ValNorm Quantifies Semantics to Reveal Consistent Valence Biases Across Languages and Over Centuries

Autumn Toney-Wails
Georgetown University
`autumn.toney@georgetown.edu`

Aylin Caliskan
University of Washington
`aylin@uw.edu`

Abstract

Word embeddings learn implicit biases from linguistic regularities captured by word co-occurrence statistics. By extending methods that quantify human-like biases in word embeddings, we introduce ValNorm, a novel intrinsic evaluation task and method to quantify the valence dimension of affect in human-rated word sets from social psychology. We apply ValNorm on static word embeddings from seven languages (Chinese, English, German, Polish, Portuguese, Spanish, and Turkish) and from historical English text spanning 200 years. ValNorm achieves consistently high accuracy in quantifying the valence of non-discriminatory, non-social group word sets. Specifically, ValNorm achieves a Pearson correlation of $\rho = 0.88$ for human judgment scores of valence for 399 words collected to establish pleasantness norms in English. In contrast, we measure gender stereotypes using the same set of word embeddings and find that social biases vary across languages. Our results indicate that valence associations of non-discriminatory, non-social group words represent widely-shared associations, in seven languages and over 200 years.

1 Introduction

New transparency-enhancing methods for static word embedding evaluation incorporate cross-disciplinary techniques that can quantify widely-accepted, intrinsic characteristics of words (Bakarov, 2018; Faruqui et al., 2016; Schnabel et al., 2015; Hollenstein et al., 2019). A promising approach for developing transparent intrinsic evaluation tasks is to evaluate word embeddings through the lens of *cognitive lexical semantics*, which captures the social and psychological responses of humans to words and language (Hollenstein et al., 2019; Osgood, 1964; Osgood et al., 1975). Such an approach could provide a representativeness evaluation method for word embeddings used in quantifying and studying biases.

This paper presents the computational approach ValNorm, that accurately quantifies the valence dimension of biases and affective meaning in word embeddings, to analyze widely shared associations of non-social group words. Implicit biases, as well as the intrinsic pleasantness or goodness of things, namely valence, have been well researched with human subjects (Greenwald et al., 1998; Russell, 1983; Russell and Mehrabian, 1977). Valence is one of the principal dimensions of affect and cognitive heuristics that shape attitudes and biases in humans (Harmon-Jones et al., 2013). Valence is described as the affective quality referring to the intrinsic attractiveness/goodness or averseness/badness of an event, object, or situation (Frijda et al., 1986; Osgood et al., 1957). For word embeddings, we define valence bias as the semantic evaluation of pleasantness or unpleasantness that is associated with words (e.g., *kindness* is associated with pleasantness and *torture* is associated with unpleasantness).

Word embedding evaluation tasks are methods to measure the quality and accuracy of learned word vector representations from a given text corpus. The two main types of evaluation tasks are *intrinsic evaluation*, which analyzes and interprets the semantic or syntactic characteristics of word embeddings (e.g., word similarity), and *extrinsic evaluation*, which measures how well word embeddings perform on downstream tasks (e.g., part-of-speech tagging, sentiment classification) (Wang et al., 2019; Tsvetkov et al., 2016). We focus on intrinsic evaluation, specifically the semantic quality of word embeddings that have been shown to learn human-like biases, such as gender and racial stereotypes (Caliskan et al., 2017).

Our intrinsic evaluation task, ValNorm, accurately quantifies the valence dimension of biases and affective meaning in word embeddings. Simply, ValNorm provides a statistical measure of the pleasant/unpleasant connotation of a word. We

validate ValNorm's ability to quantify the semantic quality of words by implementing the task on word embeddings from seven different languages (Chinese, English, German, Polish, Portuguese, Spanish, and Turkish) and over a time span of 200 years (English only). Our results showcase that non-discriminatory non-social group biases introduced by Bellezza et al. (1986) are consistent across cultures and over time; people agree that loyalty is pleasant and that hatred is unpleasant. Additionally, we use the Word Embedding Association Test (WEAT) (Caliskan et al., 2017) to measure the difference between social group biases and non-discriminatory biases in seven languages.

We compare ValNorm to six widely used, traditional intrinsic evaluation tasks that measure how semantically similar two words are to each other and how words relate to each other. All intrinsic evaluation tasks, including ValNorm, measure the correlation of the computed scores to human-annotated scores. We implement all intrinsic evaluation tasks on seven word embedding sets (English only), which were trained using four different embedding algorithms and five different training text corpora (see Figure 1) to ensure that the ValNorm results are not model or corpus specific. ValNorm achieves Pearson correlation coefficients (ρ) in the range $[0.82, 0.88]$ for the seven English word embedding sets, outperforming the six traditional intrinsic evaluation tasks we compare our results to.

We summarize our three main contributions: **1)** We quantify semantics, specifically the valence dimension of affect (pleasantness/unpleasantness) to study the valence norms of words and present a permutation test to measure the statistical significance of our valence quantification, **2)** we introduce ValNorm, a new intrinsic evaluation task that measures the semantic quality of word embeddings (validated on seven languages), and **3)** we establish widely-shared associations of valence across languages and over time. Extended methodology, results, dataset details are in the appendices; the open source repository will be made public.

2 Related Work

Derived from the Implicit Association Test (IAT) in social psychology, Caliskan et al. defined the Word Embedding Association Test (WEAT) and the Word Embedding Factual Association Test (the single-category WEAT), to measure implicit biases in word embeddings (Caliskan et al., 2017). The WEAT has two tests that measure non-social group (e.g., flowers) biases and seven tests that measure social group (e.g., gender, race) biases. The social group WEATs have been widely studied in the natural language processing (NLP) domain, as understanding social group biases is important for society. The single-category WEAT (SC-WEAT) measured gender bias in occupations and androgynous names, which highly correlate with gender statistics (Caliskan et al., 2017).

SC-WEAT resembles a single-category association test in human cognition (Caliskan and Lewis, 2020; Guo and Caliskan, 2021; Karpinski and Steinman, 2006). SC-WEAT also shares similar properties with lexicon induction methods, which automatically extract semantic dictionaries from textual corpora without relying on large-scale annotated data for training machine learning models (Hatzivassiloglou and McKeown, 1997). Riloff and Wiebe (2003); Turney and Littman (2003) apply lexicon induction methods for sentiment, polarity, orientation, and subjectivity classification.

In prior work, the classification of valence is not evaluated in the context of measuring the quality of word embeddings or quantifying valence norms.

Lewis and Lupyan (2020) investigate the distributional structure of natural language semantics in 25 different languages to determine the gender bias in each culture. While Lewis and Lupyan analyze bias across languages, they focus specifically on the social group of gender, and not on widely shared associations across languages. Garg et al. quantify gender and ethnic bias over 100 years to dynamically measure how biases evolve over time (Garg et al., 2018). Similarly, Garg et al. do not measure widely shared associations over time, they only measure social group biases.

Predicting affective ratings of words from word embeddings has proven to be a more complex task than computing word similarity, and is typically approached as a supervised machine learning problem (Li et al., 2017; Teofili and Chhaya, 2019; Wang et al., 2016). Affect ratings of words computed from word embeddings can improve NLP tasks involving sentiment analysis and emotion detection (Ungar et al., 2017; Mohammad, 2016), thus, designing an intrinsic evaluation task that estimates the valence association of a word is significant.

Traditional word embedding intrinsic evaluation tasks use word similarity, word analogy, or word categorization to measure linguistic properties cap-

tured by word embeddings (Schnabel et al., 2015; Bakarov, 2018). Word similarity and word analogy tasks use cosine similarity to measure semantic similarity of the vector representations of words in the evaluation task. Word similarity tasks compare the cosine similarity to a human-rated similarity score through Pearson or Spearman correlation (Kirch, 2008; Dodge, 2008); the correlation coefficient provides the accuracy metric for semantic similarity learned by word embeddings. Word analogy tasks output matching words based on vector arithmetic and accuracy is the metric of correct word selection (Mikolov et al., 2013). Since there is no standardized approach to evaluate word embeddings, we focus on the five most commonly used word similarity tasks WordSim (353 word pairs), RareWord (2,034 word pairs), MEN (3,000 word pairs), SimLex (999 word pairs), SimVerb (3,500 word pairs) (Finkelstein et al., 2001; Luong et al., 2013; Bruni et al., 2014; Hill et al., 2015; Gerz et al., 2016; Turney and Pantel, 2010), and the word analogy task from Mikolov et al. (2013) which contains 8,869 semantic and 10,675 syntactic questions.

3 Datasets

We use two main sources of data: 1) word embeddings and 2) human-annotated validation datasets.

Word Embeddings: We choose six widely-used, pre-trained word embedding sets in English, listed in Table 1, to compare ValNorm's performance on different algorithms (GloVe, fastText, word2vec) and training corpora (Common Crawl, Wikipedia, OpenSubtitles, Twitter, and Google News) (Pennington et al., 2014; Bojanowski et al., 2016; Mikolov et al., 2013; Grave et al., 2018). We include a seventh word embedding set, ConceptNet Numberbatch, since it is comprised of an ensemble of lexical data sources and is claimed to be less prejudiced in terms of ethnicity, religion, and gender (Speer et al., 2017). ConceptNet Numberbatch's results on social group and non-social group association tests provide a unique insight into valence norms for word embeddings, since the social group biases have been intentionally lowered.

We use the 300-dimensional, pre-trained fastText word embeddings prepared for 157 languages for our seven languages of interest from five branches of language families that have different syntactic properties (Chinese, English, German, Polish, Portuguese, Spanish, and Turkish) (Grave et al., 2018).

Algorithm	Corpus	# Tokens
fastText skipgram	Common Crawl	600B
	OpenSubtitles 2018	22.2B
fastText CBOW	Common Crawl	
	Wikipedia	600+ B
GloVe	Common Crawl	840B
	Twitter	27B
Numberbatch	-	-
word2vec	Google News	100B

300-dimensional embeddings (except the 200-dimensional GloVe Twitter)

Table 1: Details of word embeddings used in experiments; the corpus and token count for Numberbatch word embeddings are blank since they are formed using lexical information from ConceptNet, OpenSubtitles 2016, GloVe, and word2vec (Speer et al., 2017).

For longitudinal valence analysis, we use historical word embeddings from Hamilton et al. (2016) trained on English text between 1800 and 1990. Each word embedding set covers a 10-year period.

Validation Datasets (English): We choose three validation, human-annotated datasets of varying size for our experiments in English. All human-rated valence scores are reported as the mean.

Bellezza et al. (1986) compiled a vocabulary list of 399 words to establish norms for pleasantness, imagery, and familiarity. College students rated words on pleasantness versus unpleasantness, which corresponds to cognitive representation of valence. The Affective Norms for English Words (ANEW) dataset is a widely used resource in sentiment analysis for NLP tasks. ANEW contains 1,034 vocabulary words and their corresponding valence, arousal, and dominance ratings. Psychology students were asked to rate words according to the Self-Assessment Manikin (SAM), on a scale of 1 (unhappy) to 9 (happy) (Bradley and Lang, 1999). Warriner et al. (2013) extended ANEW to 13,915 vocabulary words by adding words from more category norms (e.g., taboo words, occupations, and types of diseases). 1,827 Amazon Mechanical Turk workers rated words on the SAM scale of 1 to 9. Warriner et al. (2013) note that valence scores were comparatively similar among responses.

The three human-annotated datasets have 381 common words in their respective vocabularies[1]. Using this subset of 381 words, we measure the

---

[1]The missing words are 'affectionate', 'anxiety', 'capacity', 'comparison', 'constipation', 'disappointment', 'easter', 'epilepsy', 'hitler', 'inconsiderate', 'magnificent', 'me', 'nazi', 'prosperity', 'reformatory', 'sentimental', 'tuberculosis', 'woman'.

Pearson correlation (ρ) of human valence scores across all three datasets to assess the inter-rater reliability. Our measurements result in $\rho \geq 0.97$ for all combinations of comparison. This high correlation indicates a strong inter-rater reliability for valence scores of words and signals widely shared associations, since each dataset was collected from a different year (1995, 1999, and 2013) with different groups of participants from various backgrounds.

Validation Datasets (Cross-linguistic):

ANEW has been adapted to many languages in order to interpret affective norms across cultures. We select five adaptations of ANEW: German, Polish, Portuguese, Spanish, and Turkish. We found these sets to be most complete (included majority of the ANEW vocabulary) and representative of various language structures (e.g., Turkish is a non-gendered language). We also include an affective norm Chinese dataset that contains a large overlapping vocabulary, but is not an ANEW adaptation.

The Polish, Portuguese, Spanish, and Turkish adaptations of ANEW use the original set of words, translated by experts to ensure accurate cross-linguistic results, and collected human-rated valence scores on the SAM 1 to 9 point scale for unhappy/happy according to the original ANEW study (Imbir, 2015; Soares et al., 2012; Redondo et al., 2007; Kapucu et al., 2018). The German adaptation of ANEW is an extension of the Berlin Affected Word List (BAWL)(Vo et al., 2009) and was relabeled as Affective Norms for German Sentiment Terms (ANGST) (Schmidtke et al., 2014). Valence scores were collected on a -3 (unhappy) to 3 (happy) point scale (Schmidtke et al., 2014). The Chinese Valence-Arousal Words (CVAW) contains 5,512 words rated by four expert annotators who were trained on the Circumplex Model of Affect, which is one the foundational methodologies for affective meaning of words (Russell and Mehrabian, 1977; Yu et al., 2016). The annotators assigned sentiment scores on a 1 (negative) to 9 (positive) point scale accordingly (Yu et al., 2016).

	EN	DE	PL	PT	ES	TR
CN	0.91	0.85	0.88	0.88	0.89	0.83
EN	1.0	0.95	0.94	0.95	0.95	0.87
DE		1.0	0.92	0.94	0.95	0.87
PL			1.0	0.93	0.93	0.85
PT				1.0	0.96	0.88
ES					1.0	0.88

Table 2: Pearson correlation coefficient (ρ) of human scores in the validation sets across languages.

We identify the words in common across the seven cross-linguistic datasets, and we check the variance in human-annotated valence scores for this subset of 143 words[2]. The top five words with the **least** amount of variance in valence are *terrific* ($\sigma^2 = 3.6 \times 10^{-4}$), *loyal* ($\sigma^2 = 4.7 \times 10^{-4}$), *humor* ($\sigma^2 = 6.9 \times 10^{-4}$), *hatred* ($\sigma^2 = 7.4 \times 10^{-4}$), and *depression* ($\sigma^2 = 9.2 \times 10^{-4}$). The top five words with the **most** amount of variance in valence are *execution* ($\sigma^2 = 4.9 \times 10^{-2}$), *party* ($\sigma^2 = 4.1 \times 10^{-2}$), *vomit* ($\sigma^2 = 3.4 \times 10^{-2}$), *malaria* ($\sigma^2 = 2.7 \times 10^{-2}$), and *torture* ($\sigma^2 = 2.6 \times 10^{-2}$). The overall variance of valence for all words are low.

4 Methods

To measure social group biases and valence norms, we use the Word Embedding Association Test (WEAT) and the single-category Word Embedding Association Test (SC-WEAT).

4.1 Association Tests: WEAT & SC-WEAT

The WEAT and SC-WEAT compute an effect-size statistic (Cohen's d) (Cohen, 2013) measuring the association of a given set of target words or a single vocabulary word between two given attribute sets in a semantic vector space composed of word embeddings. The WEAT measures the differential association between two sets of target words and two sets of polar attribute sets, and the SC-WEAT measures the association of a single word to the two sets of polar attributes. Stimuli representing target social groups and polar attributes used in the WEAT are borrowed from the IATs designed by experts in social psychology. Table 3 provides the equations to compute the effect sizes for WEAT and SC-WEAT and their respective p-values; the p-values represent the significance of the effect sizes. $|d| \geq 0.80$ represents a biased association with high effect size (Cohen, 2013), with a one sided p-value ≤ 0.05 or p-value ≥ 0.95 representing a statistically significant effect size.

We use the WEAT and SC-WEAT to quantify biases and measure statistical regularities in text corpora. We extend the SC-WEAT to precisely measure valence using pleasant/unpleasant evaluative attribute sets provided by Greenwald et al.

---

[2]This word set is mainly limited by the Chinese dataset. See Table 5 for the number of words contained in each language's dataset.

Test	Effect Size (es measured in Cohen's d)	p-value
WEAT	$es(X, Y, A, B) = \frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std-dev}_{w \in X \cup Y} s(w, A, B)}$	$\Pr_i[s(X_i, Y_i, A, B) > s(X, Y, A, B)]$
SC-WEAT	$es(\vec{w}, A, B) = \frac{\text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})}{\text{std-dev}_{x \in A \cup B} \cos(\vec{w}, \vec{x})}$	$\Pr_i[s(\vec{w}, A_i, B_i) > s(\vec{w}, A, B)]$

Association: $s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})$

Cosine similarity: $\cos(\vec{a}, \vec{b})$ denotes the cosine of the angle between the vectors $\vec{a}$ and $\vec{b}$.

Target words: $X = [\vec{x_1}, \vec{x_2}, \ldots, \vec{x_m}]$ and $Y = [\vec{y_1}, \vec{y_2}, \ldots, \vec{y_m}]$ are the two equal-sized (m) sets of target stimuli.

Attribute words: $A = [\vec{a_1}, \vec{a_2}, \ldots, \vec{a_n}]$ and $B = [\vec{b_1}, \vec{b_2}, \ldots, \vec{b_n}]$ are the two equal-sized (n) sets of attributes.

SC-WEAT: $\vec{w}$ is the single target stimulus in SC-WEAT.

Permutation test: $(X_i, Y_i)_i$ and $(A_i, B_i)_i$ denote all the partitions of $X \cup Y$ and $A \cup B$ into two sets of equal size. Random permutations of these sets represent the null hypothesis as if the biased associations did not exist so that we can perform a statistical significance test by measuring the unlikelihood of the null hypothesis, given the effect size of WEAT or SC-WEAT.

Table 3: WEAT and SC-WEAT effect size equations and their corresponding statistical significance in p-values.

(1998) (as opposed to male/female from Caliskan et al. (2017)). We design our intrinsic evaluation task, ValNorm, around this valence quantification method. We extend all methods (WEAT, SC-WEAT, ValNorm) to six non-English languages using native speaker translations of the word sets. In all of our experiments we use the defined sets of stimuli from (Caliskan et al., 2017) to ensure that our experiments provide accurate results. Following WEAT, each word set contains at least 8 words to satisfy concept representation significance. Accordingly, the limitations of not following WEAT's methodological robustness rules, which are analyzed by Ethayarajh et al. (2019), are mitigated.

4.2 Statistical Significance of Valence Quantification

Caliskan et al. do not present a p-value for the SC-WEAT effect size[3]. Thus, we define the one-sided p-value of SC-WEAT where $\{(A_i, B_i)\}_i$ represents the set of all possible partitions of the attributes $A \cup B$ of equal size to represent the null hypothesis. The null hypothesis is, for a given stimulus $\vec{w}$, computing the SC-WEAT effect size using a random partition of the attribute words $\{(A_i, B_i)\}_i$ represents the empirical distribution of effect sizes in case there were no biased associations between the stimulus and the attribute sets. Accordingly, the permutation test measures the unlikelihood of the null hypothesis for SC-WEAT.

4.3 ValNorm: An Intrinsic Evaluation Task

Our intrinsic evaluation task uses the SC-WEAT with pleasant and unpleasant attribute sets to represent the valence dimension of affect[4]. ValNorm's output is the Pearson's correlation value when comparing the computed valence scores to the human-rated valence scores from a ground truth validation dataset. We define the ValNorm task as:

1. **Assign** the *word* column from the validation dataset to W, the set of target word vectors.

2. **Assign** the pleasant attribute words to A, the first attribute set, and assign the unpleasant attribute words to B the second attribute set.

3. **Compute** SC-WEAT effect size and p-value for each $\vec{w} \in W$ using the given word embedding set.

4. **Compare** SC-WEAT effect sizes to the human-rated valence scores using Pearson's correlation to measure the semantic quality of word embeddings.

For non-English languages we include a preliminary step to ValNorm, where we translate the pleasant/unpleasant attribute word sets from English to the given language and verify the translations with native speakers.

4.4 Discovering Widely-Accepted Non-Social Group Associations

We investigate the existence of widely-accepted non-social group biases by implementing the *flowers-insects-attitude*, *instruments-weapons-attitude*, and *gender-science* WEATs defined by Caliskan et al. (2017) on word embeddings from

---

[3]Caliskan et al. measure the p-value of the correlation between the SC-WEAT computed gender association scores and their corresponding ground truth values obtained from annual U.S. Census and Labour Bureau statistics.

[4]https://github.com/autumntoney/ValNorm

our seven languages of interest[5]. Both non-social group attitude tests are introduced as 'universally accepted stereotypes' in the original paper that presents the IAT (Greenwald et al., 1998). Thus, these are baseline biases that we expect to observe with high effect size in any representative word embeddings. We use the *gender-science* WEAT results, which measures social group biases, to compare with our non-social group bias tests' results. In this way, we can identify if social and non-social group biases have consistent results across languages to infer if our non-social group bias results indicate universality.

Additionally, we implement ValNorm on the six non-English word embedding sets and historical word embeddings from 1800–1990.

5 Experiments

We conduct three main experiments to 1) quantify the valence statistics of words in text corpora, 2) evaluate our intrinsic evaluation task, ValNorm and 3) investigate widely-shared non-social group valence associations across languages and over time.

5.1 Quantifying Valence Using SC-WEAT

We use the SC-WEAT to quantify valence norms of words by measuring a single word's relative association to pleasant versus unpleasant attribute sets. We use the same word sets of 25 pleasant and 25 unpleasant words used in Caliskan et al. (2017) flowers-insects-attitude bias test. These attribute word sets were designated by experts in social psychology to have consistent valence scores among humans (Greenwald et al., 1998). We run the SC-WEAT on the seven sets of word embeddings listed in Section 3, and we evaluate each word embedding set using valence lexica.

5.2 Evaluating ValNorm

We run ValNorm on the seven English word embedding sets, using Bellezza's Lexicon, ANEW, and Warriner's Lexicon as the target word set respectively. We measure the correlation of the ValNorm scores to the corresponding set of human-rated scores. We compare ValNorm's results to the results from six traditional intrinsic evaluation

tasks on the seven English word embedding sets. This evaluation compares six traditional evaluation tasks to three implementations of ValNorm across seven sets of word embeddings, trained using four different algorithms and five different text corpora.

To investigate the significance of training corpus size for word embeddings, we sample 5 bin sizes (50%, 10%, 1%, 0.1%, and 0.001%) of the OpenSubtitles 2018 corpus and train word embeddings according to Paridon and Thompson (2019)'s method to generate subs2vec (fastText skipgram 300-dimensional word embeddings). We choose the OpenSubtitles corpus for this experiment since it reflects human communication behavior more closely than a structured written corpus, such as Wikipedia or news articles, making it a more appropriate corpus for capturing semantic content (Paridon and Thompson, 2019).

There are 89,135,344 lines in the cleaned and deduplicated OpenSubtitles corpus text file, which we round to 89,000,000 to make our sample size bins neat. For each bin size we randomly sample, without replacement, the designated number of lines in the text corpus file. We generate word embeddings for each sample size and run the five word similarity intrinsic evaluation tasks and the ValNorm evaluation task to analyze the significance of corpus size on word embedding quality.

5.3 Analyzing Widely Shared Associations

We use the WEAT to quantify valence associations of non-social groups (flowers, insects, instruments, and weapons) and to quantify social group (male/female) associations to science and arts. We hypothesize that valence biases will remain consistent across word embeddings, and that social group biases will change. Gender bias scores in word embeddings may vary depending on culture and language structure (e.g., Turkish pronouns are gender-neutral). We compare the result differences from the valence association tests and the gender association test on seven different sets of English word embeddings (see Table 1) and on word embeddings from six other languages (Chinese, German, Polish, Portuguese, Spanish, and Turkish). We were unable to run these WEATs on the historical word embeddings, as their vocabularies did not contain most of the target and attribute words.

We implement ValNorm across the six non-English languages, using Bellezza's Lexicon as the target set, since all languages (except for Chinese)

---

[5] Cross-linguistic 'flowers-insects', 'instruments-weapons', and 'gender-science' WEATs have been replicated using their corresponding attribute word sets from IATs on Project Implicit's (Nosek et al., 2002, 2009) webpages in the seven languages we analyzed (Chinese, English, German, Polish, Portuguese, Spanish, and Turkish).
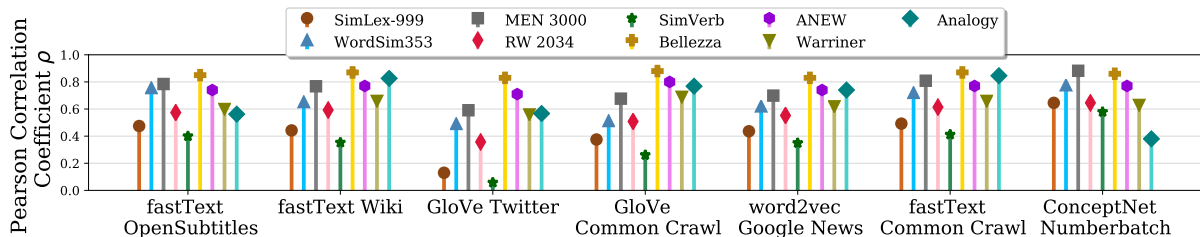
Figure 1: Comparison of nine intrinsic evaluation tasks on seven widely used word embeddings shows that ValNorm achieves the highest correlation with human-rated scores, outperforming other intrinsic evaluation metrics.

had at least 97% of the words in their ground-truth dataset (see Table 5). We also evaluate the stability of valence norms over 200 years by implementing ValNorm on historical embeddings. If valence norms are independent of time, culture, and language, they will be consistent over 200 years and across languages, making them an appropriate metric for evaluating word embeddings.

6 Results

Quantifying valence norms. We implement the SC-WEAT using valence evaluative attributes and target word sets, that are hypothesized to represent valence norms, from Bellezza's Lexicon, ANEW, and Warriner's Lexicon. Our initial experiments signalled widely shared associations of valence scores with $\rho \in [0.82, 0.88]$ for all seven English word embeddings using Bellezza's Lexicon (vocabulary size of 399[6]) as the target word set. The corresponding p-values have a Spearman's correlation coefficient greater than $\rho \geq 0.99$ to the effect sizes, indicating statistically significant results.

ValNorm performance. Figure 1 compares the performance of ValNorm using three valence lexica to five word similarity tasks and one analogy task. ValNorm using Bellezza's Lexicon overperforms all other intrinsic evaluation tasks on word embeddings trained on five corpora via four algorithms.

Widely shared associations. We compute the variance (σ^2) of the effect sizes for the flowers-insects-attitude, instruments-weapons-attitude, and gender-science WEAT bias tests across all seven language word embeddings. In Table 4, as expected based on findings in social psychology, flowers-insects-attitudes and instruments-weapons have the most consistent valence associations, with 0.13 and 0.09 variance scores respectively.

Bias Type	σ^2
flowers-insects	0.13
instruments-weapons	0.09
gender-science	0.45

Table 4: Variance of WEAT scores across 7 languages. The variances of the scores for widely accepted biases are low, whereas culture-specific social group bias scores measuring gender-science associations have high variance across languages.

validation dataset for all seven languages, providing insight into consistent valence norms across cultures. Figure 2 shows the stability of valence norms over 200 years, with low variance in scores ($\sigma^2 < 10^{-3}$), reporting the Pearson correlation coefficients for the valence association scores compared to the corresponding human-rated valence scores from Bellezza's Lexicon (compiled in 1986). Each point on the graph is labeled with the number of vocabulary words from Bellezza's lexicon that was present in the embedding's vocabulary; slight fluctuations in correlation scores may be dependent on the changes in words that were tested.

Figure 3 presents the results for the training corpus size experiment. For all intrinsic evaluation tasks the correlation score increases minimally from 50% to 100% and from 10% to 50%; ValNorm using Bellezza's Lexicon has a 0.01 and 0.03 increase respectively.

7 Discussion

In our three experiments we find evidence that word embeddings capture valence norms using ValNorm and WEAT to measure widely shared associations. These experiments show that valence norms relate to widely-shared associations, as opposed to culture specific associations, and can be used as a measurement of embedding quality across languages.

---

[6]The dataset section includes the details for words that are not included in cross-linguistic experiments.
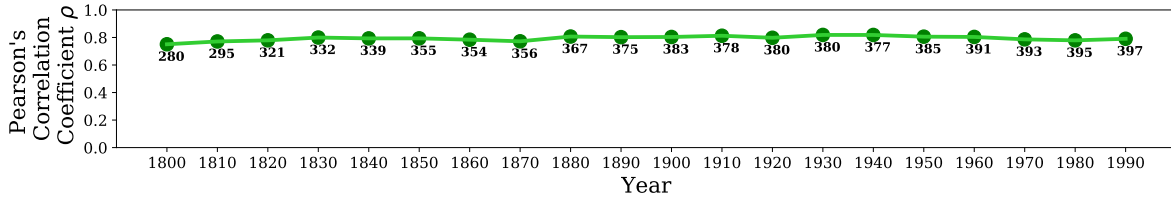
Figure 2: ρ for ValNorm using Bellezza's Lexicon and HistWords historical word embeddings (Hamilton et al., 2016). Points are labeled with the number of target words present in their vocabulary. The low variance ($\sigma^2 < 10^{-3}$) of results validate our hypothesis that valence norms are consistent over two centuries.
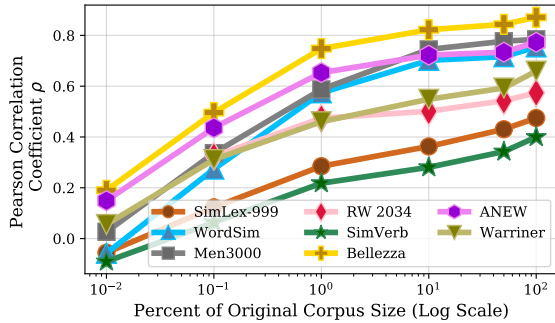

Figure 3: Performance of word embeddings trained on the OpenSubtitles2018 corpus of different sizes.

Language	N	ρ
Chinese	269	0.85
English	381	0.87
EU Portuguese	381	0.85
German	370	0.80
Polish	375	0.79
Spanish	381	0.83
Turkish	379	0.73

Table 5: Correlation coefficients (ρ) and number of target words present (N) in ValNorm for each language.

ValNorm as a new intrinsic evaluation task. Figure 1 compares our three implementations of ValNorm to six traditional intrinsic evaluation tasks, with Bellezza's Lexicon performing the highest, most likely because it is designed specifically to measure valence norms and it is smaller than the other valence lexica. ValNorm computes an effect size rather than just the cosine similarity, the metric for word similarity and word analogy tasks. Notably, ValNorm using Bellezza's Lexicon (399 valence tasks) outperforms WordSim (353 similarity tasks). Using ANEW (1,035 valence tasks) and Warriner's Lexicon (13,915 valence tasks), ValNorm consistently outperforms SimLex (999 similarity tasks), RW (2,034 similarity tasks), and SimVerb (3,5000 similarity tasks). These results suggest that ValNorm measures va-

lence accurately and consistently, regardless of the task size, whereas results of all other intrinsic evaluation tasks have high variance and lower accuracy. ValNorm's performance supports our hypothesis that valence norms are captured by word co-occurrence statistics and that we can precisely quantify valence in word embeddings.

Widely Shared Valence Associations. Measuring ValNorm using Bellezza's Lexicon on Conceptnet Numberbatch word embeddings achieves $\rho = 0.86$. This high ρ value highlights that, even when social group biases are reduced in word embeddings, valence norms remain and are independent of social group biases. The low variance of flowers-insects-attitude and instruments-weapons-attitude experiments signal widely-accepted associations for the non-social groups of flowers, insects, instruments, and weapons. Producing the highest variance of 0.45 across all languages, the gender-science experiment signals a culture and language specific association for gender social groups.

Our corpus size experiment results using ValNorm follow the same trend-line as the other intrinsic evaluation tasks. This result signals widely-shared associations, since the co-occurrence statistics of the word embeddings preserve valence and word similarity scores comparably. When quantifying bias in embeddings, ValNorm can identify if the training corpus is of a sufficient size for representative and statistically significant bias analysis.

Implementing ValNorm on seven different languages from five different language families, we find that valence norms are widely-shared across cultures. However, social group WEAT associations are not widely-shared; these results align with IAT findings from 34 countries (Nosek et al., 2009). Applying WEAT in seven languages, that belong to five branches of varying language families, shows that word embeddings capture grammatical gender along with gender bias. For example, when applying the gender-science WEAT in Polish by using

the IAT words on Poland's Project Implicit, the resulting effect size signals stereotype-incongruent associations. Further analysis of this anomaly revealed that most of the words representing science in the Polish IAT have nouns with feminine grammatical gender. However, when the grammatical gender direction is isolated and removed from the word embeddings while performing WEAT, the results move to the stereotype-congruent direction reported via IATs on the Project Implicit site (Nosek et al., 2009). These findings suggest that structural properties of languages should be taken into account when performing bias measurements that might be somehow related to some syntactic property in a language. This analysis is left to future work since it does not directly affect valence norm measurements in language.

ValNorm quantifies stable valence norms over time with $\rho \in [0.75, 0.82]$ using historical word embeddings and Bellezza's Lexicon. While semantics are certainly evolving (Hamilton et al., 2016), there are non-social group words that maintain their intrinsic characteristics at least for 200 years, as Bellezza et al. suggested, and furthermore, these words are consistent across languages.

8 Conclusion

Valence norms reflect widely-shared associations across languages and time, offering a distinction between non-social group biases and social group biases (gender, race, etc.). These valence associations are captured in word embeddings trained on historical text corpora and from various languages. We document widely-shared non-social group associations as well as culture-specific associations via word embeddings. While the social group biases we measure vary, we find that non-social group valence norms are widely-shared across languages and cultures and stable over 200 years.

We present ValNorm as a new intrinsic evaluation task which measures the quality of word embeddings by quantifying the preservation of valence norms in a word embedding set. ValNorm, which has three implementations with increasing vocabulary sizes, outperforms traditional intrinsic evaluation tasks and provides a more sensitive evaluation metric based on effect size, as opposed to the cosine similarity metric of other evaluation tasks. Computationally quantifying valence of words produce a high correlation to human-rated valence scores, indicating that word embeddings can measure semantics, particularly valence, with high accuracy.

The results of valence norms as statistical regularities in text corpora provides another layer of transparency into what word embeddings are learning during their training process.

9 Ethical Considerations

This work uses expert research in social psychology and computer and information science, specifically the Implicit Association Test (IAT) and the Word Embedding Association Test (WEAT), and applies it to the NLP domain in order to discover widely shared associations of non-discriminatory non-social group words (Greenwald et al., 1998; Caliskan et al., 2017). Prior NLP applications of the WEAT focus mainly on social group biases, since studying potentially harmful features of machine learning and artificial intelligence (AI) are important for fair and ethical implementations of AI. Our application investigates valence (pleasant/unpleasant) associations that quantify attitudes, which can be used to analyze sentiment classification or for a more specific use case of detecting targeted language (information operations/hate speech). By establishing a method to measure valence norms, we establish an AI tool that can identify if biases in a text corpus align with widely accepted valence associations or if the language in the corpus expresses shifted biases.

While our work does not focus on social group biases and attitudes, valence association can be used as an indicator of how a social group is represented in text—is the group associated with pleasantness or unpleasantness? Social group biases are not consistent across cultures and over time, making this valence bias test useful in detecting derogatory or targeted attitudes towards social groups. It is also notable that we share valence associations regardless of language and culture; everyone agrees that kindness is pleasant and that vomit is unpleasant. This distinction between discriminatory biases (against social groups) and non-discriminatory biases (against non-social groups) creates a distinction in analyzing biases and stereotypes in languages. It may be acceptable if language expresses dislike of cancer, but harmful information may propagate to downstream applications if language expresses a negative attitude towards a specific race, gender, or any social group.

10 Acknowledgements

References

Amir Bakarov. 2018. A survey of word embeddings evaluation methods. *arXiv preprint arXiv:1801.09536* (2018).

Francis S Bellezza, Anthony G Greenwald, and Mahzarin R Banaji. 1986. Words high and low in pleasantness as rated by male and female college students. *Behavior Research Methods, Instruments, & Computers* 18, 3 (1986), 299–303.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (07 2016).

Margaret M Bradley and Peter J Lang. 1999. *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical Report. Technical report C-1, the center for research in psychophysiology, University of Florida.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research* 49 (2014), 1–47.

Aylin Caliskan, Joanna Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356 (04 2017), 183–186.

Aylin Caliskan and Molly Lewis. 2020. Social biases in word embeddings and their relation to human cognition. (2020).

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Academic press.

Yadolah Dodge. 2008. Spearman rank correlation coefficient. *The Concise Encyclopedia of Statistics. Springer, New York* (2008), 502–505.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. *arXiv preprint arXiv:1908.06361* (2019).

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems With Evaluation of Word Embeddings Using Word Similarity Tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. Association for Computational Linguistics, Berlin, Germany, 30–35. https://www.aclweb.org/anthology/W16-2506

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*. 406–414.

Nico H Frijda et al. 1986. *The emotions*. Cambridge University Press.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635–E3644.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simverb-3500: A large-scale evaluation set of verb similarity. *arXiv preprint arXiv:1608.00869* (2016).

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893* (2018).

Anthony G Greenwald, Debbie E. McGhee, and Joe L. Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology* 74 6 (1998), 1464–80.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 122–133.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096* (2016).

Eddie Harmon-Jones, Philip A Gable, and Tom F Price. 2013. Does negative affect always narrow and positive affect always broaden the mind? Considering the influence of motivational intensity on cognitive scope. *Current Directions in Psychological Science* 22, 4 (2013), 301–307.

Vasileios Hatzivassiloglou and Kathleen McKeown. 1997. Predicting the semantic orientation of adjectives. In *35th annual meeting of the association for computational linguistics and 8th conference of the european chapter of the association for computational linguistics*. 174–181.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics* 41, 4 (2015), 665–695.

Nora Hollenstein, Antonio de la Torre, Nicolas Langer, and Ce Zhang. 2019. CogniVal: A Framework for Cognitive Word Embedding Evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning*. 538–549.

Kamil K Imbir. 2015. Affective norms for 1,586 polish words (ANPW): Duality-of-mind approach. *Behavior research methods* 47, 3 (2015), 860–870.

Aycan Kapucu, Aslı Kılıç, Yıldız Özkılıç, and Bengisu Sarıbaz. 2018. Turkish emotional word norms for arousal, valence, and discrete emotion categories. *Psychological reports* (2018), 0033294118814722.

7212

Andrew Karpinski and Ross B Steinman. 2006. The single category implicit association test as a measure of implicit social cognition. *Journal of personality and social psychology* 91, 1 (2006), 16.

Wilhelm Kirch (Ed.). 2008. *Pearson's Correlation Coefficient*. Springer Netherlands, Dordrecht, 1090–1091. https://doi.org/10.1007/978-1-4020-5614-7_2569

Molly Lewis and Gary Lupyan. 2020. Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature human behaviour* (2020), 1–8.

Minglei Li, Qin Lu, Yunfei Long, and Lin Gui. 2017. Inferring Affective Meanings of Words from Word Embedding. *IEEE Transactions on Affective Computing* PP (07 2017), 1–1. https://doi.org/10.1109/TAFFC.2017.2723012

Minh-Thang Luong, Richard Socher, and Christopher D Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*. 104–113.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013).

Saif M Mohammad. 2016. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*. Elsevier, 201–237.

Brian A Nosek, Mahzarin R Banaji, and Anthony G Greenwald. 2002. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice* 6, 1 (2002), 101.

Brian A Nosek, Frederick L Smyth, Natarajan Sriram, Nicole M Lindner, Thierry Devos, Alfonso Ayala, Yoav Bar-Anan, Robin Bergh, Huajian Cai, Karen Gonsalkorale, et al. 2009. National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences* 106, 26 (2009), 10593–10597.

Charles E Osgood. 1964. Semantic differential technique in the comparative study of cultures. *American Anthropologist* 66, 3 (1964), 171–200.

Charles Egerton Osgood, William H May, Murray Samuel Miron, and Murray S Miron. 1975. *Crosscultural universals of affective meaning*. Vol. 1. University of Illinois Press.

Charles Egerton Osgood, George J Suci, and Percy H Tannenbaum. 1957. *The measurement of meaning*. Number 47. University of Illinois press.

Jeroen Paridon and Bill Thompson. 2019. subs2vec: Word embeddings from subtitles in 55 languages.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

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 (2007), 600–605.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*. 105–112.

James Russell and Albert Mehrabian. 1977. Evidence for a Three-Factor Theory of Emotions. *Journal of Research in Personality* 11 (09 1977), 273–294. https://doi.org/10.1016/0092-6566(77)90037-X

James A Russell. 1983. Pancultural aspects of the human conceptual organization of emotions. *Journal of personality and social psychology* 45, 6 (1983), 1281.

David S Schmidtke, Tobias Schröder, Arthur M Jacobs, and Markus Conrad. 2014. ANGST: Affective norms for German sentiment terms, derived from the affective norms for English words. *Behavior research methods* 46, 4 (2014), 1108–1118.

Tobias Schnabel, Igor Labutov, David M. Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *EMNLP*.

Ana Paula Soares, Montserrat Comesaña, Ana P Pinheiro, Alberto Simões, and Carla Sofia Frade. 2012. The adaptation of the Affective Norms for English words (ANEW) for European Portuguese. *Behavior research methods* 44, 1 (2012), 256–269.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Tommaso Teofili and Niyati Chhaya. 2019. Affect Enriched Word Embeddings for News Information Retrieval. *arXiv preprint arXiv:1909.01772* (2019).

Yulia Tsvetkov, Manaal Faruqui, and Chris Dyer. 2016. Correlation-based Intrinsic Evaluation of Word Vector Representations. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. 111–115.

Peter D Turney and Michael L Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)* 21, 4 (2003), 315–346.

Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research* 37 (2010), 141–188.

Lyle H. Ungar, Daniel Preotiuc-Pietro, and João Sedoc. 2017. Predicting Emotional Word Ratings using Distributional Representations and Signed Clustering. In *EACL*.

Melissa LH Vo, 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 (2009), 534–538.

Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C.-C. Jay Kuo. 2019. Evaluating Word Embedding Models: Methods and Experimental Results. *ArXiv* (2019).

Jin Wang, Liang-Chih Yu, K. Lai, and Xuejie Zhang. 2016. Community-Based Weighted Graph Model for Valence-Arousal Prediction of Affective Words. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24 (11 2016), 1–1. https://doi.org/10.1109/TASLP.2016.2594287

Amy Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods* 45 (02 2013). https://doi.org/10.3758/s13428-012-0314-x

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 2016. Building Chinese Affective Resources in Valence-Arousal Dimensions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, 540–545.

Appendix

Category		Chinese Stimuli
flowers	(target)	三叶草, 兰花, 玫瑰, 水仙花, 紫丁香, 郁金香, 雏菊, 百合, 紫色, 木兰
insects	(target)	蚂蚁, 跳蚤, 蜘蛛, 臭虫, 飞, 狼蛛, 蜜蜂, 蟑螂, 蚊子, 大黄蜂
instruments	(target)	风笛, 大提琴, 吉他, 琵琶, 长号, 班卓琴, 单簧管, 口琴, 曼陀林, 喇叭, 巴松管, 鼓, 竖琴, 双簧管, 大号, 钟, 小提琴, 大键琴, 钢琴, 中提琴, 邦戈, 长笛, 喇叭, 萨克斯风, 小提琴
weapons	(target)	箭头, 俱乐部, 枪, 导弹, 矛, 斧头, 匕首, 鱼叉, 手枪, 剑, 刀, 炸药, 斧头, 步枪, 罐, 炸弹, 火器, 刀子, 滑膛枪, 催泪瓦斯, 大炮, 手榴弹, 锤, 弹弓, 鞭子
pleasant	(attributes)	抚摸, 自由, 健康, 爱, 和平, 欢呼, 朋友, 天堂, 忠诚, 乐趣, 钻石, 温和, 诚实, 幸运, 彩虹, 文凭
unpleasant	(attributes)	滥用, 崩溃, 污秽, 谋杀, 疾病, 事故, 死亡, 悲痛, 毒, 臭, 突击, 灾害, 仇恨, 污染, 悲剧, 离婚, 监狱, 贫穷, 丑陋, 癌症, 杀, 烂, 呕吐, 痛苦, 监狱

Table 6: Chinese Stimuli (Word List)

Category		English Stimuli Collected from Caliskan et al. (2017)
flowers	(target)	clover, orchid, rose, daffodil, lilac, tulip, daisy, lily, violet, magnolia
insects	(target)	ant, flea, spider, bedbug, fly, tarantula, bee, cockroach, mosquito, hornet
instruments	(target)	bagpipe,cello, guitar, lute, trombone, banjo, clarinet, harmonica, mandolin, trumpet, bassoon, drum, harp, oboe, tuba, bell, fiddle, harpsichord, piano, viola, bongo, flute, horn, saxophone, violin
weapons	(target)	arrow, club, gun, missile, spear, axe, dagger, harpoon, pistol, sword, blade, dynamite, hatchet, rifle, tank, bomb, firearm, knife, shotgun, teargas, cannon, grenade, mace, slingshot, whip
pleasant	(attributes)	caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation
unpleasant	(attributes)	abuse , crash , filth , murder , sickness , accident , death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison

Table 7: English Stimuli (Word List)

Category		German Stimuli
flowers	(target)	Klee, Orchidee, Rose, Narzisse, Flieder, Tulpe, Gänseblümchen, Lilie, Veilchen, Magnolie
insects	(target)	Ameise, Floh, Spinne, Wanze, Fliege, Tarantel, Biene, Kakerlake, Mücke, Hornisse
instruments	(target)	Dudelsack, Cello, Gitarre, Laute, Posaune, Banjo, Klarinette, Mundharmonika, Mandoline, Trompete, Fagott, Trommel, Harfe, Oboe, Tuba, Glocke, Geige, Cembalo, Klavier, Bratsche, Bongo, Flöte, Horn, Saxophon, Violine
weapons	(target)	Pfeil, Keule, Waffe, Rakete, Speer, Axt, Dolch, Harpune, Pistole, Schwert, Klinge, Dynamit, Beil, Gewehr, Panzer, Bombe, Schusswaffe, Messer, Schrotflinte, Tränengas, Kanone, Granate, Streitkolben, Schleuder, Peitsche
pleasant	(attributes)	Liebkosung, Freiheit, Gesundheit, Liebe, Frieden, Jubel, Freund, Himmel, Treue, Vergnügen, Diamant, sanft, ehrlich, glücklich, Regenbogen, Diplom, Geschenk, Ehre, Wunder, Sonnenaufgang, Familie, glücklich, Lachen, Paradies, Urlaub
unpleasant	(attributes)	Missbrauch, Absturz, Schmutz, Mord, Krankheit, Unfall, Tod, Trauer, Gift, Gestank, Angriff, Katastrophe, Hass, Umweltverschmutzung, Tragödie, Scheidung, Gefängnis, Armut, hässlich, Krebs, töten, faul, Erbrechen, Qual, das Gefängnis

Table 8: German Stimuli (Word List)

Category		Polish Stimuli
flowers	(target)	koniczyna, orchidea, róża, narcyz, liliowy, tulipan, stokrotka, lilia, fiołek, magnolia
insects	(target)	mrówka, pchła, pająk, pluskwa, latać, tarantula, pszczoła, karaluch, komar, szerszeń
instruments	(target)	dudy,wiolonczela, gitara, flet, lutnia, puzon, banjo, klarnet, harmonijka, mandolina, trąbka, fagot, bęben, harfa, obój, tuba, dzwon, skrzypce, klawesyn, fortepian, altówka, bongo, róg, saksofon, skrzypce
weapons	(target)	strzałka, buława, strzelba, pocisk, włócznia, topór, sztylet, harpun, pistolet, miecz, nóż, dynamit, toporek, karabin, czołg, bomba, broń palna, ostrze, flinta, gaz łzawiący, armata, granat, buzdygan, proca, bat
pleasant	(attributes)	pieszczota, swoboda, zdrowie, miłość, dyplom, pokój, przyjemność, dopingować, przyjaciel, niebiosa, wierny, diament, delikatny, uczciwy, fartowny, tęcza, podarunek, honor, cud, rodzina, szczęśliwy, śmiech, raj, wakacje, świt
unpleasant	(attributes)	nadużycie, wypadek, brud, zabójstwo, choroba, awaria, śmierć, smutek, trucizna, smród,atak, katastrofa, nienawiść, zanieczyszczać, tragedia, rozwód, więzienie, bieda, brzydki, rak, zgniły, wymiociny, agonia, areszt, zło

Table 9: Polish Stimuli (Word List)

Category		Portuguese Stimuli
flowers	(target)	trevo, orquídea, rosas, narciso, lilás, tulipa, margarida, lírio, tolet, magnólia
insects	(target)	formiga, pulga, aranha, percevejo, mosca, tarântula, abelha, barata, mosquito, vespa
instruments	(target)	gaita de foles, violoncelo, violão, alaúde, trombone, banjo, clarinete, harmônica, bandolim, superada, fagote, tambor, harpa, oboé, tuba, sino, rabeca, cravo, piano, viola, bongo, flauta, chifre, saxofone, violino
weapons	(target)	flecha, porrete, arma de fogo, míssil, lança, machado, punhal, arpão, pistola, espada, lâmina, dinamite, machadinha, rifle, tanque, bomba, arma de fogo, faca, espingarda, gás lacrimogêneo, canhão, granada, maça, estilingue, chicote
pleasant	(attributes)	carícia, liberdade, saúde, amor, diploma, paz, prazer, alegrar, amigo, céu, leal,diamante, gentil, honesto, sortudo, arco-íris, prenda, honra, milagre, amanhecer, família, feliz,riso, paraíso, férias
unpleasant	(attributes)	maus-tratos, colisão, imundíce, assassinato, enfermidade, acidente, morte, tristeza, veneno, fedor, assalto, desastre, ódio, tragédia, poluir, divórcio, cadeia, pobreza, feio, cancro, matar, divórcio, cadeia, pobreza, feio, cancro, matar, podre, vómito, agonia, prisão

Table 10: Portugese Stimuli (Word List)

Category		Spanish Stimuli
flowers	(target)	trébol, orquídea, rosa, narciso, lila, tulipán, margarita, lirio, violeta, magnolia
insects	(target)	hormiga, pulga, araña, ácaro, mosca, tarántula, abeja, cucaracha, mosquito, avispón
instruments	(target)	cornamusa, violonchelo, guitarra, flauta, trombón, banjo, clarinete, harmónica, mandolina, trompeta, fagot, tambor, arpa, oboe, tuba, campana, fiddle, clave, piano, viola, bongo, flute, cuerno, saxofón, violín
weapons	(target)	flecha, palo, pistola, misil, lanza, hacha, daga, arpón, espada, cuchilla, dinamitar, rifle, tanque, bomba, naja, escopeta, cañón, granada, mazo, honda, látigo
pleasant	(attributes)	caricia, libertad, salud, amor, diploma, paz, placer, ánimo, amigo, cielo, leal, diamante, delicado, honesto, afortunado, arco-iris, obsequio, honor, milagro, amanecer, familia, feliz
unpleasant	(attributes)	maltrato, choque, inmundicia, asesinato, enfermedad, accidente, muerte, pena, ponzoña, hedor, asalto, desastre, odio, contaminar, tragedia, divorcio, cárcel, pobreza, feo, cáncer, matar, podrido, vómito, agonía, prisión

Table 11: Spanish Stimuli (Word List)

Category	Turkish Stimuli
flowers (target)	yonca, orkide, gül, nergis, leylak, lale, papatya, zambak, menekşe, manolya
insects (target)	karınca, pire, örümcek, tahtakurusu, sinek, tarantula, arı, hamamböceği, sivrisinek, eşekarısı
instruments (target)	gayda,çello, gitar, ut, trombon, banço, klarnet, mızıka, mandolin, trompet, fagot, davul, arp, obua, tuba, zil, keman, harpsikord, piyano, viyola, tamtam, flüt, boynuz, saksafon, viyolin
weapons (target)	ok, cop, tabanca, mermi, mızrak, balta, hançer, zıpkın, silah, kılıç, bıçak, dinamit, nacak, tüfek, tank, bomba, silâh, bıçak, çifte, gözyaşı gazı, gülle, bombası, topuz, mancınık, kırbaç
pleasant (attributes)	okşamak, özgürlük, sağlık, sevgi, barış, neşe, arkadaş,cennet, sadık, keyif, pırlanta, kibar, dürüst, şanslı, gökkuşağı,diploma, hediye, onur, mucize, gündoğumu, aile, mutlu, kahkaha,cennet, tatil
unpleasant (attributes)	istismar, çarpmak pislik cinayet, hastalık, ölüm , üzüntü , zehir , kokuşmuş , saldırı , felaket , nefret , kirletmek , facia , boşanmak , hapishane , fakirlik , çirkin , kanser , öldürmek , çürümüş , kusmuk , ızdırap , sancı, cezaevi

Table 12: Turkish Stimuli (Word List)