Breaking Bad: Norms for Valence, Arousal, and Dominance for over 10k English Multiword Expressions

Saif M. Mohammad
National Research Council Canada
saif.mohammad@nrc-cnrc.gc.ca

Abstract

Factor analysis studies have shown that the primary dimensions of word meaning are *Valence (V)*, *Arousal (A)*, and *Dominance (D)*. Existing lexicons such as the NRC VAD Lexicon, published in 2018, include VAD association ratings for *words*. Here, we present a complement to it, which has human ratings of valence, arousal, and dominance for ∼10k English Multiword Expressions (MWEs) and their constituent words. We also increase the coverage of unigrams, especially words that have become more common since 2018. In all, the new NRC VAD Lexicon v2 now has entries for ∼10k MWEs and ∼25k words, in addition to the entries in v1. We show that the associations are highly reliable. We use the lexicon to examine emotional characteristics of MWEs, including: 1. The degree to which MWEs (idioms, noun compounds, and verb particle constructions) exhibit strong emotionality; 2. The degree of emotional compositionality in MWEs. The lexicon enables a wide variety of research in NLP, Psychology, Public Health, Digital Humanities, and Social Sciences. The NRC VAD Lexicon v2 is freely available through the project webpage: http://saifmohammad.com/WebPages/nrc-vad.html

1 Introduction

Several influential factor analysis studies have shown that the three most important, largely independent, dimensions of connotative meaning and emotions are valence (positiveness–negativeness/pleasure–displeasure), arousal (active–passive), and dominance (dominant–submissive, competent–incompetent, powerful–weak) (Osgood et al., 1957; Russell, 1980, 2003). We will refer to the three dimensions individually as *V, A,* and *D*, and together as *VAD*. Language is a powerful medium for expressing emotions (consciously and unconsciously) and language-resource work has produced large repositories of word–emotion associations and sentences annotated for emotions. However, we are not aware of any large scale work on multiword expressions (MWEs) and VAD.

MWEs have been defined with some differences in past works, but here we simply consider sequences of two or more words (often with some interesting semantic, syntactic, or functional property) as MWEs. Broadly speaking, MWEs are important in NLP, linguistics, social sciences, and psychology because their meaning is often not compositional (Smolka and Schulte im Walde, 2020) and MWEs reveal insights about the structure of language, social interaction, and cognitive processing. Yet, unlike their lexical or sentence cousins, far fewer language resources exist for MWEs.

Our work at the intersection of emotions and MWEs makes these contributions:

1. We obtained human ratings of valence, arousal, and dominance for about 10,000 common English MWEs.

2. We also obtained VAD ratings for about 25,000 English words that are not included in the NRC VAD Lexicon v1 (Mohammad, 2018). These include terms that have become more common since 2018 (such as *quarantine)* as well as words that are constituents of the MWEs included in the lexicon.

3. The scores are fine-grained real-valued numbers from -1 (lowest V, A, or D) to 1 (highest V, A, or D). We show that the annotations lead to reliable VAD score (split-half reliability scores of $r = 0.99$ for valence, $r = 0.98$ for arousal, and $r = 0.96$ for dominance.) We will refer to this lexicon as *MWE-VAD*. The new MWE and unigram annotations are

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1964–1988

added to the entries in NRC VAD Lexicon v1 to form the NRC VAD Lexicon v2.

4. We use the newly created lexicon to examine emotionality of MWEs, including:
   (a) The distributions of high- and low-valence MWEs in different types of MWEs such as idioms, noun compounds, and verb particle constructions. Similar analysis is done for arousal and dominance.
   (b) The degree of emotional compositionality in MWEs — i.e., to what extent the emotionality of an MWE can be determined from the emotionality of its constituent words?

5. Finally, we describe a number of research and application directions that can benefit from MWE-VAD. One can use MWE-VAD to study MWEs specifically and NRC VAD v2 for work on valence, arousal, and dominance, in general. NRC VAD Lexicon v2 (along with automatic translations of the English terms to over 100 languages) is made freely available for research through the project webpage.[1]

All of the annotation tasks described in this paper were approved by our institution's review board, which examined the methods to ensure that they were ethical. Special attention was paid to obtaining informed consent and protecting participant anonymity.

2 Related Work

Primary Dimensions of Meaning and Affect. Highly influential, psycholinguistics and affective science work by Osgood et al. (1957) and Russell (1980), respectively, asked human participants to rate words along dimensions of opposites such as *heavy–light, good–bad, strong–weak,* etc. Factor analysis of these judgments revealed that the three most prominent dimensions of connotative meaning and emotion are valence (*pleasure–displeasure*), arousal (*active–passive*), and dominance (*strong–weak*).

MWEs. MWE work in NLP has focused on the automatic discovery, processing, and understanding of MWEs from corpora (see surveys such as Smolka and Schulte im Walde (2020) and Constant et al. (2017)). Less work has gone into manually compiling lists of MWEs. Notably, Muraki et al. (2023a) manually compiled a lexicon of about 62,000 MWEs. They added concreteness ratings for the MWEs, as well as the frequencies of the MWEs in a subtitles corpus (Brysbaert et al., 2012). Takahashi et al. (2024) compiled a lexicon of 160,000 Japanese MWEs. Tong et al. (2024) compiled a 10k English metaphor–literal paraphrase pairs dataset. There is even less work on annotating MWEs for emotions, despite work on many *word*–emotion association lexicons. Jochim et al. (2018) crowdsourced the sentiment annotation of 5000 English idioms. Ibrahim et al. (2015) annotated 3600 Modern Standard Arabic idioms for sentiment.

Existing Affect Lexicons. The Bradley and Lang (1999) lexicon has more than 1000 words with real-valued scores of valence, arousal, and dominance. For each word, they asked annotators to rate valence, arousal, and dominance—for more than 1,000 words—on a 9-point rating scale. The ratings from multiple annotators were averaged to obtain a score between 1 (lowest V, A, or D) to 9 (highest V, A, or D). Their lexicon, called the *Affective Norms of English Words (ANEW)*, has since been widely used across many different fields of study. ANEW was also translated into non-English languages: e.g., Moors et al. (2013) for Dutch, Võ et al. (2009) for German, and Redondo et al. (2007) for Spanish. Warriner et al. (2013) created a VAD lexicon for more than 13,000 words, using a similar annotation method as for ANEW. The NRC VAD Lexicon v1 (Mohammad, 2018) is the largest manually created VAD lexicon (in any language). It has entries for about 20,000 English words. The NRC Emotion Lexicon was created by crowdsourcing and it includes association entries for about 14,000 words with eight Plutchik emotions as well as positive and negative sentiment (Mohammad and Turney, 2013, 2010).[2]

The 10K MWEs and many unigrams from VAD v2 have since been used in other annotation projects as well. The NRC WorryWords Lexicon

[1] **VAD v2:** http://saifmohammad.com/WebPages/nrc-vad.html
Emotion Dynamics Code (Vishnubhotla and Mohammad, 2022) to analyze emotions in text using emotion lexicons: https://github.com/Priya22/EmotionDynamics.

[2] http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

(Mohammad, 2024a) has real-valued scores indicating their associations with anxiety for roughly the same 44k English words and 10k MWEs. The non-neutral subset of the VAD v2 (those with valence scores lower or equal to −0.33 and those with valence scores higher or equal to 0.33) were used to create the NRC Words of Warmth (WoW) Lexicon. WoW is a list of about 31,000 English terms (∼26k unigrams and ∼5k MWEs) and real-valued scores indicating their associations with warmth, sociability, and trust—core dimensions (along with dominance aka competence) in social cognition and stereotypes (Fiske et al., 2002; Bodenhausen et al., 2012; Fiske, 2018; Abele et al., 2016; Koch et al., 2024).

Automatically Generated Affect Lexicons. There is a large body of work on automatically determining word–sentiment, word–emotion, and word–VAD associations, including: Strapparava and Valitutti (2004); Yang et al. (2007); Mohammad (2012); Mohammad and Kiritchenko (2015); Yu et al. (2015); Staiano and Guerini (2014); Bandhakavi et al. (2021); Muhammad et al. (2023) to name just a few. These methods, including those that employ large language models, often assign a real-valued score representing the degree of association. Our MWE-VAD Lexicon can enable further such work on MWEs, especially by keeping a portion as a source of seed/example words for training/few-shot learning and a held-out portion for evaluating the automatically generated lexicons.

3 Obtaining Human Ratings of Valence, Arousal, and Dominance

The keys steps in obtaining the new annotations were as follows:

1. selecting the terms to be annotated
2. developing the questionnaire
3. developing measures for quality control (QC)
4. annotating terms on a crowdsource platform
5. discarding data from outlier annotators (QC)
6. aggregating data from multiple annotators to determine the VAD association scores

We describe each of the steps below.

1. Term Selection. We wanted to include various kinds of multi-word expressions, including common phrases, light verb constructions, idiomatic constructions, etc. However, identifying MWEs from a large corpus of text is not trivial. Further, we wanted to include terms for which other linguistically interesting annotations already exist (such as concreteness ratings). Thus, for our work we chose the 10,500 most frequent MWEs compiled by Muraki et al. (2023a).

The NRC VAD Lexicon v1 includes about 20,000 common English words. However, with the passage of time, words that were less prominent earlier can become more commonly used: e.g., *quarantine, deepfake, lockdown, workstation,* and *gaslighting*. Further, we wanted to include words that were constituents of the MWEs chosen for annotation. This would allow for comparisons of the VAD scores of MWEs and their constituents. Finally, we wanted to include terms for which other linguistically interesting annotations already exists (such as concreteness and age of acquisition ratings). Therefore we included words from the Prevalence dataset (Brysbaert et al., 2019). This dataset has prevalence scores (how widely a word is known by English speakers), determined directly by asking people, for 62,000 lemmas. We included a term if it was marked as known to at least 70% of the people who provided responses for the term. From this set we removed terms that are common person names or city names; and also words already annotated for VAD in the NRC VAD lexicon. This resulted in close to 25k unigrams that we annotated for VAD.

2. VAD Questionnaires The questionnaires used to annotate the data were developed after several rounds of pilot annotations. Detailed directions, including notes directing respondents to consider predominant word sense (in case the word is ambiguous) and example questions (with suitable responses) were provided. (See Appendix.) The primary instruction and the questions presented to annotators are shown below.

VALENCE: Consider positive feelings (or positive sentiment) to be a broad category that includes:
positiveness / pleasure / goodness / happiness / greatness / brilliance / superiority / health etc.
Consider negative feelings (or negative sentiment) to be a category that includes:
negativeness / displeasure /badness / unhappiness / insignificance / terribleness / inferiority / sickness etc.
If you do not know the meaning of a word or are unsure, you can look it up in a dictionary (e.g., the Merriam Webster) or on the internet.

Quality Control

Some questions have pre-determined correct answers. If you mark these questions incorrectly, we will give you immediate feedback in a pop-up box. An occasional misanswer is okay. However, if the rate of misanswering is high (e.g., >20%), then all of one's HITs may be rejected.

Select the options that most English speakers will agree with.

Q1. <term> is often associated with:

3: very positive feelings
2: moderately positive feelings
1: slightly positive feelings
0: not associated with positive or negative feelings
-1: slightly negative feelings
-2: moderately negative feelings
-3: very negative feelings

AROUSAL: This task is about words and their association with activeness or arousal. Consider activeness or arousal to be a broad category that includes:

active, aroused, stimulated, frenzied, excited, jittery, alert, etc.

Consider inactiveness or calmness to be a broad category that includes:

inactive, calm, unaroused, passive, relaxed, sluggish, etc.

This task is not about sentiment. (For example, something can be positive and inactive (such as flower), positive and active (such as exercise and party), negative and active (such as murderer), and negative and inactive (such as negligent).

DOMINANCE: This task is about words and their association with dominance, competence, control of situation, or powerfulness. Consider dominance, competence, control of situation, or powerfulness to be a broad category that includes:

dominant, competent, in control of the situation, powerful, influential, important, autonomous, etc.

Consider submissiveness, incompetence, controlled by outside factors, or weakness to be a broad category that includes:

submissive, incompetent, not in control of the situation, weak, influenced, cared-for, guided, etc.

This task is not about sentiment. (For example, something can be positive and weak (such as a flower petal) and something can be negative and strong (such as tyrant).

---

3. Quality Control Measures. About 2% of the data was annotated beforehand by the authors and interspersed with the rest. These questions are referred to as *gold* (aka *control*) questions. Half of the gold questions were used to provide immediate feedback to the annotator (in the form of a popup on the screen) in case they mark them incorrectly. We refer to these as *popup gold*. This helps prevent the situation where one annotates a large number of instances without realizing that they are doing so incorrectly. It is possible, that some annotators share answers to gold questions with each other (despite this being against the terms of annotation). Thus, the other half of the gold questions were also separately used to track how well an annotator was doing the task, but for these gold questions no popup was displayed in case of errors. We refer to these as *no-popup gold*.

4. Crowdsourcing. We setup the annotation tasks on the crowdsourcing platform, *Mechanical Turk*. In the task settings, we specified that we needed annotations from nine people for each word. Annotators were free to provide responses to as many terms as they wished. The annotation task was approved by our institution's review board.

Demographics: Since location and culture can impact word–association norms, and because AMT workers are mostly from the US, we requested annotations from participants who live in USA and Canada. The average age of the respondents was 34 years. Among those that disclosed their gender, about 53% were female, 47% were male.[3]

5. Filtering. If an annotator's accuracy on the gold questions (popup or non-popup) fell below 80%, then they were refused further annotation, and all of their annotations were discarded (despite being paid for).

6. Aggregation. Every response was mapped to an integer from -3 (highly negative/inactive/submissive) to 3 (highly positive/active/dominant) as follows:

- highly positive/active/dominant: 3
- moderately positive/active/dominant: 2
- slightly positive/active/dominant: 1
- neither positive/active/dominant nor negative/inactive/submissive: 0
- slightly negative/inactive/submissive: -1
- moderately negative/inactive/submissive: -2
- highly negative/inactive/submissive: -3

The final score for each term is simply the average score it received from each of the annotators. The scores were then linearly transformed to the interval: -1 (highest negativeness/inactivity/submissiveness) to 1 (highest positiveness/activity/dominance). See Table 1 for summary statistics.

Version	#Words	#MWEs	#Total
NRC VAD v1 (2018)	19,839	132	19,971
MWE VAD (2025)	25,089	10,073	35,162
NRC VAD v2 (2025)	44,928	10,205	55,133

Table 1: Number of terms in the NRC VAD v1, MWE VAD, and the combined lexicon (NRC VAD v2).

4 Reliability of the Annotations

A useful measure of quality is the reproducibility of the end result—repeated independent manual annotations from multiple respondents should

[3]Respondents were shown optional text boxes to disclose their demographic information as they choose; especially important for social constructs such as gender, in order to give agency to the respondents and to avoid binary language.

Dimension	Avg. #Annot.	SHR (ρ)	SHR (r)
valence	7.83	0.98	0.99
arousal	7.96	0.97	0.98
dominance	8.06	0.96	0.96

Table 2: Average number of annotations per term and split half reliability measured through both Spearman rank (ρ) and Pearson's (r) correlations. Scores in the 0.9s indicate high reliability.

result in similar scores. To assess this reproducibility, we calculate average *split-half reliability (SHR)* over 1000 trials. SHR is a common way to determine reliability of responses to generate scores on an ordinal scale (Weir, 2005). All annotations for an item are randomly split into two halves. Two separate sets of scores are aggregated, just as described in Section 3 (bullet 6), from the two halves. Then we determine how close the two sets of scores are (using a metric of correlation). This is repeated 1000 times and the correlations are averaged. The last two columns in Table 2 show the results (split half-reliabilities). Spearman rank and Pearson correlation scores of over 0.95 for V, A, and D indicate high reliability of the real-valued scores obtained from the annotations. (For reference, if the annotations were random, then repeat annotations would have led to an SHR of 0. Perfectly consistent repeated annotations lead to an SHR of 1. Also, similar past work on word–anxiety associations had SHR scores in the 0.8s (Mohammad, 2024b).)

5 How Commonly do MWEs Convey Strong Emotionality?

MWEs can be of different types such as noun compounds (non–noun collocations), idioms/fixed expressions, particle verb constructions, etc. Each of these types is relevant to expressing strong emotionality (high and low VAD rather than just neutral VAD). For example, *breath of fresh air, over the moon, kicked the bucket*, and *cold shoulder* are idiomatic expressions conveying various levels of valence. Similarly, *make a scene, take a leap, take a seat,* and *make a wish* are light verb constructions that convey different levels of arousal. And, *power move, victory lap, support system, death spiral,* etc. are noun compounds conveying various levels of dominance. Yet, we do not know the extent to which these different types of MWEs are associated with high and low VAD: e.g., how common is high dominance association in light verb

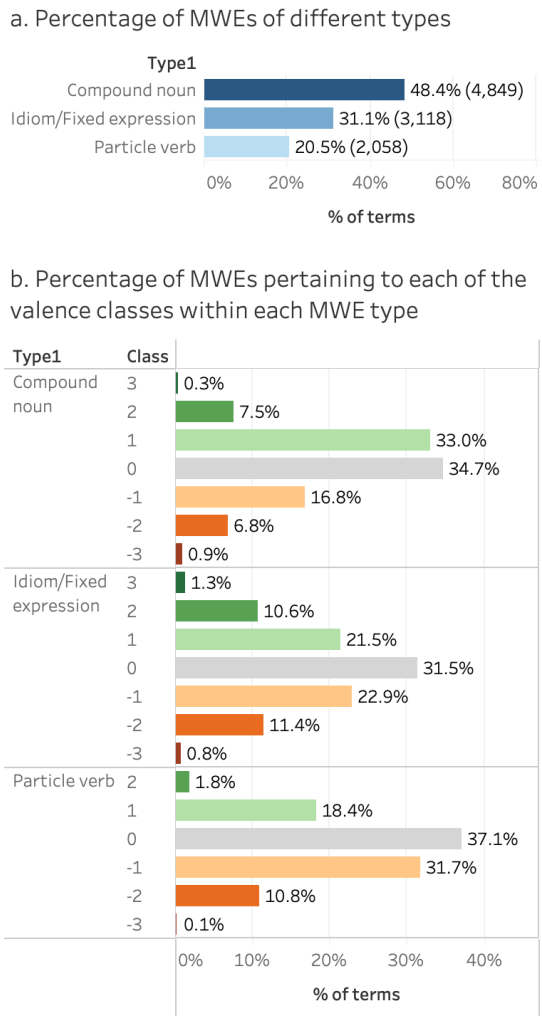


Figure 1: Distributions of MWE types.

constructions? Knowing these distributions will shed more light on how we use different types of MWEs to express emotions.

Each MWE entry in the MWE concreteness norms dataset (Muraki et al., 2023b) is marked with information about its MWE type. We make use of this to determine (a) percentage of different types of MWEs — shown in Figure 1 (a); and (b) the percentage of MWEs in various valence classes within various types of MWEs — shown in in Figure 1 (b). (The numbers within each type sum up to 100%.) Similar plots for arousal and dominance are shown in Figures 12 and 13, respectively (in the Appendix).

Results: From Figure 1 (a) we see that compound nouns are the most frequent class of MWEs in MWE-VAD (∼48%), followed by fixed expressions and then particle verbs. Observe in 1 (b) that the percentage of non-neutral MWEs varies

from about 63% in particle verb constructions to 69% in idioms and fixed expressions. Thus idioms seem to be particularly useful in terms of conveying valence. Further, in all three MWE types, the proportions for the low-valence (negative) classes are higher than the proportions for the high-valence (positive) classes. This is inline with what was found for words (Schrauf and Sanchez, 2004; Mohammad, 2018), and consistent with the hypotheses: 1. it is evolutionarily more useful to clearly identify different negative valence situations (requiring a greater vocabulary of negative words and expressions) than positive situations; and 2. human beings spend more time and more effort in thinking about negative experiences, thereby coming up with more negative words for the more careful and detailed deliberation (Schrauf and Sanchez, 2004).

The arousal and dominance distributions (shown in the Appendix) reveal that idioms and particle verb constructions have more low arousal and low dominance MWEs (than high A/D), whereas noun compounds have markedly more high arousal and high dominance expressions (than low A/D). This indicates that high-dominance and high-arousal concepts are more likely to be lexicalized (i.e., turned into fixed phrases or compounds). Power-related entities (e.g., *leader, boss, commander*) are culturally and cognitively salient and thus appear in compound forms more frequently: for example, *power play, executive order, master plan,* etc. High-arousal experiences more often result in visible actions, events, or consequences. This makes them easier to name and more likely to become lexicalized as noun compounds: e.g., *car crash, power surge, stress test*, etc.

Overall, these results show that all three types of MWEs are substantial sources of expressing high and low VAD, and that some types of MWEs are more amenable to express some types of emotionality (e.g., noun compounds more likely to express high arousal and dominance).

6 Emotional Compositionality of MWEs

MWEs are especially interesting because often their meaning is noncompositional. Thus, the emotionality of the MWE may not be predictable from the emotionality of its constituent words. Yet, little is known about the extent to which this is true. Further, we do not know the extent to which neutral words come together to create high- or low-valence/arousal/dominance MWEs.

To explore this, we focused on the 8,330 bigram (two-word sequence) MWEs. We will refer to the first word in the MWE as word1 and second as word2. For each of the dimensions (V/A/D), we partitioned them into 49 (7*7) bins corresponding to every class combination of word1 and word2: high V/A/D–high V/A/D, high V/A/D–moderate V/A/D,..., neutral–neutral,..., low V/A/D–low V/A/D. We then determined the average V/A/D scores of all the MWEs in each bin. We show the results for valence in Figure 2 (a). For arousal and dominance the results are shown in Figures 14 (a) and 15 (a), respectively, in the Appendix. For each of the 49 bins, we also computed the percentages of bigram MWEs associated with high V/A/D (Figures 2 (b), 14 (b), 15 (b)) and the percentages of bigram MWEs associated with low V/A/D (Figures 2 (c), 14 (c), 15 c)).

Results: Observe in Figure 2 (a) that as word1 and word2 (categorical) bin scores increase, the average valence score of the bigram bins also increases. This trend also exists for arousal and dominance (Figures 14 (a), 15 (a)), but it is markedly weaker. The position of the higher V/A/D word (whether word1 or word2) does not seem to impact the scores much (corresponding scores on opposite sides of the diagonal are roughly similar). Overall this suggests a marked degree of compositionality for all three dimensions, and more so for valence.

However, an examination of Figure 2 (b) reveals that: the percentage of high-valence MWEs for which both word1 and word2 are neutral is not negligible (1.66%); and markedly higher than the percentage for many other cells. Figure 2 (c) shows that the percentage of low-valence MWEs for which both word1 and word2 are neutral is 4.79% — highest among all cells. Thus a large number of low-valence MWEs have neutral constituents; showing a high amount of noncompositionality. The trends for arousal and dominance are also similar (Figures 14 and 15) showing that a large number of high and low A/D MWEs have neutral constituents.

Thus, overall, we conclude that a marked number of non-neutral MWEs are made up of neutral constituents; and that there is marked amount of emotional noncompositionality in MWEs (more so w.r.t. arousal and dominance than valence).

a. Average valence score of bigrams (words1 word2)

Word1-Class	Word2-Class: -3	-2	-1	0	1	2	3
-3	-0.70	-0.68	-0.64	-0.61	-0.57	-0.56	-0.22
-2	-0.79	-0.53	-0.39	-0.40	-0.43	-0.41	-0.14
-1	-0.63	-0.48	-0.29	-0.25	-0.14	-0.13	0.00
0	-0.65	-0.40	-0.24	-0.09	0.04	0.11	0.40
1	-0.59	-0.34	-0.10	0.03	0.16	0.23	0.45
2	-0.54	-0.32	-0.12	0.17	0.29	0.35	0.60
3	-0.47	-0.12	0.28	0.39	0.53	0.61	0.74

Avg. valence score of bigram
-0.75 0.75

b. Percentage of bigrams associated with high valence (V >= 0.33)

Word1-Class	Word2-Class: -3	-2	-1	0	1	2	3
-3	0.000	0.000	0.000	0.000	0.000	0.000	0.012
-2	0.000	0.012	0.012	0.036	0.048	0.024	0.024
-1	0.000	0.000	0.024	0.252	0.300	0.348	0.048
0	0.000	0.012	0.132	1.656	2.220	1.932	0.480
1	0.000	0.096	0.312	3.577	4.573	3.108	0.768
2	0.000	0.048	0.132	3.841	4.057	2.856	0.528
3	0.000	0.012	0.180	1.044	1.428	0.888	0.240

% terms -0.217 4.573

c. Percentage of bigrams associated with low valence (V <= -0.33)

Word1-Class	Word2-Class: -3	-2	-1	0	1	2	3
-3	0.120	0.132	0.180	0.792	0.384	0.228	0.036
-2	0.144	0.432	0.336	2.460	1.308	0.816	0.036
-1	0.156	0.552	0.540	3.253	1.104	0.756	0.036
0	0.432	1.272	1.284	4.789	1.296	0.756	0.024
1	0.432	0.996	0.780	2.556	0.636	0.324	0.012
2	0.252	0.444	0.336	0.660	0.120	0.072	0.000
3	0.024	0.072	0.012	0.000	0.012	0.000	0.000

% terms -0.227 4.789

Figure 2: Measures of Valence Compositionality.

7 Applications and Future Work

The large number of entries in the VAD Lexicon (for words and MWEs) make it useful for a number of research inquiries and applications. We list a few below.

Especially relevant to MWEs

1. High- and low-valence MWEs reflect key psychological processes related to affect, motivation, memory, attention, and social communication. These MWEs like *walk on air* (high valence) and *rock bottom* (low valence) encode strong emotional meaning beyond the sum of their constituents. They can be used to study psychological processes such as reward processing, approach motivation, social bonding, resilience/coping, and self-evaluation.

2. High-arousal MWEs such as *blow your mind*, *burst into tears*, *shake with rage*, and *on edge*, often reflect, trigger, or describe intense physiological and psychological states. They can be used to study various psychological processes such as:

   Physiological Activation (Autonomic Arousal): states of heightened bodily activation (elevated heart rate, muscle tension, increased adrenaline, etc.). Examples: *heart pounding* and *on fire* (with excitement).

   Fight-or-Flight or Freeze Responses: Many MWEs metaphorically encode evolutionary survival responses, such as fight (aggression), flight (fear), or freeze (shock). Examples: *Jump out of your skin* and *Frozen with fear*.

3. High-dominance MWEs, such as *take charge*, *lay down the law*, *crack the whip*, and *have the upper hand*, reflect psychological processes tied to a strong sense of self-agency, where the subject is the initiator of actions and outcomes. Thus they can be useful in studying social power, assertiveness, threat readiness, and competence signaling. In contrast, low-dominance MWEs, such as *at the mercy of*, *in over one's head, thrown under the bus*, and *toe the line*, reflect (and can be used to study) submission, helplessness, passivity, or external control. They often mark vulnerability or a lack of agency.

4. MWEs such as *over the moon*, *make someone's day, out of sorts*, and *keep your cool* have non-literal meanings that cannot be determined simply from the meanings of their constituents. They can be especially revealing in how people store and retrieve language chunks of meaning. Thus, both linguists and developers of artificial chat agents benefit from a large repository of MWE–VAD associations.

5. MWEs are often laden with rich connotative meaning, conveying subtle nuances of emotional intensity, politeness, formality, or social acceptability (Sag et al., 2002; Zgusta, 1967; Citron et al., 2019; Allawama et al., 2025). MWEs such as *losing control, on edge, on cloud nine*, *at peace with myself*, or *on top of the world* encode folk psychological concepts of emotional states. These expressions reveal how people naturally talk about and categorize emotion, which helps researchers build models of affect. MWEs often have an emotional punch, thereby influencing perception, recall, and judgment (Citron et al., 2019). Thus, MWEs like *war on drugs* and *family values* are used to frame complex issues in persuasive ways. Therefore, MWEs, especially those associated with high and low V/A/D, are highly relevant to studying discourse analysis and political rhetoric.

6. MWEs often draw on physical and embodied metaphors (Kacinik, 2014). Examples of emotional and embodied MWEs include: *feeling down* (valence is verticality: low valence is low and high valence is high), glow with happiness (valence is degree of light and high valence is bright whereas low valence is dark) boil with rage (anger is heat and strong arousal raises internal temperature), running on empty (body is a machine), step up (power is elevation), etc. Thus MWEs are a window into embodied and metaphorical thinking. MWEs associated with high and low V/A/D can be used to study how emotional language is grounded in sensorimotor experience and is organized metaphorically.

7. MWEs frequently reflect emotion regulation strategies, both maladaptive and adaptive (Nichter, 2010; Lee, 2017; Cole et al., 2010).

Examples of MWEs pertaining to emotion regulation, include: *bottling it up, trying to push it down, taking a deep breath,* and *letting it go*. They give insight into implicit self-regulatory processes people engage in during emotional episodes. MWEs provide linguistic evidence for how affect interacts with attention, memory, appraisal, and prediction (core components of emotion theories). MWE-VAD can be used to study emotion regulation strategies and inform theories of emotion by showing how people encode appraisals and attention patterns in everyday language.

Relevant Generally (to Words and MWEs)

1. Understanding valence, arousal, and dominance, and the underlying mechanisms; how VAD relate to our mind and body; how VAD change with age, socio-economic status, weather, green spaces, etc.
2. Determining how VAD manifest in language; how language shapes our VAD; how culture shapes the language of VAD; etc.
3. Tracking the degree of VAD towards targets of interest such as climate change, government policies, biological vectors, etc.
4. Developing automatic systems for detecting VAD; To provide features for automatic sentiment or emotion detection systems. They can also be used to obtain sentiment-specific word embeddings and sentiment-specific sentence representations.
5. MWE-VAD can be used to study emotions in story telling; its relationship with central elements of narratology such as conflict and resilience. To identify high V, A, and D words and MWEs in books and literature. To facilitate work of researchers in digital humanities. To facilitate work on literary analysis.
6. MWE-VAD is a source of gold (reference) scores, the entries in the VAD lexicon can be used in the evaluation of automatic methods of determining word–VAD associations.

Thus language resources at the intersection of MWEs and VAD are highly relevant to our understanding of a wide variety of phenomena. Note that automatic prediction of valence/sentiment/emotions from individual text instances is only a small part of the use cases. The lexicon can be used to obtain new insights on a wide variety of research questions (including those that are most directly answered by the lexicon rather than by using some ML system or LLM). Finally, even though large language models can at times be used in place of lexicons, any inferences drawn from an automatic approach requires manual validation. Portions of the manually created VAD lexicon presented here can be used to improve the generations of the LLM and held out portions can be used to validate the LLM generations.

8 Conclusions

We present here the MWE-VAD Lexicon, which has human ratings of valence, arousal, and dominance for more than 10,000 English MWEs and 25,000 unigrams. Notably the 25k unigrams are words not included in the NRC VAD Lexicon v1, and so greatly increasing coverage for unigrams. We show that the ratings are highly reliable (split-half reliability of over 0.95 for all three dimensions). We add these entries to those in NRC VAD Lexicon v1 to create v2. We use the lexicon to study the the extent to which different MWE types express strong emotionality. We also quantify the degree of emotional compositionality of MWEs with various metrics. Finally, we make a case for why language resources of MWEs associated with valence, arousal, and dominance are useful for a wide array of research inquiries and applications in Psychology, NLP, Public Health, Digital Humanities, and Social Sciences. The NRC VAD Lexicon v2 is freely available for research through the project webpage.[4]

9 Limitations

The lexicon created is one of the largest that exist with wide coverage and a large number of annotators (thousands of people as opposed to just a handful). However, no lexicon can cover the full range of linguistic and cultural diversity in emotion expression. The lexicons are restricted to words that are most commonly used in Standard American English and they capture emotion associations as judged by American native speakers of English. Annotators on Mechanical Turk are not representative of the wider US population. However, obtaining annotations from a large number of annotators (as we do) makes the lexicon more resilient to individual biases and captures more di-

[4] http://saifmohammad.com/WebPages/nrc-vad.html

versity in beliefs. We see this work as a first step that paves the way for more work using responses from various other groups of people and in various other languages. We built our lexicon using many of the principles and ideas listed in Mohammad (2023), which provides a detailed discussion of the limitations and best-practices in the creation and use of emotion lexicons.

10 Ethics and Data Statement

The crowd-sourced task presented in this paper was approved by our Institutional Research Ethics Board. The individual words and MWEs selected did not pose any risks beyond the risks of occasionally reading text on the internet. The annotators were free to do as many word and MWE annotations as they wished. The instructions included a brief description of the purpose of the task (Figures 3 through 11).

VAD assessments are complex, nuanced, and often instantaneous mental judgments. Additionally, each individual may use language to convey these assessments slightly differently. See Mohammad (2023) for a discussion of good practices and ethical considerations when using emotion lexicons. See Mohammad (2022) for a broader discussion of ethical considerations relevant to automatic emotion recognition. We discuss below notable points of discussion as well as some new and updated points especially relevant for VAD and MWE norms.

1. *Coverage:* We sampled a large number of English words from other lexical sources (which themselves sample from many sources). Yet, the words included do not cover all domains, genres, and people of different locations, socio-economic strata, etc. equally. It likely includes more of the vocabulary and MWEs used by people in the United States and with socio-economic and educational backgrounds that allow for technology access.
2. *Word Senses and Sense Priors:* Words when used in different senses and contexts may be associated with different degrees of VAD associations. The entries in in the VAD Lexicon are indicative of the associations with the predominant senses of the words. This is usually not problematic because most words have a highly dominant main sense (which occurs much more frequently than the other senses). In specialized domains, some terms might have a different dominant sense than in general usage. Entries in the lexicon for such terms should be appropriately updated or removed. Further, any conclusions using the lexicon should be made based on relative change of associations using a large number of textual tokens. For example, if there is a marked increase in low-valence words from one period to the next, where each period has thousands of word tokens, then the impact of word sense ambiguity is minimal, and it is likely that some broader phenomenon is causing the marked increase in low-valence words. (See last two bullets.)
3. *Not Immutable:* The VAD scores do not indicate an inherent unchangeable attribute. The associations can change with time (e.g., the decrease in negativeness associated with *inter-race relationships* over the last 100 years), but the lexicon entries are fixed. They pertain to the time they are created. However, they can be updated with time.
4. *Socio-Cultural Biases:* Many multiword expressions have origins and connotations in historic racism and bigotry, e.g., *sold down the river, grandfathered in,* and *black sheep*. Many have argued that, in everyday speech, choosing alternative expressions fosters more inclusiveness. On the other other hand, use of such expressions in research can shed light on the historical and social context of racism and stereotypes permeate language.
   The annotations for VAD capture various human biases. These biases may be systematically different for different socio-cultural groups. Our data was annotated by mostly US English speakers, but even within a country there are many diverse socio-cultural groups. Notably, crowd annotators on Amazon Mechanical Turk do not reflect populations at large. In the US for example, they tend to skew towards male, white, and younger people. However, compared to studies that involve just a handful of annotators, crowd annotations benefit from drawing on hundreds and thousands of annotators (such as this work).
   Our dataset curation was careful to avoid words and MWEs from problematic sources. We also asked people annotate terms based on what most English speakers think (as opposed to what they themselves think). This

helps to some extent, but the lexicon may still capture some historical VAD associations with certain identity groups. This can be useful for some socio-cultural studies; but we also caution that VAD associations with identity groups be carefully contextualized.

5. *Perceptions (not "right" or "correct" labels):* Our goal here was to identify common perceptions of WTS association. These are not meant to be "correct" or "right" answers, but rather what the majority of the annotators believe based on their intuitions of the English language.
6. *Avoid Essentialism:* When using the lexicon alone, it is more appropriate to make claims about VAD word usage rather than the VAD of the speakers. For example, *'the use of high-valence words in the context of the target group grew by 20%'* rather than *'valence in the target group grew by 20%'*. In certain contexts, and with additional information, the inferences from word usage can be used to make broader VAD claims.
7. *Avoid Over Claiming:* Inferences drawn from larger amounts of text are often more reliable than those drawn from small amounts of text. For example, *'the use of high-valence words grew by 20%'* is informative when determined from hundreds, thousands, tens of thousands, or more instances. Do not draw inferences about a single sentence or utterance from the VAD associations of its constituent words.
8. *Embrace Comparative Analyses:* Comparative analyses can be much more useful than stand-alone analyses. Often, VAD word counts and percentages on their own are not very useful. For example, *'the use of high-valence words grew by 20% when compared to [data from last year, data from a different person, etc.]'* is more useful than saying *'on average, 5 high-valence words were used in every 100 words'*.

We recommend careful reflection of ethical considerations relevant for the specific context of deployment when using the VAD lexicon.

References

Andrea E Abele, Nicole Hauke, Kim Peters, Eva Louvet, Aleksandra Szymkow, and Yanping Duan. 2016. Facets of the fundamental content dimensions: Agency with competence and assertiveness—communion with warmth and morality. *Frontiers in psychology*, 7:1810.

Ashraf Allawama, Aseel Zibin, Abdel Rahman Altakhaineh, and 1 others. 2025. Idioms as gateways to emotional expressions of sadness and joy in french. *Journal of Intercultural Communication*, 25(1):83–97.

Anil Bandhakavi, Nirmalie Wiratunga, Stewart Massie, and Deepak P. 2021. Emotion-aware polarity lexicons for twitter sentiment analysis. *Expert systems*, 38(7):e12332.

Galen V Bodenhausen, Sonia K Kang, and Destiny Peery. 2012. Social categorization and the perception of social groups. *The Sage handbook of social cognition*, pages 311–329.

Margaret M Bradley and Peter J Lang. 1999. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report, The Center for Research in Psychophysiology, University of Florida.

Marc Brysbaert, Paweł Mandera, Samantha F McCormick, and Emmanuel Keuleers. 2019. Word prevalence norms for 62,000 english lemmas. *Behavior research methods*, 51:467–479.

Marc Brysbaert, Boris New, and Emmanuel Keuleers. 2012. Adding part-of-speech information to the subtlex-us word frequencies. *Behavior research methods*, 44:991–997.

Francesca MM Citron, Cristina Cacciari, Jakob M Funcke, Chun-Ting Hsu, and Arthur M Jacobs. 2019. Idiomatic expressions evoke stronger emotional responses in the brain than literal sentences. *Neuropsychologia*, 131:233–248.

Pamela M Cole, Laura Marie Armstrong, and Caroline K Pemberton. 2010. The role of language in the development of emotion regulation.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Survey: Multiword expression processing: A Survey. *Computational Linguistics*, 43(4):837–892.

Susan Fiske, Amy Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82:878–902.

Susan T Fiske. 2018. Stereotype content: Warmth and competence endure. *Current directions in psychological science*, 27(2):67–73.

Hossam S Ibrahim, Sherif M Abdou, and Mervat Gheith. 2015. Idioms-proverbs lexicon for modern standard arabic and colloquial sentiment analysis. *arXiv preprint arXiv:1506.01906*.

Charles Jochim, Francesca Bonin, Roy Bar-Haim, and Noam Slonim. 2018. SLIDE - a sentiment lexicon of common idioms. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Natalie A Kacinik. 2014. Sticking your neck out and burying the hatchet: what idioms reveal about embodied simulation. *Frontiers in human neuroscience*, 8:689.

Alex Koch, Austin Smith, Susan T Fiske, Andrea E Abele, Naomi Ellemers, and Vincent Yzerbyt. 2024. Validating a brief measure of four facets of social evaluation. *Behavior Research Methods*, 56(8):8521–8539.

Sophia Yat Mei Lee. 2017. Figurative language in emotion expressions. In *Workshop on Chinese Lexical Semantics*, pages 408–419. Springer.

Saif Mohammad. 2012. #Emotional Tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 246–255, Montréal, Canada.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Saif M. Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

Saif M. Mohammad. 2023. Best practices in the creation and use of emotion lexicons. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia. Association for Computational Linguistics.

Saif M. Mohammad. 2024a. Worrywords: Norms of anxiety association for 44,450 english words. In *Proceedings of The Annual Conference of the Empirical Methods on Natural Language Processing (EMNLP 2024, main)*, Miami, FL.

Saif M. Mohammad. 2024b. WorryWords: Norms of anxiety association for over 44k English words. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16261–16278, Miami, Florida, USA. Association for Computational Linguistics.

Saif M. Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

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

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Agnes Moors, Jan De Houwer, Dirk Hermans, Sabine Wanmaker, Kevin Van Schie, Anne-Laura Van Harmelen, Maarten De Schryver, Jeffrey De Winne, and Marc Brysbaert. 2013. Norms of valence, arousal, dominance, and age of acquisition for 4,300 dutch words. *Behavior research methods*, 45(1):169–177.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Alipio Jorge, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, and 8 others. 2023. AfriSenti: A Twitter sentiment analysis benchmark for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore. Association for Computational Linguistics.

Emiko J Muraki, Summer Abdalla, Marc Brysbaert, and Penny M Pexman. 2023a. Concreteness ratings for 62,000 english multiword expressions. *Behavior research methods*, 55(5):2522—2531.

Emiko J Muraki, Summer Abdalla, Marc Brysbaert, and Penny M Pexman. 2023b. Concreteness ratings for 62,000 english multiword expressions. *Behavior research methods*, 55(5):2522–2531.

Mark Nichter. 2010. Idioms of distress revisited. *Culture, Medicine, and Psychiatry*, 34(2):401–416.

C.E. Osgood, Suci G., and P. Tannenbaum. 1957. *The measurement of meaning*. University of Illinois Press.

Jaime Redondo, Isabel Fraga, Isabel Padrón, and Montserrat Comesaña. 2007. The spanish adaptation of anew (affective norms for english words). *Behavior research methods*, 39(3):600–605.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword

expressions: A pain in the neck for nlp. In *International conference on intelligent text processing and computational linguistics*, pages 1–15. Springer.

Robert W Schrauf and Julia Sanchez. 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.

Eva Smolka and Sabine Schulte im Walde. 2020. *The role of constituents in multiword expressions: An interdisciplinary, cross-lingual perspective (Volume 4)*. Language Science Press.

Jacopo Staiano and Marco Guerini. 2014. Depechemood: a lexicon for emotion analysis from crowd-annotated news. *arXiv preprint arXiv:1405.1605*.

Carlo Strapparava and Alessandro Valitutti. 2004. Wordnet-Affect: An affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*, pages 1083–1086, Lisbon, Portugal.

Masahito Takahashi, Toshifumi Tanabe, Jack Halpern, and Kosho Shudo. 2024. A comprehensive japanese mwe lexicon: Jmwel. In *Recent Advances in Multiword Units in Machine Translation and Translation Technology*, pages 218–242. John Benjamins Publishing Company.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor understanding challenge dataset for llms. *arXiv preprint arXiv:2403.11810*.

Krishnapriya Vishnubhotla and Saif M. Mohammad. 2022. Tweet Emotion Dynamics: Emotion word usage in tweets from US and Canada. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4162–4176, Marseille, France. European Language Resources Association.

Melissa LH Võ, Markus Conrad, Lars Kuchinke, Karolina Urton, Markus J Hofmann, and Arthur M Jacobs. 2009. The berlin affective word list reloaded (bawl-r). *Behavior research methods*, 41(2):534–538.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Joseph P Weir. 2005. Quantifying test-retest reliability using the intraclass correlation coefficient and the sem. *The Journal of Strength & Conditioning Research*, 19(1):231–240.

Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. 2007. Building emotion lexicon from weblog corpora. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 133–136.

Liang-Chih Yu, Jin Wang, K Robert Lai, and Xue-jie Zhang. 2015. Predicting valence-arousal ratings of words using a weighted graph method. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 788–793.

Ladislav Zgusta. 1967. Multiword lexical units. *Word*, 23(1-3):578–587.

A APPENDIX

A.1 AMT Questionnaires for Valence, Arousal, and Dominance

Screenshots of the detailed instructions, sample instance (question), and examples presented to the annotators are shown in Figures 3 through 11. Participants were informed that they may work on as many instances as they wish. The annotation task was approved by our institution's IRB. The purpose of the task and how their annotations will be used was made clear, and consent was obtained.

A.2 Distribution of MWE-VAD

MWE-VAD is made freely available on the project website as a compressed file. Terms of use will require that users not re-distribute the file and not post any form of the lexicon on the web. This is to prevent the resource being included in the data scrape fed to a large language model. See full list of terms of use at the project home page.

A.3 Computational Resources and Carbon Footprint

A nice advantage of using simple lexicon-based approaches is the low carbon footprint and computational resources required. All of the experiments described in the paper were conducted on a regular personal laptop.

B Supplementary Figures

Distributions of the arousal and dominance classes in MWE types are shown in Figures 12 and 13. Metrics for arousal and dominance compositionality are shown in Figures 14 and 15.

Introduction:

1. Attempt these questions only if you are fluent in English.
2. Your responses are confidential.

Task:

Words can be associated with different degrees of positiveness or negativeness. While there is some variation from person to person, there is also a fair amount of consensus. For example, most people will agree that the term:

- *heaven and ecstacy* are often associated with being **very positive**
- *stroll and good show* are often associated with being **moderately positive**
- *tree and okay* are often associated with being **slightly positive**
- *desk and polygon* are often **not associated** with being positive or negative
- *wait and inconvenience* are often associated with being **slightly negative**
- *argumentative and stalled* are often associated with being **moderately negative**
- *death and fail* are often associated with being **very negative**

In this multiple choice task, you will be given common English terms and you have to select the options that best describe the degree of positiveness or negativeness associated with them.

Consider **positiveness** to be a broad category that includes:

- positiveness, pleasure, goodness, happiness, greatness, brilliance, superiority, health, etc.

Consider **negativeness** to be a broad category that includes:

- negativeness, displeasure, badness, unhappiness, insignificance, terribleness, inferiority, sickness, etc.

Give answers that capture what most English speakers would agree.

If you do not know the meaning of a word or are unsure, you can look it up in a dictionary (e.g., the Merriam Webster) or on the internet.

Purpose of the task:

Your responses will be used in a research study to better understand how positiveness and negativeness mainfest in language.

Quality Control:

- Responses that are not in accordance with the instructions will not be paid for.
- Some questions have pre-determined correct answers. If you mark these questions incorrectly, we will often give you immediate feedback in a pop-up box. An occasional misanswer is okay. In addition, for some questions, we record the misanswers, but do not show a popup. If the **rate of misanswering is high (e.g., >20%), then **all** of one's HITs may be rejected.**
- If you see that you are getting quite a few of the gold questions wrong (e.g. more than 2 in every 10 HITs), then do not accept more HITs.
- If you disagree with the answer for a gold HIT, include the correct response in the Feedback textbox. Note that missing an occasional gold question will not lead to the rejection of your responses.
- This quality control measure promotes fairness for those who do the task responsibly.

Notes:

- If a term has more than one meaning, consider the most common meaning.
- A rule of thumb is that a term associated with more positiveness tends to often occur in sentences that convey positiveness, whereas a term associated with more negativeness tends to often occur in sentences that convey negativeness.
- Try not to overthink the answer. Let your instinct guide you.

Figure 3: Valence Questionnaire: Detailed instructions.

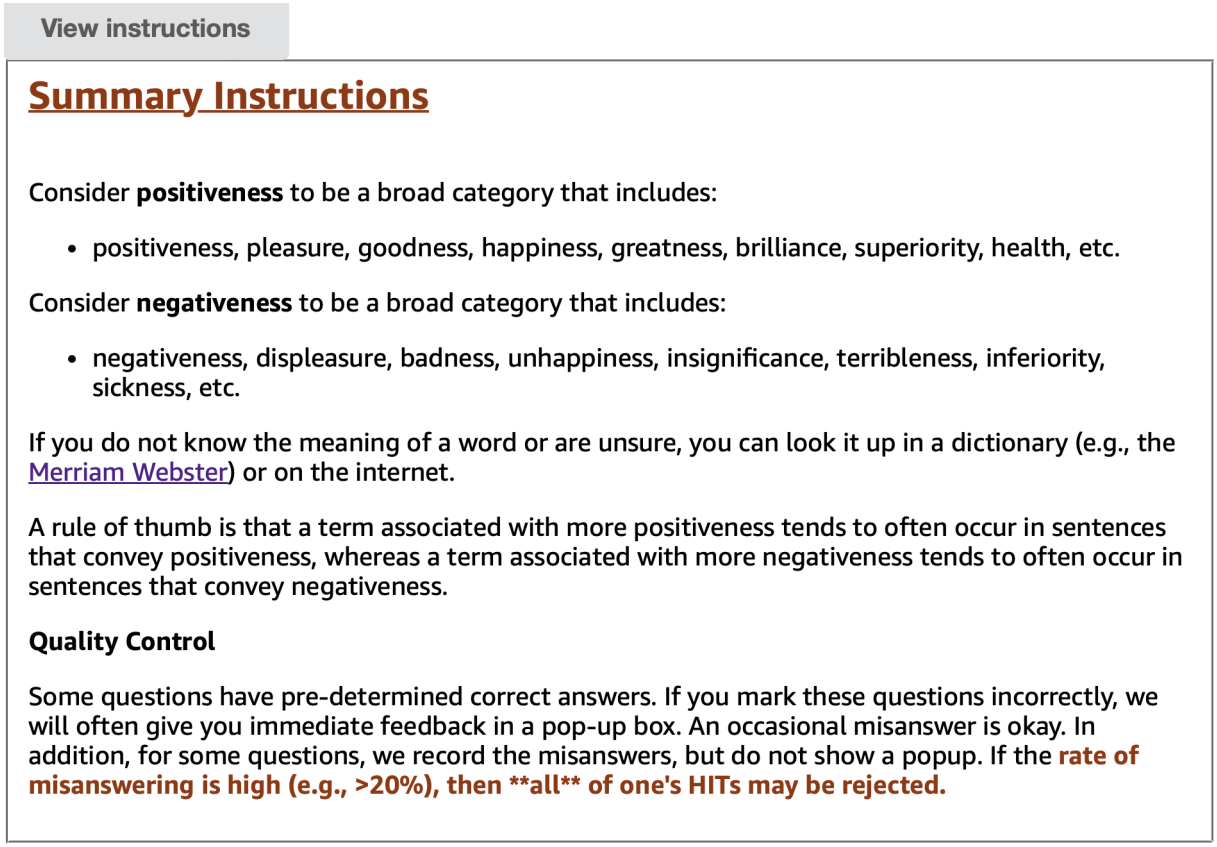


Figure 4: Valence Questionnaire: Sample question.

Very positive:

- heaven, promotion, ecstacy, vacation, success, kindly, courage, etc.

Moderately positive:

- stroll, good show, gift, slept well, favor, smooth sailing, etc.

Slightly positive:

- tree, starter, okay, some help, word play, etc.

Not associated with positiveness or negativeness:

- furniture, envelope, utencil, fyi, garage, profession, very, same, percent, etc.

Slightly negative:

- wait, inconvenience, climbing stairs, confused, lip service, worn, slow day, etc.

Moderately negative:

- argumentative, taxes, warned, stalled, subpar, minor illness, etc.

Very negative:

- death, murder, cancer, tyrant, crime, fail, crying, etc.

Figure 5: Valence Questionnaire: Examples.

Introduction:

1. Attempt these questions only if you are fluent in English.
2. Your responses are confidential.

Task:

Words can be associated with different degrees of activeness or arousal or inactiveness or calmness. While there is some variation from person to person, there is also a fair amount of consensus. For example, most people will agree that the term:

- *war zone and ecstacy* are often associated with being **very active or aroused**
- *prepare and concern* are often associated with being **moderately active or aroused**
- *wondering and meeting* are often associated with being **slightly active or aroused**
- *copper and apple* are often **not associated** with being active or aroused or inactive or calm
- *sunday and routine* are often associated with being **slightly inactive or calm**
- *garden and snug* are often associated with being **moderately inactive or calm**
- *serene and lifeless* are often associated with being **very inactive or calm**

In this multiple choice task, you will be given common English terms and you have to select the options that best describe the degree of activeness or arousal or inactiveness or calmness associated with them.

Consider **activeness or arousal** to be a broad category that includes:

- active, aroused, stimulated, frenzied, excited, jittery, alert, etc.

Consider **inactiveness or calmness** to be a broad category that includes:

- inactive, calm, unaroused, passive, relaxed, sluggish, etc.

This task is not about sentiment. (For example, something can be positive and inactive (such as serene or flower), positive and active (such as exercise and party), negative and active (such as murderer), and negative and inactive (such as negligent).

Give answers that capture what most English speakers would agree.

If you do not know the meaning of a word or are unsure, you can look it up in a dictionary (e.g., the Merriam Webster) or on the internet.

Purpose of the task:

Your responses will be used in a research study to better understand how activeness or arousal and inactiveness or calmness mainfest in language.

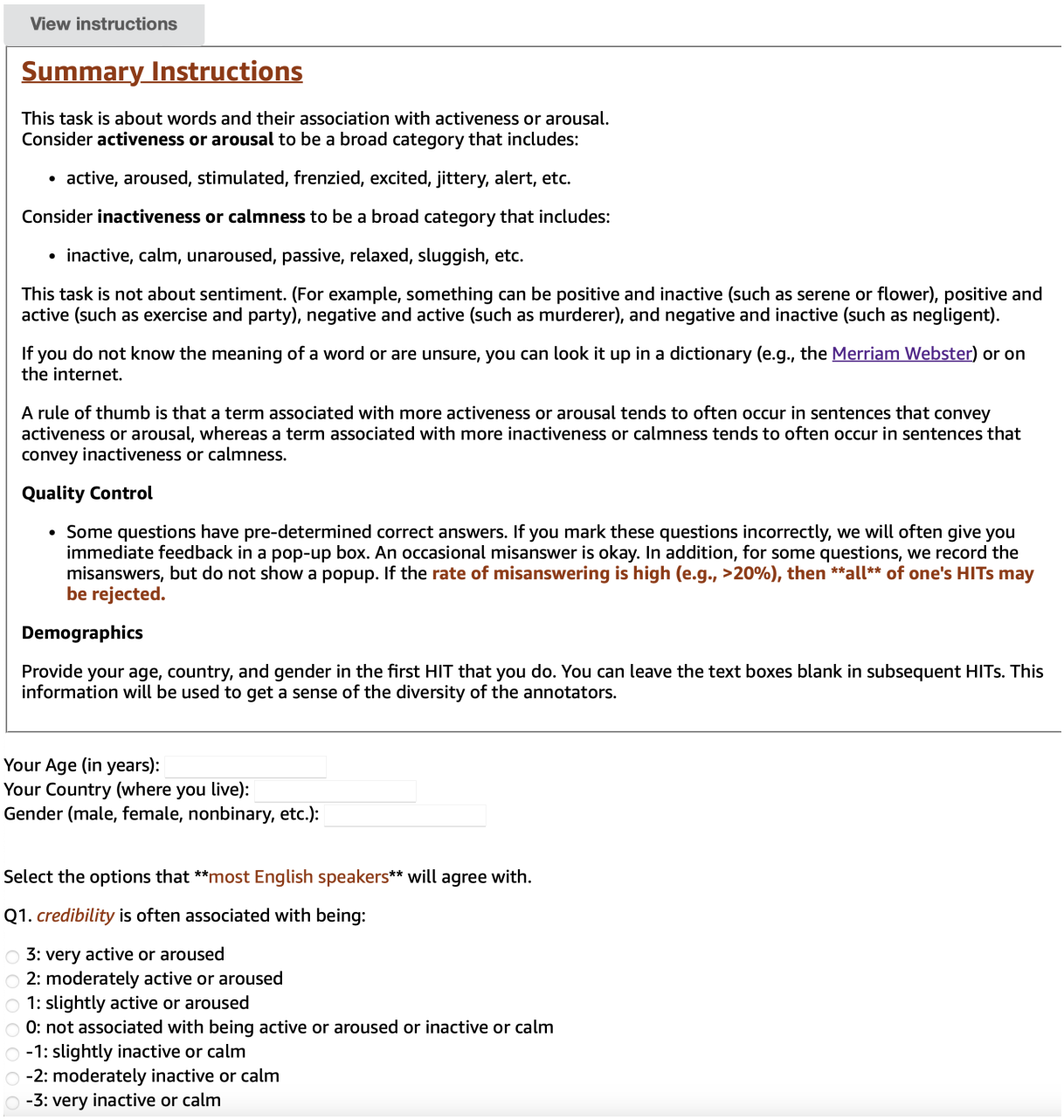
Quality Control:

- Responses that are not in accordance with the instructions will not be paid for.
- Some questions have pre-determined correct answers. If you mark these questions incorrectly, we will often give you immediate feedback in a pop-up box. An occasional misanswer is okay. In addition, for some questions, we record the misanswers, but do not show a popup. If the **rate of misanswering is high (e.g., >20%), then **all** of one's HITs may be rejected.**
- If you see that you are getting quite a few of the gold questions wrong (e.g. more than 2 in every 10 HITs), then do not accept more HITs.
- If you disagree with the answer for a gold HIT, include the correct response in the Feedback textbox. Note that missing an occasional gold question will not lead to the rejection of your responses.
- This quality control measure promotes fairness for those who do the task responsibly.

Notes:

- If a term has more than one meaning, consider the most common meaning.
- A rule of thumb is that a term associated with more activeness or arousal tends to often occur in sentences that convey activeness or arousal , whereas a term associated with more inactiveness or calmness tends to often occur in sentences that convey inactiveness or calmness.
- Try not to overthink the answer. Let your instinct guide you.

Figure 6: Arousal Questionnaire: Detailed instructions.


View instructions

Summary Instructions

This task is about words and their association with activeness or arousal.
Consider **activeness or arousal** to be a broad category that includes:

- active, aroused, stimulated, frenzied, excited, jittery, alert, etc.

Consider **inactiveness or calmness** to be a broad category that includes:

- inactive, calm, unaroused, passive, relaxed, sluggish, etc.

This task is not about sentiment. (For example, something can be positive and inactive (such as serene or flower), positive and active (such as exercise and party), negative and active (such as murderer), and negative and inactive (such as negligent).

If you do not know the meaning of a word or are unsure, you can look it up in a dictionary (e.g., the Merriam Webster) or on the internet.

A rule of thumb is that a term associated with more activeness or arousal tends to often occur in sentences that convey activeness or arousal, whereas a term associated with more inactiveness or calmness tends to often occur in sentences that convey inactiveness or calmness.

Quality Control

- Some questions have pre-determined correct answers. If you mark these questions incorrectly, we will often give you immediate feedback in a pop-up box. An occasional misanswer is okay. In addition, for some questions, we record the misanswers, but do not show a popup. If the **rate of misanswering is high (e.g., >20%), then **all** of one's HITs may be rejected.**

Demographics

Provide your age, country, and gender in the first HIT that you do. You can leave the text boxes blank in subsequent HITs. This information will be used to get a sense of the diversity of the annotators.

Your Age (in years):
Your Country (where you live):
Gender (male, female, nonbinary, etc.):

Select the options that **most English speakers** will agree with.

Q1. *credibility* is often associated with being:

- 3: very active or aroused
- 2: moderately active or aroused
- 1: slightly active or aroused
- 0: not associated with being active or aroused or inactive or calm
- -1: slightly inactive or calm
- -2: moderately inactive or calm
- -3: very inactive or calm

Figure 7: Arousal Questionnaire: Sample question.

Very active or aroused:

- attack, keyed up, rollercoaster, bungee jumping, stimulated, sprint, exam, war zone, etc.

Moderately active or aroused:

- thinking, prepare, concern, compute, waiting, alarm clock, etc.

Slightly active or aroused:

- minor issue, wondering, meeting, etc.

Not associated with activeness or arousal or inactiveness or calmness:

- hat, zebra, apple, body, honesty, copper, etc.

Slightly inactive or calm:

- sunday, routine, staycation, etc.

Moderately inactive or calm:

- garden, snug, unconcerned, happy go lucky, bath tub, etc.

Very inactive or calm:

- lifeless, serene, sluggish, bored, depressed, peaceful, silence, asleep, spa, etc.

Figure 8: Arousal Questionnaire: Examples.

Introduction:

1. Attempt these questions only if you are fluent in English.
2. Your responses are confidential.

Task:

Words can be associated with different degrees of dominance, competence, control of situation, or powerfulness or submissiveness, incompetence, controlled by outside factors, or weakness. While there is some variation from person to person, there is also a fair amount of consensus. For example, most people will agree that the term:

- *trumphant* is often associated with being **very dominant, competent, in control of the situation, or powerful**
- *healthy* is often associated with being **moderately dominant, competent, in control of the situation, or powerful**
- *somewhat useful* is often associated with being **slightly dominant, competent, in control of the situation, or powerful**
- *desk* is often **not associated** with being dominant, competent, in control of the situation, or powerful or submissive, incompetent, not in control of the situation, or weak
- *hazy* is often associated with being **slightly submissive, incompetent, not in control of the situation, or weak**
- *minimum wage* is often associated with being **moderately submissive, incompetent, not in control of the situation, or weak**
- *homeless* is often associated with being **very submissive, incompetent, not in control of the situation, or weak**

In this multiple choice task, you will be given common English terms and you have to select the options that best describe the degree of dominance, competence, control of situation, or powerfulness or submissiveness, incompetence, controlled by outside factors, or weakness associated with them.

Consider **dominance, competence, control of situation, or powerfulness** to be a broad category that includes:

- dominant, competent, in control of the situation, powerful, influential, important, autonomous, etc.

Consider **submissiveness, incompetence, controlled by outside factors, or weakness** to be a broad category that includes:

- submissive, incompetent, not in control of the situation, weak, influenced, cared-for, guided, etc.

This task is not about sentiment. (For example, something can be positive and weak (such as a flower petal) and something can be negative and strong (such as tyrant).

Give answers that capture what most English speakers would agree.

If you do not know the meaning of a word or are unsure, you can look it up in a dictionary (e.g., the Merriam Webster) or on the internet.

Purpose of the task:

Your responses will be used in a research study to better understand how dominance, competence, control of situation, or powerfulness and submissiveness, incompetence, controlled by outside factors, or weakness mainfest in language.

Quality Control:

- Responses that are not in accordance with the instructions will not be paid for.
- Some questions have pre-determined correct answers. If you mark these questions incorrectly, we will give you immediate feedback in a pop-up box. We will keep track of your answers for these gold questions. **If you mark too many of these incorrectly, it will lead to the rejection of **all** your HITs.**
- If you see that you are getting quite a few of the gold questions wrong (e.g. more than 2 in every 10 HITs), then do not accept more HITs.
- If you disagree with the answer for a gold HIT, include the correct response in the Feedback textbox. Note that missing an occasional gold question will not lead to the rejection of your responses.
- This quality control measure promotes fairness for those who do the task responsibly.

Notes:

- If a term has more than one meaning, consider the most common meaning.
- A rule of thumb is that a term associated with more dominance, competence, control of situation, or powerfulness tends to often occur in sentences that convey dominance, competence, control of situation, or powerfulness, whereas a term associated with more submissiveness, incompetence, controlled by outside factors, or weakness tends to often occur in sentences that convey submissiveness, incompetence, controlled by outside factors, or weakness.
- Try not to overthink the answer. Let your instinct guide you.

Figure 9: Dominance Questionnaire: Detailed instructions.

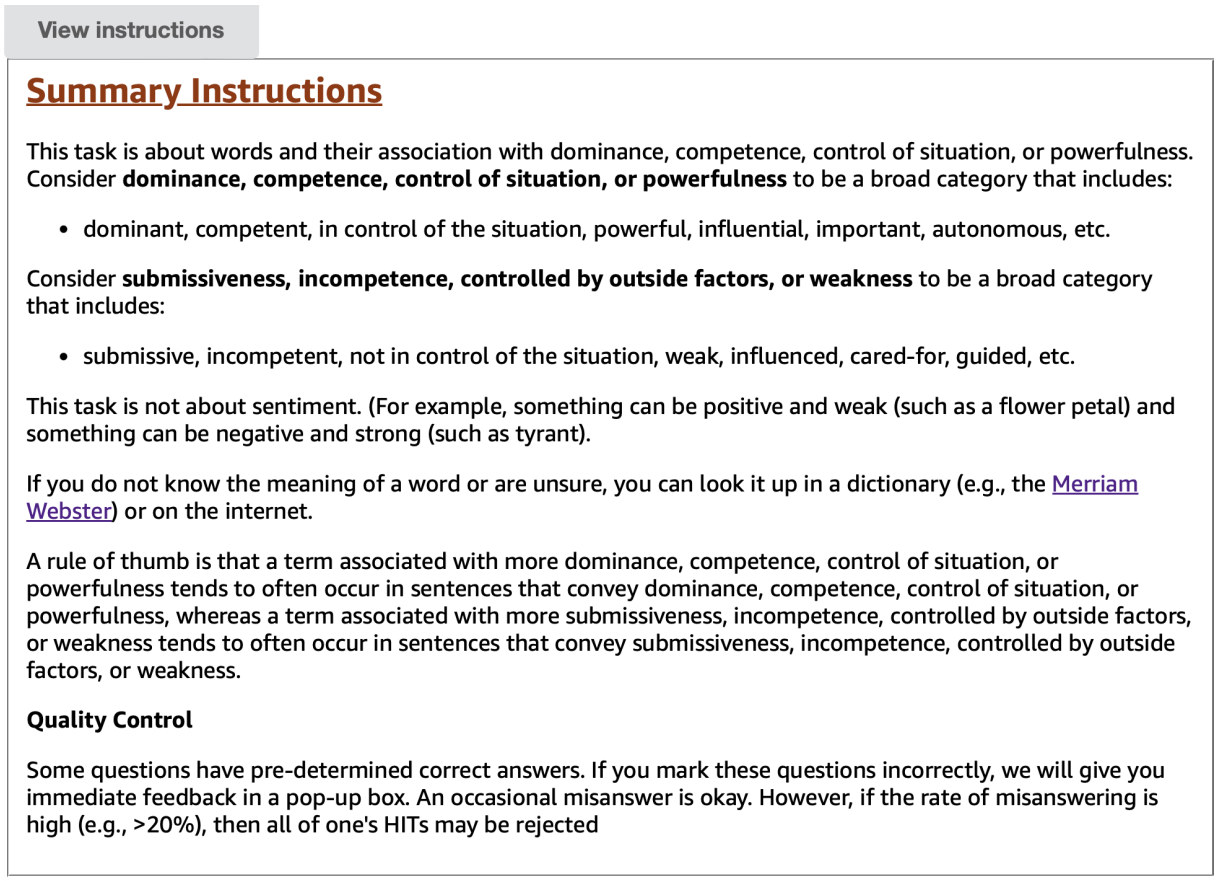


Figure 10: Dominance Questionnaire: Sample question.

Very dominant, competent, in control of the situation, or powerful:

- supreme, trumphant, governor, unflinching, resourceful, giant

Moderately dominant, competent, in control of the situation, or powerful:

- healthy, capable, driving, organize, propel

Slightly dominant, competent, in control of the situation, or powerful:

- keep at it, somewhat useful, increase, illuminate, yoga

Not associated with dominance, competence, control of situation, or powerfulness or submissiveness, incompetence, controlled by outside factors, or weakness:

- desk, hat, zebra, orange, beach, sunny, body, now

Slightly submissive, incompetent, not in control of the situation, or weak:

- lessened, smelly, hazy, frown, stranger

Moderately submissive, incompetent, not in control of the situation, or weak:

- sad, minimum wage, unsure, storm, run out, rolloercoaster

Very submissive, incompetent, not in control of the situation, or weak:

- weakness, pauper, cancer, helpless, lost, slave

Figure 11: Dominance Questionnaire: Examples.

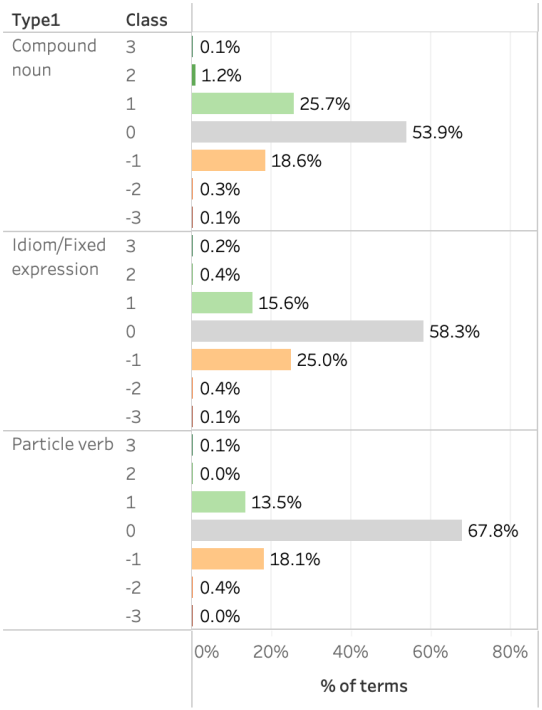


Figure 12: Percentage of MWEs pertaining to each of the **arousal** classes within each MWE type.

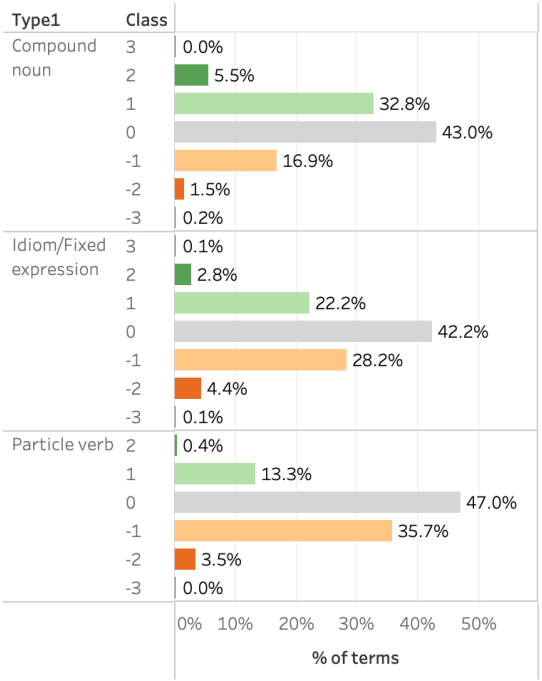


Figure 13: Percentage of MWEs pertaining to each of the **dominance** classes within each MWE type.

a. Average arousal score of bigrams (words1 word2)

Word1-Class	Word2-Class -3	-2	-1	0	1	2	3
-3		-0.40	-0.43	-0.41	-0.26	-0.25	0.09
-2	-0.35	-0.23	-0.12	-0.10	0.03	0.06	0.38
-1	-0.32	-0.12	-0.09	-0.04	0.10	0.10	0.30
0	-0.35	-0.08	-0.04	0.00	0.11	0.14	0.37
1	-0.41	0.04	0.07	0.08	0.20	0.23	0.35
2		0.12	0.13	0.17	0.26	0.29	0.46
3		0.41	0.27	0.30	0.48	0.36	0.47

Avg. arousal score of bigram
-0.75 0.75

b. Percentage of bigrams associated with high arousal (A >= 0.33)

Word1-Class	Word2-Class -3	-2	-1	0	1	2	3
-3		0.000	0.000	0.000	0.000	0.000	0.000
-2	0.000	0.012	0.144	0.120	0.132	0.228	0.060
-1	0.000	0.120	0.853	1.418	0.985	1.118	0.156
0	0.000	0.156	1.238	2.187	1.226	1.214	0.192
1	0.012	0.228	1.574	1.767	0.901	1.094	0.072
2		0.120	0.973	1.574	0.541	0.769	0.132
3		0.048	0.192	0.361	0.084	0.144	0.036

% terms -0.103 2.187

c. Percentage of bigrams associated with low arousal (A <= -0.33)

Word1-Class	Word2-Class -3	-2	-1	0	1	2	3
-3		0.072	0.168	0.168	0.048	0.024	0.000
-2	0.048	0.685	1.550	0.937	0.084	0.120	0.000
-1	0.108	1.286	4.471	2.668	0.144	0.264	0.024
0	0.084	0.781	2.512	2.019	0.084	0.108	0.000
1	0.072	0.108	0.409	0.373	0.024	0.012	0.000
2		0.012	0.120	0.132	0.012	0.000	0.000
3		0.000	0.000	0.000	0.000	0.000	0.000

% terms -0.212 4.471

Figure 14: Measures of Arousal Compositionality.

a. Average dominance score of bigrams (words1 word2)

Word1-Class	Word2-Class -3	-2	-1	0	1	2	3
-3			-0.64	-0.44	-0.51	-0.59	
-2		-0.27	-0.27	-0.25	-0.25	-0.12	-0.03
-1	-0.35	-0.22	-0.13	-0.13	-0.05	-0.03	0.11
0	-0.65	-0.21	-0.08	-0.01	0.14	0.13	0.44
1		-0.26	0.01	0.07	0.22	0.24	0.42
2	-0.32	-0.04	0.09	0.19	0.31	0.37	0.61
3			0.31	0.27	0.48	0.46	0.56

Avg. dominance score of bigram
-0.75 0.75

b. Percentage of bigrams associated with high dominance (D >= 0.33)

Word1-Class	Word2-Class -3	-2	-1	0	1	2	3
-3			0.000	0.000	0.000	0.000	
-2		0.000	0.012	0.048	0.024	0.048	0.000
-1	0.000	0.024	0.468	1.056	0.576	0.576	0.012
0	0.000	0.072	1.128	3.818	3.061	2.185	0.120
1		0.000	1.345	3.565	3.205	2.173	0.120
2	0.000	0.060	0.780	2.269	1.657	1.405	0.108
3			0.096	0.168	0.132	0.132	0.024

% terms -0.181 3.818

c. Percentage of bigrams associated with low dominance (D <= -0.33)

Word1-Class	Word2-Class -3	-2	-1	0	1	2	3
-3			0.036	0.156	0.072	0.048	
-2		0.108	1.104	1.248	0.576	0.192	0.000
-1	0.024	0.468	3.241	4.526	0.840	0.708	0.000
0	0.024	0.468	2.869	4.442	0.612	0.612	0.000
1		0.228	0.984	1.621	0.312	0.168	0.000
2	0.012	0.084	0.216	0.336	0.048	0.012	0.000
3			0.000	0.024	0.000	0.000	0.000

% terms -0.214 4.526

Figure 15: Measures of Dominance Compositionality.