens, P., Tuerlinckx, F., Russell, J.A., and Barrett, L.F. (2013). The relation between valence and arousal in subjective experience. *Psychological Bulletin, 139*, 917-940.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1990). Emotion, attention, and the startle reflex. *Psychological review*, *97*(3), 377.

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

Lakoff, R.T. 1975. *Language and Woman's Place*. New York: Harper & Row.

Liu, B. (2012). *Sentiment analysis and opinion mining*. San Rafael, CA: Morgan & Claypool.

Michel, J., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., The Google Books Team, Pickett, J.P., Hoiberg, D., Clancy, D., Norbig, P., Orwant, J., Pinker. S., Nowak, M.A., & Aiden, E.L. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*, 176-182.

Mulac, A., Bradac, J.J., & Gibbons, P. (2001). Empirical support for the gender-as-culture hypothesis: An intercultural analysis of male/ female language differences. *Human Communication Research, 27*, 121–52.

Newman, M.L., Pennebaker, J.W., Berry, D.S., & Richards, J.M. (2002). Lying words: Predicting deception from linguistics styles. Personality and Social Psychology Bulletin, 29(5), 665-675. doi: 10.1177/0146167203251529

Pennebaker, J.W. & King, L.A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*(6), 1296-1312.

Pennebaker, J.W. & Stone, L.D. (2003). Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology, 85*(2), 291-301. doi: 10.1037/0022-3514.85.2.291

Power, M.J. (2006). The structure of emotion: An empirical comparison of six models. *Cognition & Emotion, 20*(5), 694-713. doi: 10.1080/02699930500367925

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

Rozin, P. & Royzman, E.B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*(4), 296-320.

Rozin, P., Berman, L., & Royzman, E. (2010). Biases in use of positive and negative words across twenty natural languages. *Cognition and Emotion, 24*(3), 536-548.

Rude, S.S., Gortner, E. & Pennebaker, J.W. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion, 18*(8), 1121-1133.

Russell, J.A. (1991) Culture and the categorization of emotions. *Psychological Bulletin, 110 (3)*, 426-450.

Schnoebelen, T., & Kuperman, V. (2010). Using Amazon mechanical turk for linguistic research. *Psihologija*, *43*(4), 441-464.

Schrauf, R. W., & Sanchez, J. (2004). The preponderance of negative emotion words in the emotion lexicon: A cross-generational and cross-linguistic study. *Journal of Multilingual and Multicultural Development, 25(2-3)*, 266-284.

Semin, G. R., & Fiedler, K. E. (1992). *Language, interaction and social cognition*. Sage Publications, Inc.

Sharot, T. (2011). The optimism bias. *Current Biology*, *21*(23), R941-R945.

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

Steyvers, M., & Tenenbaum, J. B. (2005). The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive science*, *29*(1), 41-78.

Suitner, C., & Maass, A. (2008). The role of valence in the perception of agency and communion. *European Journal of Social Psychology*, *38*(7), 1073-1082.

Suitner, C. & Maass, A. (2008). The role of valence in the perception of agency and communion. *European Journal of Social Psychology, 38*, 1073-1082. doi: 10.1002/ejsp.525

The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

Unkelbach, C., Fiedler, K., Bayer, M., Stegmuller, M., & Danner, D. (2008). Why positive information is processed faster: The density hypothesis. *Journal of Personality and Social Psychology, 95*(1), 36-49. doi: 10.1037/0022-3514.95.1.36

Unkelbach, C., von Hippel, W., Forgas, J.P., Robinson, M.D., Shakarchi, R.J., & Hawkins, C. (2010). Good things come easy: Subjective exposure frequency and the faster processing of positive information. *Social Cognition, 28*(4), 538-555.

Unkelbach, C. (2012). Positivity advantages in social information processing. *Social and Personality Psychology Compass, 6*(1), 83-94. doi: 10.1111/j.1751-9004.2011.00407.x

Van Heuven, W.J.B., Mandera, P., Keuleers, E., & Brysbaert, M. (in press). Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*.

Vrij, A. 2000. *Detecting Lies and Deceit: The Psychology of Lying and the Implications for Professional Practice.* Chichester, UK: Wiley.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 1-17. doi: 10.3758/s13428-012-0314-x

Weintraub, W. (1981). *Verbal behavior: Adaptation and psychopathology*. New York: Springer.

Weintraub, W. (1989). *Verbal behavior in everyday life*. New York: Springer.

Wurm, L.H. (2007). Danger and usefulness: An alternative framework for understanding rapid

evaluation effects in perception? *Psychonomic Bulletin & Review, 14(6)*, 1218-1225.

Zajonc, R.B. (1968). Attitudinal effects of mere exposure. Journal of Personality and Social Psychology Monograph Supplement, 9(2), 1-27.

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). The educator's word frequency guide/Touchstone Applied Science Associates. Inc, Brewster, NY.

CHAPTER 7: General Discussion

Thesis Summary

The purpose of this dissertation is to address limitations in previous studies looking at the interplay of language and emotion, and set the stage for future explorations. A paucity of emotional word norms has hampered the selection of stimuli both in terms of numbers and degree of control over contributing factors, leading to issues with generalizabiilty of results. In particular, the impact of arousal and its interaction with valence was often left untested. In addition, perhaps because small numbers of stimuli restricted statistical power of analyses or simply because of the newness of this field, little consideration has been given to group and individual differences.

Therefore, our starting point was to collect a large set of emotional word norms. Chapter 2 presented results from a study that collected nearly 14,000 word ratings across three emotional dimensions – valence, arousal, and dominance (Warriner et al., 2013). We were able to confirm the reliability of our ratings via high correlations with other smaller sets, including several translated from other languages. Valence was more consistent across participants and across studies than arousal and dominance. That combined with the fact that valence and dominance were strongly correlated led us to focus primarily on valence and arousal in subsequent analyses and studies. With 1,827 participants contributing to the dataset and an average of 18 ratings per word, we were able to calculate means for population sub-groups including gender, education level, and age. As a result, we were able to make observations such as the fact that among the less educated, younger participants provided higher ratings than older while there was no interaction with age for the more educated. Such distinctions have never been made in previous sets of norms. We also found that gender differences in ratings within certain semantic domains patterned themselves after previously studied attitude differences in areas such as attitudes towards diseases, occupations, and sexual activity. This provided a confirmation of the validity of our dataset and an intriguing potential for easy to obtain emotional ratings serving as stand-ins for more difficult to acquire attitude assessments.

This set of norms would be even stronger with a few more participants to provide a stronger balance within subgroups. Continuing to add more words would, of course, always be beneficial. However, we did base stimuli collection on lemmas from the SUBTLEX-US corpus (Brysbaert & New, 2009) with the highest frequency. By attributing ratings of lemmas to inflected word forms, our dataset can account for 28,724 words, a substantial percentage of an average English speaker's vocabulary. It is, at the very least, 13 times bigger than the largest previous set. In addition, choosing words based purely on frequency allowed us to select an unbiased sample of the English language (although the English language itself may have inherent biases – see Chapter 6) in contrast to previous norms such as ANEW (Bradley & Lang, 1999) which were selected for their emotionality and thus not representative of the natural distribution of affect among words. One distinct advantage of this collection over researchers collecting ratings within their own study designs is that the samples of words seen by each participant were completely mixed, drawn at random from the whole list. In an unpublished pilot study, we found that when word lists were focused on a semantic domain or strongly weighted on an emotional dimension, participants' ratings varied significantly from the patterns seen in the larger study. This suggests that word list effects have the potential to skew results and that using a set of norms, such as ours, might be a way to avoid this issue.

In the vein of large, crowd-sourced rating studies, we also collected concreteness ratings for over 60,000 English words, nearly 40, 000 of which were known by at least 85% of raters and reported in Chapter 3 (Brysbaert et al., 2013). Concreteness is a measure of how much a word can be perceived with the senses and is one of the semantic variables engendering significant recent interest in relation to word processing (e.g. Connell & Lynott, 2012; Kousta, Vigliocco, Vinson, Andrews & Campo, 2011). Previous sets of norms, while larger than for other variables, remained small at approximately 9,000 (Coltheart, 1981) and focused too strongly on visual perception (Connell & Lynott, 2012). Again, with such a large participant pool, we were able to be unbiased in our sample and calculate mean ratings for subgroups by age, education, and gender. By measuring how well words were known, we were able to identify words which future studies of additional variables should focus on. We found that high frequency (which has been used as a selection criterion in previous collections) and percent known did not always correspond. For example, people are likely to know what the word 'oxygen' means but less likely to use it in everyday conversations.

While none of the other studies in this dissertation directly deal with concreteness, it is an important variable for future studies of affect. In the past, concrete words have been shown to have an advantage over abstract words in lexical decision times (Binder, Westbury, McKiernan, Possing, & Medler, 2005; Bleasdale, 1987; Kroll & Merves, 1986; Schwanenflugel & Stowe, 1989). The common explanation has been that concrete words have a richer sensorimotor representation (Paivio, 2007; Schwanenflugel, 1991). Kousta et al. (2011), however, showed that when imageability and context availability were controlled, abstract words had an advantage over concrete words. They went on to show that abstract words have a richer affective representation, thus concrete or abstract words will have advantages depending on what other aspects of word meaning are controlled or available. In support of this idea, Newcombe, Campbell, Siakaluk, and Pexman (2012) showed that higher imageability and body-object interaction ratings correlated with faster and more accurate categorization of concrete nouns but slower and less accurate categorization for abstract nouns. In contrast, higher emotional ratings correlated with faster and more accurate categorization for abstract nouns but slower and less accurate categorization for concrete nouns. The processing of abstract nouns also varies according to how strongly they evoke emotional experience (Siakaluk, Knol, & Pexman, 2014). As such, there is growing evidence that concreteness and emotion interact and both play important roles in word processing. The development of these two large sets of norms should prove to be a critical resource for advancing this area of study.

As an increasing number of large collections of norms are established, researchers are merging their data to reveal new and important insights into the variables that affect word processing. For example, the English Lexicon Project (Balota et al., 2007) contains lexical decision and naming averages to over 40,000 words. Combining this dataset with our newly established set of emotional norms (Warriner et al., 2013), we were able to investigate how emotion influences these basic measures of word recognition (see Chapter 4, Kuperman, Estes, Brysbaert, & Warriner, 2014). Past studies had attempted to determine this relationship but limited norms and differences in controls led to conflicting conclusions (Estes & Adelman, 2008, Kousta et al., 2009, Vinson, Ponari, & Vigliocco, 2014). We showed that valence and arousal both have monotonic and independent effects on responses with increasing valence leading to quicker reaction times and increased arousal leading to slower reaction times. Both of these effects were strongest among low-frequency words. We were able to demonstrate that previously reported and conflicting patterns were due to sampling issues and failure to consider interactions with frequency. Another novel observation we made is that these effects were consistent even at the individual level. We were also able to show that the effects of valence and arousal on word recognition measures persisted even after other semantic variables were accounted for.

Our results argue that negative or threatening words are not categorically prioritized over neutral ones, as suggested by the automatic vigilance hypothesis (Estes & Adelman, 2008). Instead, we show that fine discriminations between levels of affect are made quickly enough to influence word recognition measures. Our results also contrast with the view that both positive and negative words are prioritized over neutral words (Kousta et al., 2009) and the view that valence and arousal interact such that negative paired with high arousal and positive paired with low arousal are prioritized over other combinations (Larsen et al., 2008).

Each of these rejected hypotheses are attempts to explain the motivations behind reactions to emotional stimuli. In the categorical view, distinguishing aversive (negative) and appetitive (positive) stimuli from each other is critical for survival. Negative stimuli, regardless of how negative, capture attention, preparing a person to engage in avoidance behavior. When valence is not relevant to the task, this recruitment of resources is not helpful and must be suppressed, thereby leading to categorically slower responses. Those who found an inverse-U pattern suggested that both aversive and appetitive stimuli are noticed quicker because both avoiding danger and acquiring resources is important for survival. In this case, however, one must assume that activating these motivational systems speeds up responses as the researchers found that both positive and negative words led to faster responses. The interaction observed between valence and arousal was also explained by motivation in that negative, high arousal words preferentially map onto danger while positive, low arousal words map onto rewards.

There has been a significant amount of previous research on approach and avoidance motivation and how they undergird emotion, specifically valence. In Chapter 5, we reviewed how this link has been primarily demonstrated via a congruency effect – people are faster to make approach-like responses to positive stimuli than to negative and vice versa for avoidance-like responses (Chen & Bargh, 1999; De Houwer et al., 2001; Stins et al., 2011). Measuring of a congruency effect requires the stimuli to be dichotomized, however, we have shown that both in ratings and in word recognition response times, valence and arousal are continuous. People do not categorically distinguish between positive and negative, not even simply as a first stage. Instead, they make fine distinctions very early on in processing, as demonstrated by the interaction between proportional responses and frequency which is known to be processed quickly (see Chapter 4). This raises the question of whether approach and avoidance behaviors are likewise graded. Is the activation of these motivational systems all or nothing, or do stronger stimuli engender a stronger response?

As a first step towards investigating these questions, we implemented a new method by which people indicated how close or how far they wanted to be from a given word via the movement of an on-screen manikin (Chapter 5). Thus, instead of simply moving towards or away, we measured actual distance from the stimuli. We found that people consistently indicated that they would move closer to a stimulus as valence increased and further as valence decreased. Arousal did not play a role either independently nor in interaction with valence. We found, again, an effect of frequency in that higher frequency words were approached more readily than less frequent words, even after controlling for valence.

In examining individual differences, we found that people who stay further away on average were more sensitive to both valence and frequency. People generally moved closer to words that had been rated more pleasant by those of their same gender. There was a similar, albeit much weaker tendency for people to also move closer to words which were rated more arousing by their same gender. Highly sociable people tended to move closer to all words while, on average, shy people moved further away.

With this method, we've established that approach and avoidance motivations are activated in degrees. That degree is linked to valence, frequency, gender, and individual differences such as sociability and shyness. There was little evidence that arousal had an effect. It is possible that while valence determines direction and distance, arousal determines speed or force. We did test whether there were effects on first click reaction times or overall time from stimuli onset to trial completion and did not find any significant patterns, however, our instructions were lax enough that participants may not have responded with the urgency necessary to see such effects. We also know from Chapter 4, that arousal contributes much less to word recognition times than does valence. As such, the effects of arousal on approach and avoidance distances may be too small to detect with this method.

The preponderance of research on emotion and cognition has focused on how the emotion inherent in a stimulus affects behavior, attempting to draw inferences about the resulting patterns. For example, we observe that reactions are slower to negative than to positive words and we conclude that we must have an evolutionary drive to avoid danger. We find corroboration for our inferences in other behavioral patterns such as the tendency to move towards positive and away from negative stimuli. But there is an additional source of information about the motivational underpinnings of emotion that has been less examined. In Chapter 6, we explored what the distribution of emotion across the breadth of the English language can tell us about human experience. While a positivity bias, or a tendency for positive words to be used more frequently, was first observed in the 1960's (Boucher & Osgood, 1969; Zajonc, 1969), this area of research has similarly been hampered by small sets of norms. Using the set from Warriner et al. (2013) as a proxy for language in general, we found that bias within English was best viewed as bi-dimensional. The majority of word types are positive or negative and of high-arousal, or mid-valence and of low-arousal. Based on the idea that words are tools, created to identify and distinguish between important elements of human experience, one must conclude that these combinations are characteristic of what we find ourselves needing to communicate about. We need to have a way to talk about dangers and thrills as well as the mundane stuff of everyday life. We are content to have less words that describe low valence, low arousal (depression, shame), high-valence low-arousal (serenity, comfort), and mid-valence high arousal events. Interestingly, this functional relationship is the same as that found in subjective reports of affect (Kuppens et al., 2013) meaning that we have more words that reflect the very states that people find themselves in more often. In terms of frequency, extremely positive words regardless of arousal and extremely negative, high arousing words are used more often than others. This may reflect communicative behavior that is geared towards fostering positive relationships and warning of threats. As such, there is evidence within the distribution of the English language itself that affect is organized around the poles of reward/appetitive/approach and danger/aversive/avoidance.

Limitations

As already mentioned, rating studies could always benefit from additional participants and stimuli. While our concreteness dataset is quite comprehensive and our affective dataset is significantly larger than any previous collection, there remain words that were not included. The number of words is primarily a matter of utility in studies that select subsets to use as stimuli, but could hamper the generalization from dataset to the English language as a whole such as in our characterization of the distribution of affect in Chapter 6. The usage of these norms in other studies makes assumptions that may or may not be justified such as the extension of an affective rating for a given lemma to its various inflected forms (e.g. from *walk* to *walked* and *walking*) and a lack of distinction between alternate meanings of homophones (from a river *bank* to a financial *bank*). At the moment, the cost of thoroughly investigating these assumptions is thought to outweigh the cost of making them.

The participants who contributed the ratings were residents of the US whose native language was English. There was significant variety in gender, age, and education level which should increase the generalizability of our ratings to the general population of English speakers. However, there are cultural differences in word use and meaning among Anglophones, for example, between speaker of American vs. British English which may need to be taken into account. In addition, the majority of research that uses these ratings is performed with undergraduate students for which our dataset may be less accurate. In Chapter 2, we did find differences in ratings between both age and education level groups. However, in Chapter 5, when we compared ratings collected at time of test with the set from Warriner et al. (2013), the results were the same, showing that for at least that study, the differences were not important.

Another factor that has been shown to affect studies of this kind is mood state. To pick a few examples - having a valence vs. and arousal focus can actually affect whether a discrete or dimensional model of emotions fit self-reports best (Barrett, 1998), mood can determine what word associations people make (Gilet & Jallais, 2011) and mood affects the ability to make semantic coherence judgments (Bolte, Goschke, & Kuhl, 2003). In Chapters 2 and 5, we did ask two mood related questions about participants' happiness and arousal level at the time of test, however, we did not find significant differences related to these answers and thus did not report them. With regards to the collection of ratings, it would be the hope that having a sufficient number of raters would balance out any outliers caused by mood. With regards to studies such as the slider in Chapter 5, it will become important to screen for both mood and mental health issues, especially if this is to become a tool for assessing disorders.

In the slider study, in particular, sample size could have affected our results. We were attempting to control for and test a large number of factors, particularly in Experiment 3 in which we introduced personality measures. This may have reduced our power to be able to detect significant relationships between variables. We reported the effects that were stable and significant across studies, however, there were effects that would show up in some but not all experiments depending on what variables were or were not included in the models. Including more participants would enable us to determine whether these were of actual interest or statistical anomalies. In addition, due to the female skewed nature of our subject pool, we restricted Experiment 3 to females only. It is very possible that gender and personality interact in how they affect approach and avoidance choices and will thus be important to investigate in the future.

Implications and Significance

The contribution of large sets of norms – both emotion and concreteness – to the field is a significant contribution in its own right. As of the time of writing, twenty-three published papers have cited the affective ratings and the dataset has been downloaded over 1000 times. With these norms, researchers have the ability to select larger and more appropriate sets of stimuli for their studies from a set without bias which covers the entire spectrum of both valence and arousal. With the number of norms now available, researchers' ability to control for important factors such as frequency or other semantic variables will be improved. Combining these norms with other datasets will make it possible to test assumptions about relationships between factors believed to affect word recognition. Fields, such as sentiment analysis, that mine texts for meaning, will be able to rely less on computational extensions of small sets of norms increasing the reliability and validity of their inquiries.

However, there are several additional implications of the research in this dissertation beyond just the collection of norms. For example, Chapters 2, 4 and 5 in this dissertation established that valence, as associated with words, can be measured via a scale and impacts word recognition in a continuous, not a categorical, manner. The majority of studies up until now have dichotomized stimuli into positive or negative and sometimes neutral, typically using ANOVA to analyze their results. Doing so fails to capture the full spectrum of relationships. If people are sensitive to small differences in both valence and arousal when engaging in basic word recognition, they will most likely be sensitive to those differences in other tasks. While the relationship between each emotional dimension and reaction time was monotonic in this instance, it does not mean it will always be monotonic. One cannot test dichotomized extremes and simply infer the middle.

The method introduced in Chapter 5 for measuring approach and avoidance via distance estimates validated the connection between valence and motivational systems. It demonstrates, for the first time, that approach and avoidance can be activated to a fuller or lesser extent depending on the nature of the stimulus. This requires an adjustment to theories about why we experience emotion. Rather than a discrete impulse to either seek resources or flee from danger, our experiences may range from minor to extreme such as slight aversion or an overwhelming desire. It is even possible that approach and avoidance are not two ends of a single dimension but separately activated such that there could be conflict. A typical example is wanting ice cream but knowing it’s bad for us. There is some evidence that when valence and arousal are mismatched, such as being high in both valence and arousal, responses are slower perhaps indicating conflict (Purkis, Lipp, Edwards, & Barnes, 2009; Robinson, Storbeck, Meier, & Kirkeby, 2004). Being able to measure degrees of approach and avoidance in this way opens the door for a host of experiments to test these possibilities and more (see Future Directions).

Importantly, the papers in this dissertation are among the first to take a serious look at individual and group differences in terms of how emotion affects word processing. There is evidence that individuals differ in how much they focus on positives versus negatives, or how sensitive they are to reward or punishment (Gray & Tallman, 1987; Higgins, 1997). Additionally, while valence and arousal in general relate to each other in an inverse-U fashion, individual subjective experience of these two dimensions can pattern differently (Kuppens et al., 2013). In fact, it has been suggested that such differences are a matter of personality (Elliot & Thrash, 2010). There has been some investigation of these differences in research on approach and avoidance motivation. For example, Puca, Rinkenauer, & Breidenstein (2006) showed that people who scored high in avoidance motivation executed more forceful avoidance movements in response to aversive stimuli but found no differences in reaction time. However, the many studies involved in collecting norms and using them in reaction time based studies have largely ignored the possibility of such fundamental differences. If ratings are going to form such a critical component of research on emotion, in word processing or otherwise, it behooves us to consider whether those ratings significantly differ based on individual characteristics or group membership. In Chapter 2, we observed differences in patterns of emotional ratings among groups based on gender, age, and education level. While not analyzed specifically, this same information was collected with our concreteness ratings (Chapter 3) and is available for researchers to use. In Chapter 4, we confirmed that the effect of valence and arousal on word recognition times is consistent at both the group and individual level, something that has not been considered before. In

Chapter 5, we showed that valence impacts approach and avoidance in the same direction regardless of gender, but that its effect varies in magnitude by gender, by individual, and specifically by how shy or sociable a person scored. Our findings underscore the importance of considering such variables in future studies.

Future Directions

As mentioned briefly earlier, continued collection of emotional norms would be beneficial, both in terms of extra ratings for the words already in our dataset and in terms of extending our dataset to cover all words that were rated as known by at least 85% of people. Most research has proceeded under the assumption that valence is bipolar, but this is still debated (Barrett & Russell, 1998; Cacioppo, Gardner & Berntson, 1997; Rafaeli & Revelle, 2006). It would be valuable to do a study comparing whether two sets of unipolar ratings predict lexical decision and naming times better than our current set of bipolar ratings. We showed that valence accounted for 2% of the variance in lexical decision times which leaves significant variance yet unexplained. Perhaps breaking valence apart would provide stronger explanatory power.

Of course, resources need to continue to be applied to the collection of additional norms. While we have provided a large set of emotional norms and a nearly comprehensive set of concreteness norms, there are other variables that would be of interest in the continued quest to understand word processing. Concreteness is a compound variable, a measure which attempts to capture the degree to which all senses are engaged by a given concept. Imageability is a related but someone distinct concept that relates to how easily an image of a referent can be brought to mind (Altarriba, Bauer & Benvenuto, 1999). However, small sets of norms for each sensory domain have also shown promise (Amsel, Urbach & Kutas, 2012;

Connell & Lynott, 2012). Wurm & Vakoch's (2000) danger and usefulness ratings offer an intriguing parallel to the motivational biases that underlay affective language. Another suggested variable involves the degree to which a person can interact with an object, referred to as body-object interaction ratings (Tillotson, Siakaluk, & Pexman, 2008). Further study on how these variables relate to valence and arousal, to approach and avoidance would offer important insight into the role of semantic, sensory, and connotational meaning in word processing.

The role of arousal in word processing needs to be examined more closely. Up until now, it has often been confounded with extremes of valence. Word types in English do typically show a pairing of high and low valence with high arousal, however, word usage does not. High valence words are used more frequently regardless of arousal (see Chapter 6). Future experiments need to take into account the many possible combinations of these dimensions. We have shown that arousal has its own independent effect on word recognition times (see Chapter 4). Although this effect was small, it was still significant. That said, it is still unclear what role arousal plays. As discussed, valence naturally pairs with approach and avoidance motivation. We showed that people determine how close or far they want to be from a word based on how positive or negative it is. However, we did not find a significant effect of arousal (see Chapter 5). This does not mean that arousal has no effect, only that its effect does not show up in distance ratings. Arousal has been shown to be linked more strongly than valence with more automatic responses of the autonomic system (Hofmann, et al., 2009; Kissler, et al., 2007). Perhaps arousal is less able to be captured in conscious, deliberate tasks such as ratings but would be captured with time-based measures such as how quickly a participant's arm muscles engaged in moving the mouse. In general, arousal has been far less studied than valence and its role in word processing deserves more attention.

In Chapter 4, we raised the question of whether the effects of emotion we saw on word recognition times were due to lexico-semantic factors or decision-response factors. We suggested that one lexicosemantic explanation might be the existence of more positive than negative word types which would engender greater affective priming between positive words thereby giving them a speed advantage. In Chapter 6, we confirmed a bias towards positive word types in English. While negative words have a disadvantage when being recognized due to a lack of priming, their rarity would explain their advantage in memory tests (Kensinger & Corkin, 2003a, 2003b). However, we also noted that a decision-response explanation made sense of the fact that emotion had a smaller effect on naming times than on lexical decision times because the former is less susceptible to decision making processes. It also made sense of why negative words are responded to more slowly when valence is irrelevant but more quickly when valence is central to the task at hand. If possible threat grabs attention, disengaging from that attention to respond in the former case would take time while in the latter case, the attention would facilitate responses (Estes & Verges, 2008). While it is likely that both lexicosemantic and decision response factors play a role in the effects observed, it will be important to find a way to identify their unique contributions. A better understanding of how each affect word processing will help make sense out of seemingly conflicting result patterns.

With regards to the slider method introduced in Chapter 5, there are many directions in which we could go. Possible future studies include placing words on either end of the scale to introduce conflict. For example, if a positive high arousal word and a positive low arousal word were placed opposite to each other, would we then see an effect on arousal in terms of which word participants chose to move towards at the cost of moving away from the other word? Measuring approach and avoidance separately would provide insight into whether there could be conflict between emotional dimensions. Can we characterize words that people want to both approach and avoid? Do people choose not to move at all in relation to these words when approach and avoidance are measures as opposites? Pictures and words have evoked different patterns of responding in cognitive paradigms (De Houwer & Hermans, 1994; Kensinger & Schacter, 2006). Would distance measures likewise differ in this paradigm? There has been some research into whether valence and arousal have effects only when the stimuli in question is personally relevant (Harmon-Jones, Lueck, Fearn, & Harmon-Jones, 2006; Sakaki, Niki & Mather, 2012; Tomaszczyk, Fernandes & MacLeod, 2008). By altering the nature of the manikin that is moved and our instructions concerning identifying with that manikin, we can test this possibility. For example, we could use an inanimate object as the moveable point or change the manikins age or gender to make it easier or more difficult to identify with. Developmentally, this slider could be used to determine whether emotion is linked to approach and avoidance in the same way in children or older adults. Is it an innate response that shows up early or is it developed later as children learn cultural values associated with words or with the scenes depicted in pictures? Does the connection fade with age as perhaps the evolutionary drive to obtain reproductive success is lessened? We showed that some people are more extreme in their responses than others and that shyness predicted distance. Could this method also be used to detect pathology such as extreme shyness in children?

Determining how emotion affects the processing of words in isolation is one aspect. However, language is typically transmitted and received in phrases and sentences. As such, research on how emotion impacts language processing must begin to include natural reading and eventually auditory paradigms. There have been a few studies that have used eye-tracking to measure the impact of emotion on fixation durations. Scott, et al. (2012) found that neutral words were always fixated longer than positive words. They were also fixated longer than negative words, but only when both were low in frequency. Negative words were fixated longer than positive words but only when both were high in frequency. Note that both positive and negative words in this study were high in arousal. Bayer, Sommer & Schacht (2010) found that negative high arousal words embedded in sentences evoked a stronger late positive complex than neutral words. Pupil size which can be measured along with fixation duration has been shown to increase in response to positive and negative stimuli compared to neutral (Bradley, Miccoli, Escrig, & Lang, 2008; Kuchinke, Vo, Hoffman, &Jacobs, 2007; Partala & Surakka, 2003). There is certainly far more to investigate about how emotion affects not only the word being fixated but also the words around it. A large set of emotional word norms will make it easier to form sentences with effective controls and expand this area of research.

Conclusion

As stated in the introduction, the purpose of this dissertation was to set the groundwork for inquiries into the interplay of language and emotion, specifically how emotion impacts word processing. By establishing several large sets of word norms and showing what can be discovered when they are combined with behavioral data, I have demonstrated new potential and raised several issues that need to be addressed in future studies. I proposed and tested a new method for connecting research on emotional word processing to underlying motivational systems. I also showed how the English language itself has been organized around these systems.

References

Adelman, J. S. (Ed.) (2012). *Visual word recognition, volume 1: Models and methods, orthography and phonology*. Hove, England: Psychology Press.

Adelman, J. S., Marquis, S. J., Sabatos-DeVito, M. G., & Estes, Z. (2013). The unexplained nature of reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1037-1053.

Adelman, J. S., Sabatos-DeVito, M. G., Marquis, S. J., & Estes, Z. (2014). Individual differences in reading aloud: A mega-study, item effects, and some models. *Cognitive Psychology, 68*, 113–160.

Algom, D., Chajut, E., & Lev, S. (2004). A rational look at the emotional Stroop phenomenon: A generic slowdown, not a Stroop effect. *Journal of Experimental Psychology: General*, *133*(3), 323-338.

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

Amaral, D. G. & Price, J. L. (1984). Amygdalo-cortical projections in the monkey (Macaca fascicularis). *Journal of Comparative Neurology*, *230*(4), 465-496.

Amaral, D. G., Behniea, H., & Kelly, J. L. (2003). Topographic organization of projections from the amygdala to the visual cortex in the macaque monkey. *Neuroscience*, *118*(4), 1099-1120.

Amsel, B. D., Urbach, T. P., & Kutas, M. (2012). Perceptual and motor attribute ratings for 559 object concepts. *Behavior Research Methods*, *44*(4), 1028-1041.

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, *116*(3), 463-498.

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review*, *4*(4), 439-461.

Aquino, J. M., & Arnell, K. M. (2007). Attention and the processing of emotional words: Dissociating effects of arousal. *Psychonomic Bulletin & Review*, *14*(3), 430-435.

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*. DOI: 10.1177/1948550611399154.

Baayen, R. H. and Schreuder, R. (eds), *Morphological structure in language processing*, Berlin: Mouton.

Baayen, R.H. (2010). A real experiment is a factorial experiment? *The Mental Lexicon 5*(1), 149-157.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*(2), 283-316.

Balota, D. A., Ferraro, F. R., & Connor, L. T. (1991). On the early influence of meaning in word recognition: A review of the literature. In P.J. Schwanenflugel (Ed.), *The psychology of word meanings* (pp. 187-222). Hillsdale, NY: Lawrence Erlbaum.

Balota, D. A., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition*, *29*(4), 639-647.

Balota, D. A., Yap, M. J., & Cortese, M. J. (2006). Visual word recognition: The journey from features to meaning (A travel update). In M. Traxler & M. A. Gernsbacher (Eds.), *Handbook of Emotions, 2nd Edition* (pp. 91-115). New York, NY: Guildford Press.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445-459.

Barber, H. A., Otten, L. J., Kousta, S. T., & Vigliocco, G. (2013). Concreteness in word processing: ERP and behavioral effects in a lexical decision task. *Brain and Language*, *125*(1), 47-53.

Barrett, L. F. (1998). Discrete emotions or dimensions? The role of valence focus and arousal focus. *Cognition and Emotion*, *12*, 579-599.

Barrett, L. F. (2004). Feelings or words? Understanding the content in self-report ratings of emotional experience. *Journal of Personality and Social Psychology*, *87*, 266-281.

Barrett, L. F. (2006a). Emotions as natural kinds? *Perspectives on Psychological Science*, *1*, 28-58.

Barrett, L. F. (2006b). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review*, *10*, 20-46.

Barrett, L. F., & Niedenthal, P. M. (2004). Valence focus and the perception of facial affect. *Emotion*, *4*(3), 266.

Barrett, L.F. & Russell, J. A. (1998). Independence and bipolarity in the structure of current affect. *Journal of Personality and Social Psychology*, *74*(4), 967-984.

Bayer, M., Sommer, W., & Schacht, A. (2010). Reading emotional words within sentences: The impact of arousal and valence on event-related potentials. *International Journal of Psychophysiology*, *78*(3), 299-307.

Binder, J., Westbury, C., McKiernan, K., Possing, E., & Medler, D. (2005). Distinct brain systems for processing concrete and abstract concepts. *Cognitive Neuroscience, Journal of*, *17*(6), 905-917.

Bleasdale, F. A. (1987). Concreteness-dependent associative priming: Separate lexical organization for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(4), 582-594.

Bocanegra, B. R. & Zeelenberg, R. (2009). Dissociating emotion-induced blindness and hypervision. *Emotion*, *9*(6), 865-873.

Bolte, A., Goschke, T., & Kuhl, J. (2003). Emotion and intuition effects of positive and negative mood on implicit judgments of semantic coherence. *Psychological Science*, *14*(5), 416-421.

Boucher, J. & Osgood, C. E. (1969). The pollyanna hypothesis. *Journal of Verbal Learning and Verbal Behavior*, *8*(1), 1-8.

Boyle, G. J. (1986). Higher-order factors in the Differential Emotions Scale (DES-III). *Personality and Individual Differences*, *7*(3), 305-310.

Bradley, M. M. & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings* (pp. 1-45). Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

Bradley, M. M. & Lang, P. J. (2000). Affective reactions to acoustic stimuli. *Psychophysiology*, *37*(2), 204-215.

Bradley, M. M., Codispoti, M., Cuthbert, B. N., & Lang, P. J. (2001). Emotion and motivation I: defensive and appetitive reactions in picture processing. *Emotion*, *1*(3), 276-298.

Bradley, M. M., Cuthbert, B. N., & Lang, P. J. (1996). Lateralized startle probes in the study of emotion. *Psychophysiology*, *33*(2), 156-161.

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, *45*(4), 602-607.

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.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*(3), 904-911.

Cacioppo, J. T., Berntson, G. G., Larsen, J. T., Poehlmann, K. M., & Ito, T. A. (2000). The psychophysiology of emotion. In M. Lewis & J.M. Haviland-Jones (Eds.), *Handbook of Emotions, 2nd Edition* (pp. 173-191). New York, NY: Guildford Press.

Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1997). Beyond bipolar conceptualizations and measures: The case of attitudes and evaluative space. *Personality and Social Psychology Review*, *1*(1), 3-25.

Carver, C. S. & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology*, *67*(2), 319-333.

Chen, M. & Bargh, J. A. (1999). Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus. *Personality and Social Psychology Bulletin*, *25*(2), 215-224.

Citron, F. M. (2012). Neural correlates of written emotion word processing: A review of recent electrophysiological and hemodynamic neuroimaging studies. *Brain and Language*, *122*(3), 211-226.

Citron, F. M., Gray, M. A., Critchley, H. D., Weekes, B. S., & Ferstl, E. C. (2014). Emotional valence and arousal affect reading in an interactive way: neuroimaging evidence for an approach-withdrawal framework. *Neuropsychologia*, *56*, 79-89.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, *33*(4), 497-505.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204-256.

Connell, L., & Lynott, D. (2012). Strength of perceptual experience predicts word processing performance better than concreteness or imageability. *Cognition*, *125*(3), 452-465.

Conner, T. & Barrett, L. F. (2005). Implicit self-attitudes predict spontaneous affect in daily life. *Emotion*, *5*(4), 476-488.

Cortese, M. J.,& Schock, J. (2013). Imageability and age of acquisition effects in disyllabic word recognition. *The Quarterly Journal of Experimental Psychology*, *66*(5), 946-972.

Cosmides, L. & Tooby, J. (2000). Evolutionary psychology and the emotions. In M. Lewis & J.M. Haviland-Jones (Eds.), *Handbook of Emotions, 2nd Edition* (pp. 91-115). New York, NY: Guildford Press.

Damasio, A. R. (1998). Emotion in the perspective of an integrated nervous system. *Brain Research Reviews*, *26*(2), 83-86.

de Groot, A. M. (1989). Representational aspects of word imageability and word frequency as assessed through word association. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(5), 824-845.

De Houwer, J. D., & Hermans, D. (1994). Differences in the affective processing of words and pictures. *Cognition & Emotion*, *8*(1), 1-20.

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

De Houwer, J., Thomas, S., & Baeyens, F. (2001). Association learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological Bulletin*, *127*(6), 853-869.

Doerksen, S. & Shimamura, A. P. (2001). Source memory enhancement for emotional words. *Emotion*, *1*(1), 5-11.

Dolan, R. J. (2002). Emotion, cognition, and behavior. *Science*, *298*(5596), 1191-1194.

Dresler, T., Mériau, K., Heekeren, H. R., & van der Meer, E. (2009). Emotional Stroop task: effect of word arousal and subject anxiety on emotional interference. *Psychological Research*, *73*(3), 364-371.

Duncan, S., & Barrett, L. F. (2007). Affect is a form of cognition: A neurobiological analysis. *Cognition and Emotion*, *21*(6), 1184-1211.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*(3-4), 169-200.

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, *128*(2), 203-235.

Elliot, A. J., & Thrash, T. M. (2010). Approach and avoidance temperament as basic dimensions of personality. *Journal of Personality*, *78*(3), 865-906.

Estes, Z. & Adelman, J. S. (2008). Automatic vigilance for negative words is categorical and general. *Emotion, 8*(4), 453-457.

Estes, Z., & Verges, M. (2008). Freeze or flee? Negative stimuli elicit selective responding. *Cognition*, *108*(2), 557-565.

Feldman, L. A. (1995). Valence focus and arousal focus: Individual differences in the structure of affective experience. *Journal of personality and social psychology*, *69*(1), 153-166.

Feldman, L. B. and Basnight-Brown, D. (2007). The role of morphology in visual word recognition: Graded semantic influences due to competing senses and semantic richness of the stem.In Grigorenko, E. L. & Naples, A. (Eds.) *Single-word reading: Cognitive, behavioral and biological perspectives.* Mahwah, NJ: Lawrence Erlbaum Ass.

Feng, M. C., Courtney, C. G., Mather, M., Dawson, M. E., & Davison, G. C. (2011). Age-related affective modulation of the startle eyeblink response: Older adults startle most when viewing positive pictures. *Psychology and aging*, *26*(3), 752-760.

Ferré, P., Guasch, M., Moldovan, C., & Sánchez-Casas, R. (2012). Affective norms for 380 Spanish words belonging to three different semantic categories. *Behavior Research Methods*, *44*(2), 395-403.

Fox, E., Russo, R., Bowles, R., & Dutton, K. (2001). Do threatening stimuli draw or hold visual attention in subclinical anxiety? *Journal of Experimental Psychology: General*, *130*(4), 681.

Freese, J. L., & Amaral, D. G. (2005). The organization of projections from the amygdala to visual cortical areas TE and V1 in the macaque monkey. *Journal of Comparative Neurology*, *486*(4), 295-317.

Fridlund, A. J., Ekman, P., & Oster, H. (1987). Facial expressions of emotion. In Siegman, A.W. & Feldstein, S. (Eds.), *Nonverbal behaviour and communication, 2nd edition* (pp. 143-223). Hillsdale, NY: Lawrence Erlbaum.

Friedman, B.H. (2003). Idiodynamics vis a vis psychophysiology: An idiodynamic portrayal of cardiovascular reactivity. *Journal of Applied Psychoanalytic Studies, 5*, 425–441.

Gerhand, S. & Barry, C. (1999b). Age of acquisition, word frequency, and the role of phonology in the lexical decision task. *Memory & Cognition*, *27*(4), 592-602.

Gerhand, S., & Barry, C. (1999a). Age-of-acquisition and frequency effects in speeded word naming. *Cognition*, *73*(2), B27-B36.

Ghashghaei, H. T. & Barbas, H. (2002). Pathways for emotion: Interactions of prefrontal and anterior temporal pathways in the amygdala of the rhesus monkey. *Neuroscience*, *115*(4), 1261-1279.

Gilet, A. L. & Jallais, C. (2011). Valence, arousal and word associations. *Cognition and Emotion*, *25*(4), 740-746.

Gray, L. N. & Tallman, I. (1987). Theories of choice: Contingent reward and punishment applications. *Social Psychology Quarterly, 50*, 16-23.

Hanley, J. R., Hunt, R. P., Steed, D. A., & Jackman, S. (2013). Concreteness and word production. *Memory & Cognition*, *41*(3), 365-377.

Harmon-Jones, E., Lueck, L., Fearn, M., & Harmon-Jones, C. (2006). The effect of personal relevance and approach-related action expectation on relative left frontal cortical activity. *Psychological Science*, *17*(5), 434-440.

Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist*, *52*(12), 1280-1300.

Hofmann, M. J., Kuchinke, L., Tamm, S., Võ, M. L., & Jacobs, A. M. (2009). Affective processing within 1/10th of a second: High arousal is necessary for early facilitative processing of negative but not positive words. *Cognitive, Affective, & Behavioral Neuroscience*, *9*(4), 389-397.

Horstmann, G., Scharlau, I., & Ansorge, U. (2006). More efficient rejection of happy than of angry face distractors in visual search. *Psychonomic Bulletin & Review*, *13*(6), 1067-1073.

Hulme, C., Stuart, G., Brown, G. D., & Morin, C. (2003). High-and low-frequency words are recalled equally well in alternating lists: Evidence for associative effects in serial recall. *Journal of Memory and Language*, *49*(4), 500-518.

Juhasz, B. J. (2005). Age-of-acquisition effects in word and picture identification. *Psychological Bulletin*, *131*(5), 684-712.

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. *The Quarterly Journal of Experimental Psychology*, *64*(9), 1683-1691.

Kagan, J. (2007). *What is emotion?* Birmingham, NY: Vail-Ballou Press.

Kahan, T. A, & Hely, C. D. (2008). The role of valence and frequency in the emotional Stroop task. *Psychonomic Bulletin & Review*, *15*(5), 956-960.

Kensinger, E. A. & Corkin, S. (2003a). Effect of negative emotional content on working memory and long-term memory. *Emotion*, *3*(4), 378-393.

Kensinger, E. A. & Corkin, S. (2003b). Memory enhancement for emotional words: Are emotional words more vividly remembered than neutral words? *Memory & Cognition*, *31*(8), 1169-1180.

Kensinger, E. A., & Schacter, D. L. (2006). Amygdala activity is associated with the successful encoding of item, but not source, information for positive and negative stimuli. *The Journal of Neuroscience*, *26*(9), 2564-2570.

Kissler, J., Herbert, C., Peyk, P., & Junghofer, M. (2007). Buzzwords: Early cortical responses to emotional words during reading. *Psychological Science*, *18*(6), 475-480.

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*(1), 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*(3), 473-481.

Krieglmeyer, R. & Deutsch, R. (2010). Comparing measures of approach–avoidance behaviour: The manikin task vs. two versions of the joystick task. *Cognition and Emotion*, *24*(5), 810-828.

Krieglmeyer, R., De Houwer, J., & Deutsch, R. (2011). How farsighted are behavioral tendencies of approach and avoidance? The effect of stimulus valence on immediate vs. ultimate distance change. *Journal of Experimental Social Psychology*, *47*(3), 622-627.

Kroll, J. F., & Merves, J. S. (1986). Lexical access for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(1), 92-107.

Kuchinke, L., Võ, M. L. H., Hofmann, M., & Jacobs, A. M. (2007). Pupillary responses during lexical decisions vary with word frequency but not emotional valence. *International Journal of Psychophysiology*, *65*(2), 132-140.

Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. B. (2014). Emotion and Language: Valence and Arousal Affect Word Recognition. *Journal of Experimental Psychology: General, 143*(3), 1065-1081.

Kuppens, P., Tuerlinckx, F., Russell, J. A., & Barrett, L. F. (2013). The relation between valence and arousal in subjective experience. *Psychological Bulletin*, *139*, 917-940.

Lane, R. D., Chua, P. M., & Dolan, R. J. (1999). Common effects of emotional valence, arousal and attention on neural activation during visual processing of pictures. *Neuropsychologia*, *37*(9), 989-997.

Lang, P. J. (1995). The emotion probe: Studies of motivation and attention. *American Psychologist*, *50*(5), 372-385.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1990). Emotion, attention, and the startle reflex. *Psychological Review*, *97*(3), 377-395.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1997). Motivated attention: Affect, activation, and action. In P.J. Lang, Simons, R.F. & Balaban, M. *Attention and orienting: Sensory and motivational processes* (pp. 97-135), New York: Psychology Press.

Lang, P. J., Greenwald, M. K., Bradley, M. M., & Hamm, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, *30*(3), 261-273.

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*(4), 445-452.

LeDoux, J. (1996). *The emotional brain: The mysterious underpinnings of emotional life*. New York: Touchstone.

Lench, H. C., Flores, S. A., & Bench, S. W. (2011). Discrete emotions predict changes in cognition, judgment, experience, behavior, and physiology: A meta-analysis of experimental emotion elicitations. *Psychological Bulletin*, *137*(5), 834.

Lewis, P. A., Critchley, H. D., Rotshtein, P., & Dolan, R. J. (2007). Neural correlates of processing valence and arousal in affective words. *Cerebral Cortex*, *17*(3), 742-748.

Mathewson, K. J., Arnell, K. M., & Mansfield, C. A. (2008). Capturing and holding attention: The impact of emotional words in rapid serial visual presentation. *Memory & Cognition*, *36*(1), 182-200.

Mauss, I. B. & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition and Emotion*, *23*(2), 209-237.

Mauss, I. B., Levenson, R. W., McCarter, L., Wilhelm, F. H., & Gross, J. J. (2005). The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion*, *5*(2), 175-190.

Mayer, J. D. & Gaschke, Y. N. (1988). The experience and meta-experience of mood. *Journal of Personality and Social Psychology*, *55*(1), 102-111.

McKenna, F. P. & Sharma, D. (1995). Intrusive cognitions: An investigation of the emotional Stroop task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(6), 1595-1607.

McKenna, F. P. & Sharma, D. (2004). Reversing the emotional Stroop effect reveals that it is not what it seems: The role of fast and slow components. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 382-392

Mesquita, B. & Markus, H. R. (2004). Culture and emotion: Models of agency as sources of cultural variation in emotion. In A.S.R. Manstead, N. Frijda, & A. Fischer (Eds), *Feelings and emotions: The Amsterdam symposium* (p. 341-358). New York: Cambridge University Press.

Mulligan, K. & Scherer, K. R. (2012). Toward a working definition of emotion. *Emotion Review*, *4*(4), 345-357.

Murphy, F. C., Nimmo-Smith, I. A. N., & Lawrence, A. D. (2003). Functional neuroanatomy of emotions: a meta-analysis. *Cognitive, Affective, & Behavioral Neuroscience*, *3*(3), 207-233.

New, B. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review*, *13*(1), 45-52.

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

Nishiyama, R. (2013). Dissociative contributions of semantic and lexical-phonological information to immediate recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 642-648.

Osgood, C. E. (1953). *Method and theory in experimental psychology*. New York: Oxford University Press.

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

Paivio, A. (2007). *Mind and its evolution: A dual coding theoretical approach*. New York: Psychology Press.

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International journal of human-computer studies*, *59*(1), 185-198.

Phan, K. L., Wager, T., Taylor, S. F., & Liberzon, I. (2002). Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *Neuroimage*, *16*(2), 331-348.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*(1), 56-115.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience*, *1*(3), 3-33.

Pratto, F. & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of Personality and Social Psychology, 61*, 380-391.

Puca, R. M., Rinkenauer, G., & Breidenstein, C. (2006). Individual differences in approach and avoidance movements: How the avoidance motive influences response force. *Journal of Personality*, *74*(4), 979-1014.

Purkis, H. M., Lipp, O. V., Edwards, M. S., & Barnes, R. (2009). An increase in stimulus arousal has differential effects on the processing speed of pleasant and unpleasant stimuli. *Motivation and Emotion*, *33*(4), 353-361.

Quinlan, P. T. (2003). Visual feature integration theory: past, present, and future. *Psychological Bulletin*, *129*(5), 643-673.

Rafaeli, E., & Revelle, W. (2006). A premature consensus: are happiness and sadness truly opposite affects? *Motivation and Emotion*, *30*(1), 1-12.

Raman, I., Baluch, B., & Besner, D. (2004). On the control of visual word recognition: Changing routes versus changing deadlines. *Memory & cognition*, *32*(3), 489-500.

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*(11), 1472-1484.

Rolls, E.T., Tovee, M.J., Purcell, D.G., Stewart, A.L. & Azzopardi, P. (1994). The responses of neurons in the temporal cortex of primates, and face identification and detection. *Experimental Brain Research, 101*, 473-484.

Rotshtein, P., Malach, R., Hadar, U., Graif, M., & Hendler, T. (2001). Feeling or features: different sensitivity to emotion in high-order visual cortex and amygdala. *Neuron*, *32*(4), 747-757.

Rozin, P., Berman, L., & Royzman, E. (2010). Biases in use of positive and negative words across twenty natural languages. *Cognition and Emotion*, *24*(3), 536-548.

Ruch, W. (1995). Will the real relationship between facial expression and affective experience please stand up: The case of exhilaration. *Cognition & Emotion*, *9*(1), 33-58.

Russell, J. A. & Carroll, J. M. (1999). On the bipolarity of positive and negative affect. *Psychological Bulletin*, *125*(1), 3-30.

Russell, J. A. & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, *11*(3), 273-294.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145-172.

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, *57*(3), 493-502.

Sakaki, M., Niki, K., & Mather, M. (2012). Beyond arousal and valence: The importance of the biological versus social relevance of emotional stimuli. *Cognitive, Affective, & Behavioral Neuroscience*, *12*(1), 115-139.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, *44*(4), 695-729.

Schwanenflugel, P. J. (1991). Contextual constraint and lexical processing. *Advances in Psychology*, *77*, 23-45.

Schwanenflugel, P. J., & Stowe, R. W. (1989). Context availability and the processing of abstract and concrete words in sentences. *Reading Research Quarterly*, *24*(1), 114-126.

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*(3), 783-792.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523-568.

Sharot, T. & Phelps, E. A. (2004). How arousal modulates memory: Disentangling the effects of attention and retention. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(3), 294-306.

Shweder, R. A. (1993). The cultural psychology of the emotions. In Lewis, M. & Haviland-Jones, J.M. (Eds.), *Handbook of Emotions, 2nd Edition* (pp. 417-431). New York, NY: Guildford Press.

Siakaluk, P. D., Knol, N., & Pexman, P. M. (2014). Effects of Emotional Experience for Abstract Words in the Stroop Task. *Cognitive Science*. doi: 10.1111/cogs.12137

Siakaluk, P. D., Pexman, P. M., Aguilera, L., Owen, W. J., & Sears, C. R. (2008). Evidence for the activation of sensorimotor information during visual word recognition: The body–object interaction effect. *Cognition*, *106*(1), 433-443.

Stins, J. F., Roelofs, K., Villan, J., Kooijman, K., Hagenaars, M. A., & Beek, P. J. (2011). Walk to me when I smile, step back when I'm angry: emotional faces modulate whole-body approach–avoidance behaviors. *Experimental Brain Research*, *212*(4), 603-611.

Storbeck, J., Robinson, M. D., & McCourt, M. E. (2006). Semantic processing precedes affect retrieval: The neurological case for cognitive primacy in visual processing. *Review of general psychology*, *10*(1), 41-55.

Strain, E., Patterson, K., & Seidenberg, M. S. (1995). Semantic effects in single-word naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(5), 1140-1154.

Tillotson, S. M., Siakaluk, P. D., & Pexman, P. M. (2008). Body—object interaction ratings for 1,618 monosyllabic nouns. *Behavior Research Methods*, *40*(4), 1075-1078.

Tomaszczyk, J. C., Fernandes, M. A., & MacLeod, C. M. (2008). Personal relevance modulates the positivity bias in recall of emotional pictures in older adults. *Psychonomic Bulletin & Review*, *15*(1), 191-196.

Treiman, R. & Chafetz, J. (1987). Are there onset-and rime-like units in printed words? In M.Coltheart (Ed.), *Attention and performance: The psychology of reading* (pp. 281-298). Hillsdale, NJ: Lawrence Erlbaum.

Unkelbach, C., von Hippel, W., Forgas, J. P., Robinson, M. D., Shakarchi, R. J., & Hawkins, C. (2010). Good things come easy: Subjective exposure frequency and the faster processing of positive information. *Social Cognition*, *28*(4), 538-555.

van Dantzig, S., Pecher, D., & Zwaan, R. A. (2008). Approach and avoidance as action effects. *The Quarterly Journal of Experimental Psychology*, *61*(9), 1298-1306.

Vigliocco, G., Vinson, D. P., Lewis, W., & Garrett, M. F. (2004). Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology*, *48*(4), 422-488.

Vinson, D., Ponari, M., & Vigliocco, G. (2014). How does emotional content affect lexical processing? *Cognition & Emotion*, *28*(4), 737-746.

Wallbott, H. G., & Scherer, K. R. (1991). Stress specificities: differential effects of coping style, gender, and type of stressor on autonomic arousal, facial expression, and subjective feeling. *Journal of Personality and Social Psychology*, *61*(1), 147-156.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191-1207.

Watson, D. & Clark, L. A. (1994). Emotions, moods, traits, and temperaments: Conceptual distinctions and empirical findings. In P. Ekman & R.J. Davidson (pp. 89-93), *The nature of emotion.* New York: Oxford University Press.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063-1070.

Watson, D., Wiese, D., Vaidya, J., & Tellegen, A. (1999). The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence. *Journal of Personality and Social Psychology*, *76*(5), 820-838.

Wentura, D., Rothermund, K., & Bak, P. (2000). Automatic vigilance: The attention-grabbing power of approach-and avoidance-related social information. *Journal of Personality and Social Psychology*, *78*(6), 1024-1037.

Williams, J. M. G., Mathews, A., & MacLeod, C. (1996). The emotional Stroop task and psychopathology. *Psychological Bulletin*, *120*(1), 3-24.

Wurm, L. & Vakoch, D.A. (2000). The adaptive value of lexical connotation in speech perception. *Cognition and Emotion, 14(2), 177-191.*

Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language*, *60*(4), 502-529.

Yates, M., Locker, L., & Simpson, G. B. (2004). The influence of phonological neighborhood on visual word perception. *Psychonomic Bulletin & Review*, *11*(3), 452-457.

Zajonc, R.B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology, 9*(2), 1-27.

Zhang, D., He, W., Wang, T., Luo, W., Zhu, X., Gu, R., Li, H. & Luo, Y. J. (2014). Three stages of emotional word processing: an ERP study with rapid serial visual presentation. *Social Cognitive and Affective Neuroscience*, doi: 10.1093/scan/nst188.

Zuckerman, M., & Lubin, B. (1985). *Manual for the MAACL-R: The Multiple Affect Adjective Check List Revised*. Educational and Industrial Testing Service.