This version of the manuscript was accepted and will appear in the *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Nymph Piss and Gravy Orgies: Local and Global Contrast Effects in Relational Humor

Cynthia S. Q. Siew[1], Tomas Engelthaler[2], and Thomas T. Hills[2]

[1]Department of Psychology, National University of Singapore

[2]Department of Psychology, University of Warwick

Corresponding author:

Cynthia S. Q. Siew, PhD

9 Arts Link, Block AS4

Department of Psychology

National University of Singapore

Singapore 117570

Email: cynthia@nus.edu.sg

Abstract

How does the relation between two words create humor? In this paper, we investigated the effect of global and local contrast on the humor of word pairs. We capitalized on the existence of psycholinguistic lexical norms by examining violations of expectations set up by typical patterns of English usage (global contrast) and within the local context of the words within the word pairs (local contrast). Global contrast was operationalized as lexical-semantic norms for single-words and local contrast was operationalized as the orthographic, phonological, and semantic distance between the two words in the pair. Through crowdsourced (Study 1) and best-worst (Study 2) ratings of the humor of a large set of word pairs (i.e., compounds), we find evidence of both global and local contrast on compound-word humor. Specifically, we find that humor arises when there is a violation of expectations at the local level, between the individual words that make up the word pair, even after accounting for violations at the global level relative to the entire language. Semantic variables (arousal, dominance, concreteness) were stronger predictors of word pair humor whereas form-related variables (number of letters, phonemes, letter frequency) were stronger predictors of single-word humor. Moreover, we also find that semantic dissimilarity increases humor, by defusing the impact of low-valence words—making them seem more amusing—and by enhancing the incongruence of highly imageable pairs of concrete words.

Keywords: compound-word humor, semantic similarity, phonological distance

Introduction

The most prominent theories of humor argue that humor is fundamentally relational. These are perhaps most well-represented by absurdity theories, incongruity resolution theories, and most recently benign violation theory. Kant (1914) claimed that, “In everything that is to excite a lively convulsive laugh there must be something absurd”. The Latin *absurdus* means ‘out of tune’ and thus for a thing to be out of tune, there must be another tune for comparison—or at least a background hum. Absurdity is therefore relational, a violation of some expectation set up by the context. Or, as Schopenhauer [1883] (1969) (for an overview, see Roeckelein, 2006) put it, the “ludicrous” requires a “contrast.” Gallows humor, the kind that creates humor out of dark or life-threatening situations, is a good example: have you heard the one about the holocaust survivor who goes to use the toilets while visiting the Auschwitz-Birkenau Memorial only to be asked to pay to use them. The elderly survivor rolls up his sleeve revealing a tattooed number and says, “The last time I was here, I didn’t have to pay.” (Richman, n.d.)

This humor of contrast has been proposed by some to be the output of a faulty-logic detection system (Minsky, 1981). The feeling of humor highlights the curious underlying logic of a situation and therefore calls into action our cognitive resources (Hurley et al., 2011). This theory is summarized in the benign violation theory (McGraw & Warren, 2010), which makes the simple prediction that humor requires stimuli that violate our expectations—somehow catching us off guard—while simultaneously being unthreatening. In a more general sense, humor is therefore a kind of mid-to-high valence entropy, a form of positive surprisal. Notably, entropy has been used successfully to quantify humor (Westbury et al., 2016).

If we want to take apart what is funny about the Auschwitz survivor at the museum (if it is even funny to you at all), the challenge is to describe the many dimensions along which the

situation represents a contrast. This is not trivial. Humor *in the wild* can be absurd along many dimensions and isolating what those are is challenging qualitative work. Several recent articles have tried to take the fun out of humor research and examine it by focusing on the humor of individual words, what might be considered the "fruit-fly" of humor (Engelthaler & Hills, 2018; Westbury & Hollis, 2019). What that research clearly demonstrated is that people can reliably evaluate the humor of single words. For example, which is funnier, the word *porridge* or the word *oatmeal*? Most people agree that *porridge* is funnier than *oatmeal*. This may at first glance appear to violate a relational theory of humor because it is not obvious what the context is for a word on its own. However, the data from Engelthaler and Hills (2018) suggest that the violation may be as simple as word frequency. Lower frequency words tend to be rated as more humorous than higher frequency words; inverse frequency is the strongest predictor of single word humor. Westbury and Hollis (2019) go on to show that low probability orthographic or phonological structure are also well correlated with humor of individual words, further suggesting that single word humor is the outcome of a cognitive process for entropy detection.

The natural extension of single word humor is to ask if these results scale up to multiword humor. In this article we address this question by building upon the prior work of Engelthaler and Hills (2018) and Westbury and Hollis (2019), making a simple alteration of their prior research on single words, by adding a second word. Now instead of facing our participants with the task of rating individual words, like *cage* (which is not particularly funny on its own) or *cabbage* (only mildly funnier), our participants are faced with rating the humor of *cabbage cage*, which is arguably funnier than either word alone. But why?

Compound words are combinations of two words into a single unit (for example, "school bus"), and offer a sizeable set of dimensions along which the single-word constituents that make

up the compound could differ. In addition to the individual word features studied in prior research (such as valence, concreteness, and length), compound words offer additional *relational measures*, allowing us to contrast one word directly in relation to the expectations set up by another word. In this present paper, we attempted to conduct a large-scale investigation into the humor of word pairs, building on both early investigations of word-pair humor by Godkewitsch (1974) as well as more recent attempts by Westbury and Hollis (2021) and Kang (2016).

Given the potential diversity of the set of relational measures, we focus here on a reduced set of relational measures which include *form*—using orthographic and phonological distance between the two words—and *semantic distance*—using a large-scale corpus analysis of semantic space, which identifies words with similar related meanings. Form measures allow us to examine expectations set up by the way a word looks and sounds, such as the phonological similarity between *moose* and *ooze,* which share an orthographic and phonological 'oo' (/u:/). Semantic similarity allows us to examine expectations set up by the semantic context, such as the semantic leap formed when the word *apron* is followed by the word *forehead,* as opposed to the semantic familiarity set up by following the word *power* by the word *influence.*

Using this simplified set of comparisons, we are able to address in what way the relation between two words creates humor. This contrast sets up a refinement of previous hypotheses. In one sense, we may expect the *contrast* to be between the two words themselves. One word sets up an expectation that is then violated by the second word, and the violation of that expectation leads to a contrast. We refer to this as *local contrast.* In the studies below, we operationalize local contrast as the orthographic, phonological, and semantic distance between the two words in the compound (i.e., the word-pair predictors).

However, there is another useful sense of contrast set up by the prior work of Westbury et al. (2016), which measured entropy of letter strings (based on individual letters, letter pairs, and letter triplets) that made up nonsense words. In this case, the expectation for a single word (or pair of words) is based on the entire English language. We call this *global contrast.* Similar results may be inferred from the correlations between single-word humor and low frequency and low probability orthography and phonology found in Engelthaler and Hills (2018) and Westbury and Hollis (2019), respectively.

Here we operationalize global contrast as the lexical-semantic norms for single-words (i.e., word-level predictors). The word pairs we use are extremely low frequency and absent from the corpora we examine. Estimating the global contrast of the compound as a whole is not possible. Instead, the single word norms, collected or computed for thousands of English words, represent the global expectation (the background "hum") surrounding that particular form-based or semantic feature. To illustrate this point, consider the following example of a single-word-level predictor—word frequency, which is how many times a word occurs in natural language corpora. If frequency is a significant predictor of humor such that *less* frequently occurring words are funnier, this would constitute a violation of global expectations because encountering less frequently occurring words is unexpected given one's experience with language. Hence, if we observe that the single-word measures are predicting humor and we can establish that this is a violation of expectations set up by typical patterns of English usage, then we may conclude that humor can by driven at the level of global contrast, as observed for single word humor. In addition, if we also observe an effect based on the distances between the two words in the compound, then we may conclude that local contrast is also playing a role.

In the two studies we describe next, we find evidence for both of these effects. In Study 1, we first examined local and global contrast effects for a large set of randomly generated word pairs using large crowd-sourced population of participants. Study 2 is a pre-registered follow-up to Study 1, which selects a specific set of word pairs based on the predictive contrasts observed in Study 1, and then uses Hollis' (2018) best-worst scaling to rank these word pairs for humor.

Study 1

Because the number of possible word pairs that could be generated from even a limited set of words (i.e., from the Engelthaler and Hills (2018) single-word humor norms containing $4{,}997^2 \sim 25$ million pairs) was very large, we deliberately adopted an approach that crowdsourced humor ratings from volunteers who viewed randomly generated pairs of words on a web application.

Method

R Shiny application

We created an R Shiny application to collect humor ratings of word pairs using the *shiny* R library (Chang et al., 2020). The application is hosted on the RStudio server and can be accessed at https://csqsiew.shinyapps.io/humorous_phrases/. The R code used to create the application can be found on the first author's Github page (https://github.com/csqsiew/shinyhumor).

Once the application was loaded, a pair of words was randomly selected from the 4,997 words in the Engelthaler and Hills (2018)'s humor norms (available on https://github.com/tomasengelthaler/HumorNorms). The visitor was asked to decide if the word

pair was humorous or not by clicking on one of the two buttons labelled “Humorous” (left side) and “Humorless” (right side; see Figure 1). There was no time limit for the visitor to respond. Once the response was submitted, a new pair of words was randomly generated. Visitors were able to continue responding to as many of these word pairs as they wished, and were free to stop at any time (*N.B.*, this was also clearly indicated at the bottom of the application). When the visitor exited the application, this triggered a function that recorded all word pairs shown to the visitor and their responses for each word pair (coded as 1 for “Humorous” and 0 for “Humorless”), and saved the data to the first author’s personal Dropbox account. The data was never saved to the R Studio server and no other identifying information was collected from the visitor in order to ensure complete anonymity. Ethics approval for Study 1 was obtained from the University of Warwick.

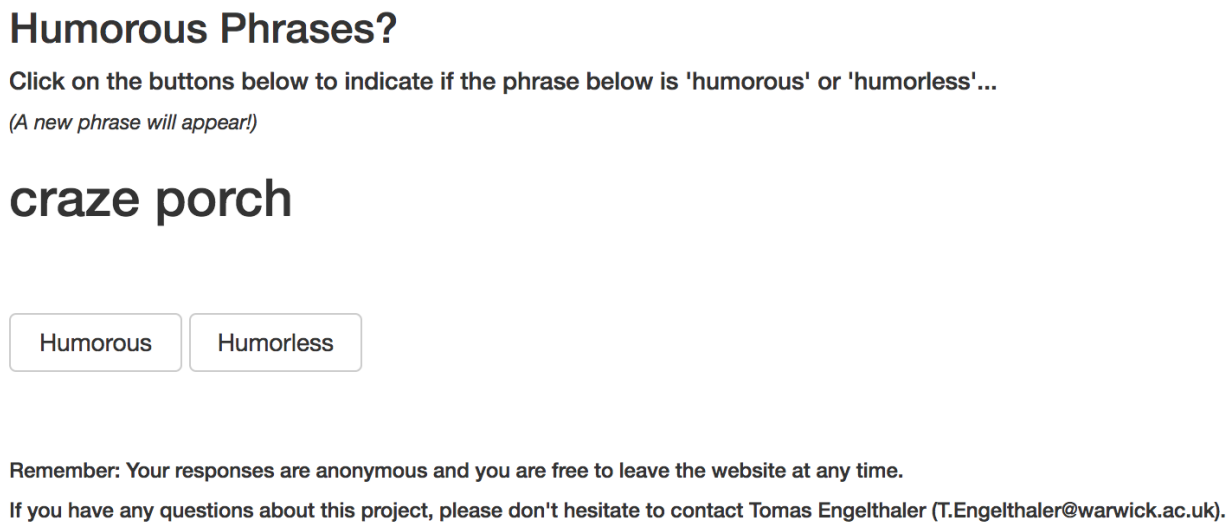

Figure 1. Screenshot of application.

The application was officially launched on 23rd October 2017. Data collection was facilitated by promoting the application through the third author’s popular science blog, and

through word of mouth and social media. The data compiled for all analyses described in the remainder of the paper included all responses collected from 23rd October 2017 to 27th May 2020 (dates inclusive). The raw data from this period is freely available on the Open Science Framework repository for this paper (see Authors' Note).

Predictors

We were interested in examining how characteristics of the words in the compound (i.e., word-level predictors representing *global contrast*) and the relationships between the two words in the word pair (i.e., word-pair predictors representing *local contrast*) influenced the probability that the word pair was rated as humorous or not. Each of these predictors is described in further detail below.

Global contrast: Word-level predictors

Word-level predictors can be classified into two groups: A set of predictors describing the word-form characteristics of individual words (i.e., based on its orthographic and phonological features, frequency in the language) and a second set of predictors describing the lexico-semantic characteristics of individual words. Table A1 in the Appendix shows the descriptive statistics and correlations among the word-level predictors.

Form predictors

1. *Orthographic length* or *Number of letters*. This was obtained by counting the number of letters in the word's orthographic form.

2. *Phonemic length* or *Number of phonemes*. Phonological transcriptions were obtained from the English Lexicon Project (ELP; Balota et al., 2007; http://elexicon.wustl.edu/). Characters indicating stress and syllable boundaries were removed, and "2-character"

segments were converted to a single character so that the length of the phonological transcription directly corresponded to the number of phonemic segments.

3. *Log letter probability.* Following Westbury and Hollis (2019), we included log letter probability as a predictor. This measure represented the logged average probability of the letter strings in each word, computed based on approximately 4.5 billion characters of English text (Lyons, n.d.).

4. *Log phoneme probability.* Following Westbury and Hollis (2019), we also included log phoneme probability as a predictor. This measure represented the logged average probability of the phonemic strings in each word, computed based on phoneme frequencies from Blumeyer (2012).

5. *Log frequency.* Frequency values were obtained from the ELP; specifically, the subtitle (SUBTLEX) frequencies based on the $SUBTLEX_{US}$ corpus (Brysbaert & New, 2009).[1]

Semantic predictors

6. Single-word *humor* ratings from Engelthaler and Hills (2018). Words with high humor ratings were perceived to be humorous (e.g., *booty, tit*), as compared to words with low humor ratings (e.g., *gunshot, torture*).

7. *Valence* ratings from the Warinner et al. (2013) affective norms. Valence refers to the pleasantness of a word. Words with high valence are associated with positive affect (e.g., *excited, relaxing*), whereas words with low valence are associated with negative affect (e.g., *rapist, murder*).

8. *Arousal* ratings obtained from the Warinner et al. (2013) affective norms. Arousal refers to the intensity of the emotion invoked by the word. Words with high arousal elicit

greater emotional intensity (e.g., *erection, terrorism*), whereas words with low arousal elicit low levels of emotional intensity (e.g., *grain, librarian*).

9. *Dominance* ratings obtained from the Warinner et al. (2013) affective norms. Dominance refers to the degree of control exerted by a word. Words with high dominance are words that participants perceive to be able to exert high control on (e.g., *successful, smile*), whereas words with low dominance are words that participants perceive to be unable to exert control over (e.g., *dementia, lobotomy*).

10. *Concreteness* ratings obtained from Hollis et al. (2017)'s extrapolated concreteness values. We used the Hollis norms instead of the commonly used Brysbaert et al. (2014) concreteness norms in order to minimize the number of words that did not have concreteness ratings in the Brysbaert norms. Hollis et al. (2017) used skip-gram vector representations to infer concreteness for over 70,000 words from human judgments of concreteness and has been shown to have high validity. Concreteness refers to the extent to which a word's referent was concrete or abstract. Words with high concreteness ratings have highly concrete referents (e.g., *yarn, museum*), whereas words with low concreteness ratings have highly abstract referents (e.g., *liberty, nifty*).

Local contrast: Word-pair predictors

This set of predictors consisted of 3 "distance" predictors representing the orthographic distance, phonological distance, and semantic similarity between the two words within the word pair. Orthographic and phonological distance would be classified as *form* word pair predictors, semantic similarity would be a *semantic* word pair predictor.

11. Orthographic distance and 12. Phonological distance. The orthographic distance between two words in a given word pair was the Levenshtein distance between the letter

strings. The phonological distance between 2 words in a given word pair was the Levenshtein distance between the phonological transcriptions. Levenshtein distance refers to the number of substitutions, additions, or deletions (of letters/phonemes) needed to convert one string into another string, and has been previously used to quantify phonological and orthographic similarity among words (Suárez et al., 2011; Yarkoni et al., 2008).

13. Semantic similarity. The semantic similarity between two words was computed based on the word embeddings developed by Li et al. (2019). Each word is initially represented as a 50,000-dimensional vector, encoding the number of times a word co-occurs with the 50,000 most frequent words in the English language. These vectors were derived from the Google Ngrams database of 5-grams for the year 2000 (Michel et al., 2011), which lists the frequency of 5-grams in ~ 4% of published books for that year. We scan through the frequency list to construct a high dimensional vector on a per-word basis, defining 'co-occurrence' as any time two words appear in the same 5-gram, multiplied by the frequency of the respective 5-gram. Positive pointwise mutual information (PPMI) was computed for each pair of words before reducing the dimensions of the word embeddings to 300 using singular value decomposition. More details about the training procedure and justification can be found in Li et al. (2019). The semantic similarity of compound words was computed via taking the cosine similarity of these word embeddings.

Results

Characteristics of the crowdsourced data

A total of 55,100 valid ratings from 597 unique sessions were obtained during the data collection period specified in the Methods section. The number of ratings obtained from each unique session ranged from 1 to 1,487, with a mean of 92.3 ratings (*SD* = 160.4) and median of 37 ratings. Note that the 597 unique sessions did not necessarily come from 597 independent visitors to the application because it was possible for the same person to visit the website on separate occasions and this would register as separate sessions. As we did not collect further information about the visitors there was no way of knowing how many times this occurred.

Out of the 55,100 ratings, 13,341 (24.2%) were “Humorous” and 41,759 (75.8%) were “Humorless”. This is consistent with the positive skew observed in the Engelthaler and Hills (2018) humor norms, where the majority of words were rated as humorless. These ratings were provided for a total number of 55,047 unique word pairs, with 56 word pairs shown twice. Note that the frequency of each of the unique word pairs generated by the application was 0 in the Touchstone Applied Science Associates, Inc. (TASA) corpus used to develop The Educator's Word Frequency Guide (http://lsa.colorado.edu/spaces.html). Hence, while TASA bigram frequency was not informative, it is at least controlled for, because these compounds did not occur in the corpus or were at least of very low frequency in naturally occurring language.

After compiling all the word norms, measures, similarities, and phonological transcriptions from various sources and databases (see Method), we excluded words for which part of the information was unavailable. This resulted in a set of 4,411 words out of the original 4,997 words (88.3%) from the humor norms. Based on this set, we were able to compute all the word-level and word-pair predictors for 43,059 out of 55,100 word pairs (78.1%).

Linear regression of single-word humor norms

To provide a useful comparison with the compound words, we first ran an independent linear regression on the individual words used in our study, which represent a subset (4,411 out of 4,997) of the humor word norms provided by Engelthaler and Hills (2018). Humor ratings of individual words from the original single-word humor norms by Engelthaler and Hills were regressed on single-word norms (i.e., the word-level form and semantic predictors). Number of letters, number of phonemes, letter frequency, word frequency, valence, and arousal were significant predictors of single-word humor (see Table 1).

Logistic regression of compound word ratings

As the outcome variable was binary (i.e., whether the word pair was humorous or humorless), a logistic regression model was implemented with the following predictors: number of letters, number of phonemes, (log) letter frequency, (log) phoneme frequency, (log) word frequency, humor, valence, arousal, dominance, concreteness, orthographic distance, phonological distance, and semantic similarity. For each compound, the mean of word-level predictors (i.e., number of letters, number of phonemes, (log) letter frequency, (log) phoneme frequency, (log) word frequency, humor, valence, arousal, dominance, concreteness, for the first and second word in the compound) was computed and included as predictors. Note that the overall result did not change when word-level predictors were included separately for each word in the compound. All predictors were mean-centered and scaled prior the logistic regression.

In addition, to be as conservative as possible, the full model was submitted to a stepwise forward and backward search procedure (by eliminating and adding 1 variable at a time) that aimed to minimize AIC by only including the optimal set of predictors in the final model. We also conducted LASSO regression such that the coefficients of predictors with smallest effect

sizes were suppressed to 0. A summary of the fixed effects of the predictors for each of the models is shown in Table 1 below.

As this is a logistic regression model with a binary DV, note that standardized odd ratios (ORs) are provided instead of the typical regression coefficients. ORs *greater than 1* indicate that higher values of the predictor were associated with *higher* probability of the compound rated as humorous. ORs *less than 1* indicate that higher values of the predictor were associated with *lower* probability of the compound rated as humorous.

Discussion

When compared with the results from the single-word humor regression model, more semantic variables (arousal, dominance, concreteness) were significant predictors of compound word humor whereas form-related variables (number of letters, phonemes, letter frequency) tended to be stronger predictors of single-word humor. Overall, the results are consistent across the various models—compounds containing funny, highly arousing, concrete, less dominating, low frequency words tend to be rated as humorous. In addition, compounds with lower orthographic and phonological distance were more likely to be rated as humorous in the linear regression model, and compounds with lower semantic similarity were more likely to be rated as humorous in both the linear and LASSO models. To put it in another way, we find that even after controlling for the influence of global contrast (as operationalized via the inclusion of lexical-semantic measures of single words), local contrast between the two words affected compound word humor as well.

Furthermore, if we compare the predictors retained in the LASSO regression against the significant predictors in the linear regression model, it is clear that semantic predictors are the

core contributors of compound-word humor. Finally, when contrasted with the results of the linear regression predicting single-word humor, it appears that semantic variables (arousal, dominance, concreteness) were stronger predictors of compound-word humor whereas form-related variables (number of letters, phonemes, letter frequency) were stronger predictors of single-word humor.

Table 1. Summary of Study 1 and Study 2 regression model results. Dark grey rows = higher values are associated with *greater* humor. Light grey rows = higher values are associated with *less* humor.

Predictors	Single word humor		Study 1					Study 2				
			Full model		Stepwise search			Full model		Stepwise search		
	Std. b	p	OR	p	OR	p	LASSO (Std. b)	Std. b	p	Std. b	p	LASSO (Std. b)
Humor			1.47	**<0.001**	1.47	**<0.001**	0.349	0.042	**<0.001**	0.042	**<0.001**	0.042
Valence	0.25	**<0.001**	0.97	**0.036**	0.97	**0.035**		-0.004	0.096	-0.004	0.098	
Arousal	0.10	**<0.001**	1.08	**<0.001**	1.08	**<0.001**	0.016	0.007	**<0.001**	0.007	**<0.001**	0.003
Dominance	0.02	0.254	0.92	**<0.001**	0.92	**<0.001**	-0.056	-0.007	**0.012**	-0.007	**0.011**	-0.006
Concreteness	0.02	0.166	1.18	**<0.001**	1.18	**<0.001**	0.098	0.011	**<0.001**	0.011	**<0.001**	0.007
No. of letters	-0.16	**<0.001**	1.06	**0.031**	1.06	**0.042**		0.009	**<0.001**	0.009	**<0.001**	
No. of phonemes	-0.06	**0.02**	1.04	0.204	1.05	0.081		x		x		x
Letter frequency	-0.16	**<0.001**	0.94	**<0.001**	0.94	**<0.001**	-0.007	0.001	0.553			
Phoneme frequency	-0.01	0.735	1.01	0.433				x		x		x
Word frequency	-0.41	**<0.001**	0.93	**<0.001**	0.93	**<0.001**	-0.049	-0.008	**0.002**	-0.009	**0.002**	-0.011
Orthographic distance			0.92	**<0.001**	0.92	**<0.001**		-0.007	0.063	-0.007	0.071	-0.001
Phonological distance			0.94	**0.004**	0.93	**0.002**		-0.013	**<0.001**	-0.013	**<0.001**	-0.011
Semantic similarity			0.90	**<0.001**	0.90	**<0.001**	-0.033	-0.004	0.10	-0.004	0.10	-0.0005

Study 2

The results of Study 1 showed that funny word pairs tend to (i) contain funny, highly arousing, concrete, less dominating, low frequency words, and (ii) have lower orthographic and phonological distance and lower semantic similarity from one another in the local context. The aim of Study 2 was to validate the results from Study 1 by collecting humor estimates for a new set of word pairs. Study 2 was pre-registered and details can be found at this link: https://osf.io/b8ftw.

Method

Stimuli selection

First, the predicted probability that a given word pair would be rated as funny was computed for ~16 million word pairs (representing the number of possible pairwise permutations of the words used to generate random word pairs in the previous study) using the regression weights derived from the predictors in the full logistic regression model in Study 1. These word pairs were then sorted based on their predicted probabilities, or predicted humor rating (PHR), and sampled such that the distribution of PHR in the set of selected compounds was as uniform as possible and with the criteria that no two words were ever repeated in the sample. This resulted in a final sample of 732 compounds, constructed from 1464 unique words. Table A2 in the Appendix shows the descriptive statistics and correlations among the word-level and word-pair predictors for the 732 compounds.

Best-worst scaling

Instead of collecting humor ratings, we adopted the methodology for judgments known as "best-worst scaling" first developed by Louviere and Woodworth (1990; see also Louviere,

Flynn, & Marley, 2015; Marley & Islam, 2012). We followed the specific methodology of Hollis (2018) and collected "best" and "worst" judgments of humor from a set of 4 compounds. These best-worst judgments were then used to compute rank order information for the set of 732 compounds on a latent variable (i.e., humor). Each participant was presented a group of 4 compounds and had to choose, from that set of 4, the compound that was the most humorous (the "best" judgment) and the compound that was the least humorous (the "worst" judgment). A value that conceptually corresponds to the probability that a given item will "beat" other items, such that higher values correspond to the item having a higher value on the latent variable, was computed using the "Value Scoring" algorithm described in Hollis (2018). Therefore, a compound with a high value is very humorous as it is rated as being more humorous than other compounds most of the time.

Procedure

Simulations indicated that for 732 items presented in sets of 4, a total of 5,856 trials is required to derive accurate estimates (see Hollis, 2018; Experiment 4). Since each trial contained a set of 4 compounds, each participant provided best-worst judgements for 183 trials. In order to reach 5,856 trials, a total of 32 participants were recruited via the Amazon Mechanical Turk platform and reimbursed for their participation.

All participants provided best-worst ratings for the same set of 732 word pairs, presented in sets of 4 (i.e., 183 trials). Trials were pseudo-randomized to ensure that permutations of items are not inadvertently duplicated across participants and that each participant only saw each of the 732 word pairs once. For each set of 4 word pairs, participants were instructed to first choose the funniest word pair (i.e., the "best") followed by the least funny word pair (i.e., the "worst"). Ethics approval for Study 2 was obtained from the University of Warwick.

Results

Manipulation check

To ensure that participants were doing the study properly, we conducted a participant compliance analysis using the Python scripts available at https://sites.ualberta.ca/~hollis/. This additional analysis is in line with best practices described in Hollis (2018) and also followed by Westbury and Hollis (2019). The participant compliance analysis assesses the reliability of each individual participant's ratings by comparing them to the population and returns a compliance score ranging from 0% to 100%, where high values correspond to greater compliance. Based on this analysis, mean compliance was 69.6% (SD = 9.3) and no participant had a compliance score that was less than 3 standard deviations below the mean of all participants. This indicated that participants were indeed doing the task properly and their best-worst ratings were reliable.

Correlation analysis

Value scores were obtained by using the scripts from Hollis (2018) to compute a value for each word pair based on the "value scoring" algorithm. Although many other scoring algorithms exist for best-worst scaling, the "value scoring" algorithm was shown to be the best measure based on the simulations conducted by Hollis (2018), and it has been previously used to compute value scores for the humor of individual words in Westbury and Hollis (2019). As discussed above, higher values correspond to the compound having a higher value on the latent variable of humor; hence, a compound with a high value is very humorous as it is rated as being more humorous than other compounds most of the time.

Figure 2 shows the scatterplot of predicted humor probability (based on Study 1) and value scores for 732 compounds. Value scores were highly correlated with the predicted probability estimates from our model in Study 1, $r = .79$, $p < .001$.

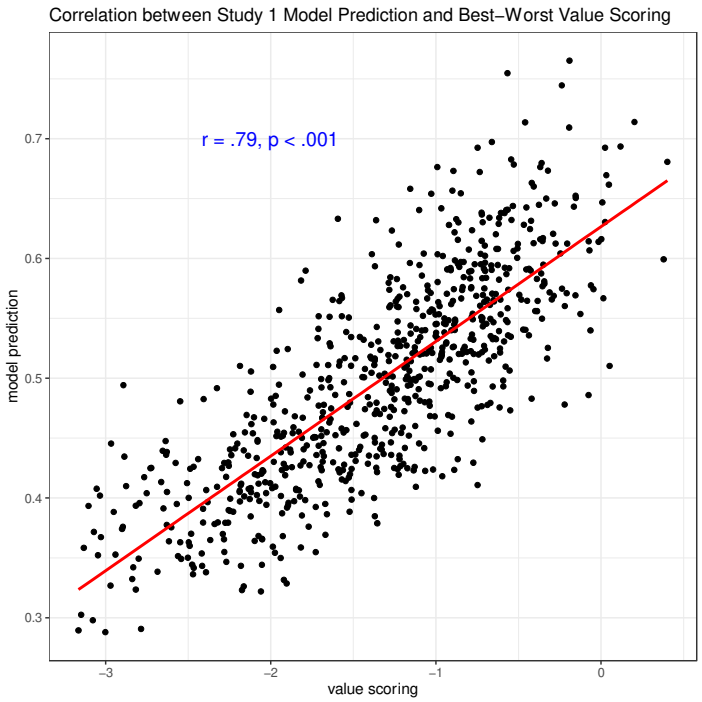

Figure 2. Scatterplot of predicted humor probability (from Study 1 model) and value score (from Study 2) for 732 compound words.

Regression analysis

We also conducted a series of regression analyses with the value scores as the dependent variable to further validate our original model (which predicted the probability that a given word pair was funny or not) against the new data collected (i.e., value scores from the set of 732 compounds) with the following predictors: number of letters, number of phonemes, (log) letter frequency, (log) phoneme frequency, (log) word frequency, humor, valence, arousal, dominance, concreteness, orthographic distance, phonological distance, and semantic similarity. Similar to Study 1, for each compound word, the mean of word-level predictors (i.e., number of letters, number of phonemes, (log) letter frequency, (log) phoneme frequency, (log) word frequency,

humor, valence, arousal, dominance, concreteness, for the first and second word in the compound) was computed and included as predictors. Note that the overall result did not change when word-level predictors were included separately for each word in the compound. All predictors were mean-centered and scaled prior the analysis.

Because Study 2 served as a confirmatory study of the results from Study 1, the set of significant predictors in the model returned by the step-wise search in the previous study was included as predictors of the humor of the 732 compounds. This list of predictors included: humor, valence, arousal, dominance, concreteness, number of letters, letter frequency, word frequency, orthographic distance, phonological distance, and semantic similarity. This model was then submitted to a stepwise forward and backward search procedure (by eliminating and adding 1 variable at a time) that aimed to minimize AIC by only including the optimal set of predictors in the final model. We also conducted LASSO regression such that the coefficients of predictors with smallest effect sizes were suppressed to 0. A summary of the fixed effects of the predictors for the linear and LASSO models is shown in Table 1.

Discussion

Overall, the pattern of findings was generally consistent across both Study 1 and 2, and across the different types of analyses (full model, stepwise, LASSO) conducted. Compound words containing funny, highly arousing, concrete, less dominating, low frequency words tend to have higher value scores, as well as compounds with lower orthographic and phonological distance and lower semantic similarity. Though semantic similarity is not significant at the .05 level in Study 2, it is nonetheless highly correlated with some of the other predictors (e.g., the zero-order correlations between semantic similarity and mean compound-word humor,

concreteness, and frequency are as follows: r_{humor} = -0.36, r_{conc} = -0.34, r_{freq} = 0.33, all *p*s < .001). The negative correlation between semantic similarity and single word humor may indicate that word pairs containing individually humorous words tend to be more semantically dissimilar than word pairs containing non-humorous words. Collectively, the results suggest that compound-word humor enjoys both global and local contrast effects.

Exploring the influence of other semantic variables on compound-word humor

In this section, we report additional analyses conducted on Study 1 and Study 2 data to explore the influence of other semantic variables on compound-word humor. This section serves two goals: First, there are various ways in which the semantics of words can be quantified. Hence, it is important to explore if additional indexes of semantic relationships between words would also predict humor. Second, the results of these analyses could provide potentially relevant points of connection from the present work of humor in language to the psycholinguistic literature on the processing of compound words, as well as prior work on humor single word.[2]

Distance to the semantic category of funny words

In Westbury and Hollis (2019)'s extensive analysis of the humor of single words, they found that a measure called *Average-CDV* emerged as a strong predictor of single-word humor. In a recent paper examining the humor of adjective-noun pairs, they also found that the Average-CDV of the noun was a strong predictor of the humor of word pairs (Westbury & Hollis, 2021). Average-CDV is a measure of how distant a word is from the general category of funny words in the semantic space obtained by computing the distance between a word's semantic vector and the average of the vectors of pre-defined word sets (detailed computation notes can be found in

Westbury & Hollis, 2019). There is a subtle but key difference between this variable and the single-word humor rating. Specifically, while two words could have similar humor ratings (e.g., ‘king’ and ‘textile’ both have an average humor rating of 2), they could still differ based on how good of a fit that word is to the broad category of “funny” concepts in the semantic space (i.e., ‘king’ has a CDV of 0.88 and ‘textile’ has a CDV of 1.18). Hence, given the previous results reported by Westbury and Hollis, it would be worthwhile to explore if including these variables may improve our models from Study 1 and Study 2.

The regression summary table for the original model that also included predictors of the Average-CDV of the first and second word in the word pair can be found in the Appendix (see Table A3 for Study 1 results and Table A4 for Study 2 results). Note that in this section, the regression models included the entire set of lexico-semantic predictors from Study 1 and 2 and for the first and second word separately. Including the CDV predictors led to significant improvement in model fit as compared to the baseline model without those predictors, Study 1: $\chi^2(2) = 278.4$, $p < .001$., and Study 2: $F(2) = 22.79$, $p < .001$. The overall pattern of results reported in Study 1 and Study 2 did not change. Both CDV1 and CDV2 were significant predictors, Study 1: CDV1: $z(37120) = -12.35$, $p < .001$; CDV2: $z(37120) = -11.38$, $p < .001$, and Study 2: CDV1: $t(706) = -4.82$, $p < .001$; CDV2: $t(706) = -4.97$, $p < .001$. Word pairs that contained words that were semantically closer to the category of humor words (i.e., low distance) were more likely to be rated as funny word pairs.

Distance to the entire compound

As seen in the psycholinguistic literature, compound words provide a rich source of linguistic stimuli for studying how people interpret the meaning of the entire expression based on

the constituents that make up the expression (Falkauskas & Kuperman, 2015; Gagné, 2001; Günther & Marelli, 2016). Compound words are made of single word constituents, comprising a modifier (e.g., police-) and a head (e.g., -man). Because the compounds in the current study are made up of two individual words randomly selected from a corpus, it would be worth exploring if mechanisms that are involved in the interpretation of a novel compound may also be implicated in the processing of compounds for their humor. In order for speakers to produce and comprehend compound words efficiently, speakers likely possess powerful meaning-composition systems that enable them to quickly combine familiar constituents into a single novel representation (Downing, 1977; Libben, 2014). This meaning-composition operation is influenced by the linguistic and semantic properties of the constituents themselves (Günther & Marelli, 2016), as well as the language experience that speakers bring to bear (Falkauskas & Kuperman, 2015).

In the present study, we showed that greater semantic dissimilarity between the words that made up the word pair led to enhanced compound-word humor. Here we examined whether the similarity between the first constituent and the entire compound (i.e., constituent1-compound), as well as the similarity between the second constituent and the entire compound (i.e., constituent2-compound), might improve our models from Study 1 and Study 2. Because these measures were obtained from the model by Günther and Marelli (2020), to be consistent the measure of semantic similarity between the two individual words was also derived from the same model rather than re-using the measure obtained from the Macroscope.

The regression summary table for the original model that also included the constituent-compound predictors can be found in the Appendix. Including the constituent-compound predictors led to marginal improvement in model fit as compared to the baseline model without

those predictors for Study 1 data, $\chi^2(2) = 5.55$, $p = .06$, but not Study 2; $F(2) < 1$, $p = .83$. The overall pattern of results reported in Study 1 and Study 2 did not change (see Table A3 for Study 1 results and Table A4 for Study 2 results). For Study 1 data, there was a small but significant effect of constituent1-compound, $z(43033) = 2.36$, $p = .02$, but the effect of constituent2-compound was not significant, $z(43033) < 1$, $p = .80$. Neither of these effects were significant predictors in the re-analysis of Study 2 data. The re-analysis of Study 1 data indicated greater similarity between the first word in the word pair to the entire word pair was associated with greater compound-word humor.

Discussion

Consistent with prior work from Westbury and Hollis, there was a strong effect of Average-CDV on humor. As a reminder, Average-CDV is a measure of how distant a word is from the general category of funny concepts. Words with a high CDV distance are further from the category of funny concepts and words with a low CDV distance are closer to this category. Although CDV and single word humor are indeed highly correlated with each other, as one might expect ($r = -0.46$, $df = 4098$, $p < .001$), the results from the regression analysis indicate that how close a given word is to the space of humorous concepts is accounting for additional variance beyond the humorous-ness of the word itself.

In contrast, the constituent-compound predictors were not as strong predictors of compound-word humor, even though they have been found to be important predictors for the perceived meaningfulness of compounds (Günther & Marelli, 2016) and in compound processing (Günther & Marelli, 2020). The small but significant effect of constituent1-compound similarity suggests that some non-negligible amount of automatic meaning construction is occurring when

processing the word pairs for humor. Furthermore, the direction of the constituent1-compound effect is in line with prior literature on compound processing that showed that this predictor successfully predicted the acceptability or meaningfulness judgments of compounds. In other words, the "meaningfulness" of the word pair could play a role in humor. Perhaps simply containing semantically dissimilar constituents is merely a prerequisite for humor—if some sort of hidden, but still meaningful, higher-order relation was discovered to also exist between the first word and the entire expression that could be yet another contributor of humor (Kang, 2016).

General Discussion

In this paper, we investigated the effect of global and local contrast on relational humor. Through crowdsourced (Study 1) and best-worst (Study 2) ratings of the humor of a large set of word pairs, we find evidence of both global and local contrast on compound-word humor. In analyses predicting compound-word humor, we observe that humor arises when there is a violation of expectations at the level of the relationship between the two words that make up the compound, even after accounting for violations at the global level relative to the entire language. When contrasted with the results of the regression predicting single-word humor, we find that semantic variables (arousal, dominance, concreteness) were stronger predictors of compound-word humor whereas form-related variables (number of letters, phonemes, letter frequency) were stronger predictors of single-word humor.

Existing theories of humor like benign violation theory provide a useful framework to evaluate these findings. First, focusing on the distance or relational predictors, funnier word pairs contain words that are orthographically and phonologically similar but semantically dissimilar. Why does greater semantic distance lead to more humor but greater orthographic and

phonological distance lead to less humor? It appears that the "type" of distance matters; specifically, it is the evaluation of the distance relative to one's expectations that is key. Given our prior experiences with language, word pairs that contain semantic leaps (such as "knapsack rapist"), as well as word pairs that are phonological tongue twisters (such as "moose ooze"), are surprising and (benignly) violate our own experience with language and multi-word phrases.

Second, the observation that semantic measures matter more when we scale our investigations of humor to multi-word phrases suggests that our expectations can flexibly shift or at least be made more or less salient depending on the context. Here context refers to whether participants are providing humor ratings to a single word or to a pair of words. In a two-word context, fluent readers reflexively attempt to construct meaning from the two words, and likely less so in a single-word context. This is supported by single-word psycholinguistic investigations that find that semantic variables are less crucial predictors of performance in single-word recognition tasks than in tasks that involve a pair of concepts/categories as in semantic categorization or classification (Goh et al., 2016; Yap et al., 2011), as well as research into how people process the meaningfulness of known and novel compounds (Günther & Marelli, 2016; 2020). This may suggest that in a two-word context, a person may hold stronger expectations about the semantics of the compound than in the single-word context such that semantic variables play a more important role in the violation of such expectations in compound-word humor than in single-word humor.

Before moving on, we wish to briefly highlight similarities and differences with a recently published paper that also looked at the humor of word pairs (Westbury & Hollis, 2021). Westbury and Hollis used best-worst scaling to measure the humor of adjective-noun pairs generated from a more focused set of funny words and examined the influence of lexical and

semantic properties that were derived computationally rather than human-generated, on humor. In the present paper, compounds were created in a highly unconstrained manner from a very large set of words, and best-worst scaling approach was used in Study 2 to validate the variables that were predictive of humor. The semantic variables that we used as predictors were obtained from large-scale norming studies (Warriner et al., 2013; Brysbaert et al., 2014). Despite these differences in approaches, our main finding does converge with that of Westbury and Hollis—word pairs containing individual words whose semantic relationship is more distant tend to be funnier word pairs.

Refining contrast theories of humor

As mentioned in the introduction, there are various classes of humor theories on the market (i.e., superiority theory, relief theory, and incongruity or contrast theory). Our results can refine and extend theories that focus on the violation of expectations as the mechanism for humor (Hurley et al., 2011; McGraw & Warren, 2010). In these theories, the core idea is that humor occurs when the stimuli violates our expectations in some way while not being too threatening. Based on this, we would expect that compounds containing words that are semantically distant, and hence surprising, would be funnier. While this was indeed what was found, additional explorations of other semantic variables inspired from the compound word literature suggest that this theory may be too simple.

Specifically, greater similarity of the first constituent (i.e., the modifier) to the entire compound was associated with greater humor. A somewhat analogous finding was also reported by Westbury and Hollis (2021) who observed that word pairs were funnier if the shared semantic neighbors of both words were dissimilar, but also if those shared semantic neighbors were *closer*

to the noun in the semantic space (the opposite relation was true for the adjective). Taken together, these results suggest the following ingredients of compound-word humor. First, containing semantically dissimilar constituents could be a prerequisite for humor as it leads to the *initial detection of the violation*. Second, a compound is likely perceived as funny if an indirect but meaningful relation also exists between constituents and between constituents and the entire expression. For instance, Westbury and Hollis (2021) observe a particular form of unexpectedness in their results where "distant neighbors of the adjective become *unexpectedly relevant* when the noun brings them into focus" (p. 14; our emphasis). In other words, another contributor of compound-word humor may involve *unexpectedly making sense of the violation*. Violations are commonplace, but ultimately the crux lies in understanding the conditions in which violations *become* funny. Going forward, leveraging on models of conceptual integration and blending (Coulson, 2001; Fauconnier & Turner, 1998) that have been influential in understanding higher order semantic processing, such as compound word processing, metaphorical and analogical processing (Gagné et al., 2010; Gentner & Markman, 1997) could help us understand the conditions in which violations become funny.

Expanding opportunities for humor

Shared humor serves a variety of functions, most prominently by uniting people around shared values and norms. Obviously this does not apply to cases where an individual *laughs at* another person, but even for superiority based theories of humor—such as Hobbes (1840)'s notion of "sudden glory" over another—the ones (or one) doing the laughing are presumably enjoying some appreciation of a sudden, or unexpected, opportunity for contrast. The results we present here demonstrate that as the context for humor expands (from one word to two) the

opportunities for contrast expand as well. Moreover, the opportunities for contrast do not only expand in the sense of global contrast, whereby a violation is made with respect to the large-scale context of all the other things an individual might experience. Even after controlling for global contrast, our results suggest an additional effect of local contrast. Of all ways that one can violate the general set of expectations set up by our day-to-day experiences, global violations that also violate themselves locally are funniest.

Amongst the most humorous word pairs we find “nymph piss,” “gravy orgy,” “moose ooze,” “crab ghetto,” “gangster pasta,” “streetcar glaze,” “knapsack rapist”, and “hippy whip.” Amongst the least humorous we find “sell bargain,” “roof darkness,” “large small,” and “fatigue daily”. This list (see also Table 2) suggests a number of potential areas for future research that move beyond our initial results. For example, *rapist* is one of lowest valence words in the English language (Warriner et al., 2013), it is also extremely unfunny (Engelthaler & Hills, 2018). However, in line with relief-based theories of humor (Freud, 1928; Spencer, 1860), combining a non-humorous word (*rapist*) with an unexpected neighbor (*knapsack*) can defuse a low-valence unfunny word and lead to something amusing. Compound-word humor allows for a closer examination of this effect by allowing us to examine exactly what kinds of words provide a defusing contrast. A second observation is that concreteness tends to be consistently predictive of compound-word humor. This may be because the capacity to visually see one concept (a *nymph*) creates a greater sense of violation when a second ‘visible’ concept appears in the same context (*piss*).

Table 2. Top 10 Most and Least Humorous word pairs from Study 2.

Least humorous		Most humorous	
sell bargain	0.288	polka hooker	0.765
conserve health	0.289	playboy parrot	0.755
power influence	0.291	penis weasel	0.745
will stay	0.298	turnip tramp	0.714
schedule year	0.303	funk fungus	0.714
insult nickname	0.322	spam scrotum	0.709
life friend	0.323	gnome bone	0.697
trouble mention	0.324	stripper hippo	0.694
workman call	0.326	rowdy bowels	0.693
large small	0.327	pansy panties	0.693

The crowdsourced compound-word humor ratings provide us with a starting point to test these ideas. We conducted a post-hoc exploratory analysis using Study 1's data by including the following interaction terms into the full model: (i) *valence x semantic similarity* to see if semantic leaps (i.e., greater semantic distance) in compounds provided a "defusing" contrast for low-valenced words and (ii) *concreteness x semantic similarity* to see if semantic distance enhanced the humorous-ness of concrete concepts. The analyses provide some support for these ideas. In the valence x semantic similarity interaction, the effect of valence on humor was non-significant for semantically dissimilar word pairs and negative for semantically similar word pairs, whereas in the concreteness x semantic similarity interaction, the effect of concreteness on humor was enhanced by semantic dissimilarity. Specific details of this post-hoc analysis along with a visual depiction of the interaction effects can be found in the Appendix and in Figure 3.

Although empirical studies are still needed to validate these exploratory findings, these patterns are intriguing as they suggest that local and global contrast can interact in interesting ways to produce relational humor.

Limitations and Future Directions

Before concluding we wish to highlight a couple of limitations in our approach. First, our lexical-semantic predictors were derived from human-generated norms collected by other researchers (e.g., Warriner et al., 2013) and in particular we used extrapolated concreteness norms to deal with missingness (Hollis et al., 2017). Westbury (2016) points out that it may not be meaningful to use a set of human generated data (i.e., humor or semantic ratings by people) to predict another set of human generated data (i.e., humor ratings by other people) because one is merely correlating two unknowns without an explicit understanding of the cognitive mechanisms that produced the data. On the other hand, Snefjella and Blank (2020) point out potential limitations in semantic norm extrapolation that aims to derive lexical-semantic norms for lexical items through purely computational means (i.e., without human input). It is clear that there are immense methodological and theoretical challenges involved in the investigation of cognitive and linguistic processes, and hence any reader should consider the implications of the present paper with these challenges and limitations in mind.

Nevertheless, the present work sets the stage for an obvious extension, which is well-known in comedy writing: the rule of three. In the rule of three, one sets up the context and expectation with the first two items, and then violates them by choosing a third item that is the humorous punch line. For example, “when you die there’s a light at the end of the tunnel. When my father dies, he’ll [1] see the light, [2] make his way toward it, and then [3] flip it off to save

electricity (Harland Williams)" (as quoted in Brown, 2005). Humor writers (e.g., Vorhaus, 1994) suggest that the first two items establish a trend, which can then be properly violated by the third item. This is an example of the local context effect of humor we demonstrate here, for which tri-grams offer a practical and ecologically valid comedic context. Perhaps, the best comedy writers are the ones who are acutely sensitive to language priors (i.e., the global context) and also acquire the skills to set up a context with local contrast—exploiting and integrating these two sources of information to create multiple pathways to humor.

Acknowledgements

C. S. Q. S was supported by the Overseas Postdoctoral Fellowship from the National University of Singapore. T.T.H. was supported by the Royal Society Wolfson Research Merit Award (WM160074) and the Alan Turing Institute.

Footnotes

1. A reviewer (Fritz Günther) noted that a potential point of concern with the current analysis was that our measures were collected from a variety of sources based on different language corpora, which will have different underlying distributional properties. To assess if our results might be an artifact of this we re-ran our analyses with word and letter frequencies obtained from the same source corpus (Günther & Marelli, 2018) and found that the pattern of results did not change. Another analysis that was conducted was to include orthotactic and phonotactic frequencies (i.e., probabilities of pairs of letters and sounds of words in the language) as additional predictors and again we found that it did not change the overall pattern of findings. We thank Fritz Günther for generously making their corpus measures available to us. These supplementary analyses can be found in the OSF page for this paper.

2. We thank the following reviewers, Fritz Günther and Chris Westbury, for suggesting the following analyses in their reviews. We also wish to note that we have also explored the influence of “taboo-ness” ratings (Reilly et al., 2020) and syntactic class structure on compound-word humor. Our overall result (i.e., evidence for both local and global contrast effects on compound-word humor) persisted. These supplementary analyses can be found in the OSF page for this paper.

Authors' note

All data and scripts can be found on the Open Science Framework: https://osf.io/wy98d

Shiny application and source code: https://csqsiew.shinyapps.io/humorous_phrases/ and https://github.com/csqsiew/shinyhumor

Pre-registration for Study 2 can be found at: https://osf.io/b8ftw

References

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

Blumeyer, D. (2012, November 11). Relative Frequencies of English Phonemes. *Cmloegcmluin*. https://cmloegcmluin.wordpress.com/2012/11/10/relative-frequencies-of-english-phonemes/

Brown, J. (2005). *The Comedy Thesaurus: 3,241 Quips, Quotes, and Smartass Remarks*. Quirk Books.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (2020). *shiny: Web Application Framework for R*. https://CRAN.R-project.org/package=shiny

Coulson, S. (2001). *Semantic Leaps: Frame-Shifting and Conceptual Blending in Meaning Construction*. Cambridge University Press.

Downing, P. (1977). On the creation and use of English compound nouns. *Language*, 810–842.

Engelthaler, T., & Hills, T. T. (2018). Humor norms for 4,997 English words. *Behavior Research Methods*, *50*(3), 1116–1124.

Falkauskas, K., & Kuperman, V. (2015). When experience meets language statistics: Individual variability in processing English compound words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1607.

Fauconnier, G., & Turner, M. (1998). Conceptual integration networks. *Cognitive Science*, *22*(2), 133–187. https://doi.org/10.1016/S0364-0213(99)80038-X

Freud, S. (1928). Original papers: Humor. *International Journal of Psychoanalysis*, *9*, 1–6.

Gagné, C. L. (2001). Relation and lexical priming during the interpretation of noun–noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(1), 236.

Gagné, C. L., Marchak, K. A., & Spalding, T. L. (2010). Meaning predictability and compound interpretation: A psycholinguistic investigation. *Word Structure*, *3*(2), 234–251. https://doi.org/10.3366/word.2010.0006

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, *52*(1), 45–56. https://doi.org/10.1037/0003-066X.52.1.45

Godkewitsch, M. (1974). Correlates of humor: Verbal and nonverbal aesthetic reactions as functions of semantic distance within adjective-noun pairs. *Studies in the New Experimental Aesthetics*, 279–304.

Goh, W. D., Yap, M. J., Lau, M. C., Ng, M. M. R., & Tan, L.-C. (2016). Semantic Richness Effects in Spoken Word Recognition: A Lexical Decision and Semantic Categorization Megastudy. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00976

Günther, F., & Marelli, M. (2016). Understanding Karma Police: The perceived plausibility of noun compounds as predicted by distributional models of semantic representation. *PloS One*, *11*(10), e0163200.

Günther, F., & Marelli, M. (2020). Trying to make it work: Compositional effects in the processing of compound “nonwords.” *Quarterly Journal of Experimental Psychology*, *73*(7), 1082–1091.

Hollis, G. (2018). Scoring best-worst data in unbalanced many-item designs, with applications to crowdsourcing semantic judgments. *Behavior Research Methods*, *50*(2), 711–729.

Hollis, G., Westbury, C., & Lefsrud, L. (2017). Extrapolating human judgments from skip-gram vector representations of word meaning. *The Quarterly Journal of Experimental Psychology*, *70*(8), 1603–1619.

Hurley, M. M., Dennett, D. C., Adams Jr, R. B., & Adams, R. B. (2011). *Inside jokes: Using humor to reverse-engineer the mind*. MIT press.

Kang, B. (2016). Metaphor and its humorousness: The case of nominal compounds in German. *Humor*, *29*(3), 359–380.

Kant, I. (1914). *Kant’s Critique of Judgement (revised)(JH Bernard, Trans.)*. London, UK: Macmillan.(Original work published in German 1790).

Li, Y., Engelthaler, T., Siew, C. S., & Hills, T. T. (2019). The Macroscope: A tool for examining the historical structure of language. *Behavior Research Methods*, 1–14.

Libben, G. (2014). The nature of compounds: A psychocentric perspective. *Cognitive Neuropsychology*, *31*(1–2), 8–25.

Louviere, J. J. and G. Woodworth (1990) Best-worst scaling: A model for largest difference judgments. Working Paper, Faculty of Business, University of Alberta.

Louviere, J. J., Flynn, T. N., & Marley, A. A. J. (2015). Best-worst scaling: Theory, methods and applications. Cambridge, England: Cambridge University Press.

Lyons, J. (n.d.). *English letter frequencies*. Retrieved July 14, 2021, from http://www.practicalcryptography.com/cryptanalysis/letter-frequencies-various-languages/english-letter-frequencies/

Marley, A. A. J., & Islam, T. (2012). Conceptual relations between expanded rank data and models of the unexpanded rank data. Journal of Choice Modelling, 5, 38-80.

McGraw, A. P., & Warren, C. (2010). Benign violations: Making immoral behavior funny. *Psychological Science*, *21*(8), 1141–1149.

Minsky, M. (1981). Jokes and their Relation to the Cognitive Unconscious. *Cognitive Constraints on Communication*, 175–200.

Molesworth, W. (1840). *The english works of Thomas Hobbes*.

Richman, D. (n.d.). Why visiting Auschwitz doesn't have to be a harrowing experience. *March of the Living UK*. Retrieved July 14, 2021, from https://www.marchoftheliving.org.uk

Roeckelein, J. E. (2006). *Elsevier's dictionary of psychological theories*. Elsevier.

Siew, C. S. Q., Hills, T., & Engelthaler, T. (2021, September 9). Humorous Phrases Generator: Relational Humor. Retrieved from osf.io/wy98d

Snefjella, B., & Blank, I. (2020). *Semantic Norm Extrapolation is a Missing Data Problem*. PsyArXiv. https://doi.org/10.31234/osf.io/y2gav

Spencer, H. (1860). *The physiology of laughter*. [Macmillan].

Suárez, L., Tan, S. H., Yap, M. J., & Goh, W. D. (2011). Observing neighborhood effects without neighbors. *Psychonomic Bulletin & Review*, *18*(3), 605–611.

Vorhaus, J. (1994). *The comic toolbox: How to be funny even if you're not*. Silman-James Press.

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

Westbury, C. (2016). Pay no attention to that man behind the curtain: Explaining semantics without semantics. *The Mental Lexicon*, *11*(3), 350–374.

Westbury, C., & Hollis, G. (2019). Wriggly, squiffy, lummox, and boobs: What makes some words funny? *Journal of Experimental Psychology: General*, *148*(1), 97.

Westbury, C., & Hollis, G. (2021). A pompous snack: On the unreasonable complexity of the world's third-worst jokes. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/cep0000234

Westbury, C., Shaoul, C., Moroschan, G., & Ramscar, M. (2016). Telling the world's least funny jokes: On the quantification of humor as entropy. *Journal of Memory and Language*, *86*, 141–156.

Yap, M. J., Tan, S. E., Pexman, P. M., & Hargreaves, I. S. (2011). Is more always better? Effects of semantic richness on lexical decision, speeded pronunciation, and semantic classification. *Psychonomic Bulletin & Review*, *18*(4), 742–750.

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

Appendix

Table A1. Descriptive statistics and correlations among the word-level predictors for words used to generate random word pairs in Study 1.

	M	SD
No. of letters	5.81	1.63
No. of phonemes	4.70	1.36
Letter frequency	0.06	0.01
Phoneme frequency	0.04	0.01
Word frequency	7.76	1.92
Humor	2.41	0.44
Valence	5.16	1.19
Arousal	4.10	0.89
Dominance	5.23	0.86
Concreteness	0.72	0.15

	No. of letters	No. of phonemes	Letter frequency	Phoneme frequency	Word frequency	Humor	Valence	Arousal	Dominance	Concreteness
No. of letters	1.00	0.84	0.16	0.19	-0.34	-0.07	0.03	0.06	0.01	-0.03
No. of phonemes	0.84	1.00	0.07	0.31	-0.28	-0.08	0.01	0.06	-0.02	-0.06
Letter frequency	0.16	0.07	1.00	0.45	0.07	-0.22	0.03	-0.04	0.06	-0.06
Phoneme frequency	0.19	0.31	0.45	1.00	0.02	-0.14	-0.01	0.01	0.01	-0.04
Word frequency	-0.34	-0.28	0.07	0.02	1.00	-0.38	0.20	0.01	0.20	-0.19
Humor	-0.07	-0.08	-0.22	-0.14	-0.38	1.00	0.10	0.04	0.02	0.11
Valence	0.03	0.01	0.03	-0.01	0.20	0.10	1.00	-0.20	0.66	0.10
Arousal	0.06	0.06	-0.04	0.01	0.01	0.04	-0.20	1.00	-0.18	-0.17
Dominance	0.01	-0.02	0.06	0.01	0.20	0.02	0.66	-0.18	1.00	0.05
Concreteness	-0.03	-0.06	-0.06	-0.04	-0.19	0.11	0.10	-0.17	0.05	1.00

Table A2. Descriptive statistics and correlation table for the word-level and word-pair predictors for 732 word pairs in Study 2.

	M	SD
No. of letters	5.75	0.97
No. of phonemes	4.67	0.89
Letter frequency	0.06	0.01
Phoneme frequency	0.04	0.01
Word frequency	7.82	1.34
Humor	2.55	0.41
Valence	5.12	0.86
Arousal	4.16	0.65
Dominance	5.19	0.59
Concreteness	0.74	0.11
Orthographic distance	4.75	1.76
Phonological distance	4.07	1.55
Semantic similarity	0.14	0.13

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
No. of letters	1.00	0.79	0.22	0.25	-0.20	-0.12	0.05	0.06	-0.01	-0.04	0.47	0.41	0.01
No. of phonemes	0.79	1.00	0.11	0.37	-0.17	-0.11	0.04	0.02	-0.02	-0.09	0.39	0.51	-0.03
Letter frequency	0.22	0.11	1.00	0.44	0.17	-0.34	0.05	-0.06	0.11	-0.15	0.21	0.25	0.11
Phoneme frequency	0.25	0.37	0.44	1.00	0.03	-0.16	0.01	0.03	0.04	-0.09	0.10	0.17	0.03
Word frequency	-0.20	-0.17	0.17	0.03	1.00	-0.50	0.25	-0.02	0.35	-0.42	0.24	0.26	0.33
Humor	-0.12	-0.11	-0.34	-0.16	-0.50	1.00	-0.01	0.15	-0.16	0.40	-0.52	-0.52	-0.36
Valence	0.05	0.04	0.05	0.01	0.25	-0.01	1.00	-0.21	0.65	0.03	0.09	0.10	0.04
Arousal	0.06	0.02	-0.06	0.03	-0.02	0.15	-0.21	1.00	-0.20	-0.11	-0.05	-0.08	-0.08
Dominance	-0.01	-0.02	0.11	0.04	0.35	-0.16	0.65	-0.20	1.00	-0.12	0.18	0.17	0.11
Concreteness	-0.04	-0.09	-0.15	-0.09	-0.42	0.40	0.03	-0.11	-0.12	1.00	-0.26	-0.29	-0.34
Orthographic distance	0.47	0.39	0.21	0.10	0.24	-0.52	0.09	-0.05	0.18	-0.26	1.00	0.82	0.19
Phonological distance	0.41	0.51	0.25	0.17	0.26	-0.52	0.10	-0.08	0.17	-0.29	0.82	1.00	0.15
Semantic similarity	0.01	-0.03	0.11	0.03	0.33	-0.36	0.04	-0.08	0.11	-0.34	0.19	0.15	1.00

Table A3. Regression model with all predictors from Study 1 combined with Average-CDV predictors (1: CDV) in column 1 and predictors of constituent-compound similarity (2: Compound) in column 2. Predictors discussed in the paper are in bold and the standard errors are in parentheses.

	Dependent variable: Humor Rating	
	(1: CDV)	(2: Compound)
humor1	0.219*** (0.015)	0.295*** (0.013)
humor2	0.183*** (0.015)	0.253*** (0.013)
valence1	0.005 (0.018)	-0.005 (0.017)
valence2	-0.033 (0.018)	-0.043** (0.017)
arousal1	0.014 (0.014)	0.053*** (0.012)
arousal2	0.021 (0.014)	0.051*** (0.012)
dominance1	-0.053** (0.018)	-0.067*** (0.016)
dominance2	-0.036* (0.018)	-0.052** (0.016)
concreteness1	0.069*** (0.014)	0.102*** (0.013)
concreteness2	0.106*** (0.014)	0.127*** (0.013)
letters1	0.071** (0.027)	0.049 (0.026)
letters2	0.056* (0.027)	0.026 (0.026)
phonemes1	0.055 (0.028)	0.036 (0.028)
phonemes2	0.042 (0.028)	0.036 (0.028)
letter freq1	-0.041** (0.015)	-0.052*** (0.014)
letter freq2	-0.029 (0.015)	-0.037** (0.014)
phoneme freq1	-0.023 (0.016)	-0.006 (0.015)
phoneme freq2	0.016 (0.016)	0.021 (0.015)
frequency1	-0.023 (0.016)	-0.047** (0.015)
frequency2	-0.046** (0.016)	-0.067*** (0.015)
orthographic distance	-0.086*** (0.023)	-0.082*** (0.022)
phonological distance	-0.083*** (0.024)	-0.064** (0.023)
similarity (w1-w2)	-0.122*** (0.014)	

cdv1	**-0.186*** (0.015)**	
cdv2	**-0.172*** (0.015)**	
similarity (w1-w2)		-0.062*** (0.014)
similarity (w1-comp)		**0.033* (0.014)**
similarity (w2-comp)		**0.004 (0.014)**
Constant	-1.357*** (0.013)	-1.329*** (0.012)
Observations	37,146	43,059
Log Likelihood	-18,532.830	-21,804.150
Akaike Inf. Crit.	37,117.660	43,660.310

Note: *p<0.05; **p<0.01; ***p<0.001

Table A4. Regression model with all predictors from Study 2 combined with Average-CDV predictors (1: CDV) in column 1 and predictors of constituent-compound similarity (2: Compound) in column 2. Predictors discussed in the paper are in bold and the standard errors are in parentheses.

	Dependent variable: Humor Rating	
	(1: CDV)	(2: Compound)
humor1	0.036*** (0.007)	0.047*** (0.006)
humor2	0.038*** (0.007)	0.054*** (0.006)
valence1	-0.003 (0.002)	-0.002 (0.002)
valence2	-0.002 (0.002)	-0.004 (0.002)
arousal1	0.002 (0.002)	0.005* (0.002)
arousal2	0.005* (0.002)	0.007** (0.002)
dominance1	-0.003 (0.003)	-0.004 (0.003)
dominance2	-0.003 (0.003)	-0.005 (0.003)
conc1	0.061*** (0.015)	0.075*** (0.015)
conc2	0.014 (0.016)	0.032 (0.017)
olen1	-0.001 (0.003)	-0.001 (0.003)
olen2	0.009** (0.003)	0.008** (0.003)
plen1	0.011*** (0.003)	0.009** (0.003)
plen2	-0.002 (0.003)	-0.003 (0.003)
lgletterfreq1	0.101 (0.197)	-0.002 (0.203)
lgletterfreq2	0.129 (0.207)	0.219 (0.214)
lgphonfreq1	-0.730*** (0.207)	-0.636** (0.215)
lgphonfreq2	-0.001 (0.207)	-0.072 (0.214)
freq1	-0.003* (0.001)	-0.003* (0.001)

freq2	-0.003* (0.001)	-0.003* (0.001)
odist	-0.003 (0.002)	-0.004 (0.002)
pdist	-0.012*** (0.003)	-0.010*** (0.003)
similarity (w1-w2)	-0.048** (0.017)	
cdv1	**-0.105*** (0.022)**	
cdv2	**-0.118*** (0.024)**	
similarity (w1-w2)		-0.040 (0.031)
similarity (w1-comp)		**0.014 (0.031)**
similarity (w2-comp)		**0.012 (0.029)**
Constant	0.545*** (0.060)	0.238*** (0.044)
Observations	732	732
R^2	0.674	0.651
Adjusted R^2	0.662	0.639
Residual Std. Error (df = 706)	0.050	0.052
F Statistic (df = 25; 706)	58.363***	52.787***
Note:		*p<0.05; **p<0.01; ***p<0.001

Exploratory analyses of humor ratings from Study 1.

We conducted a post-hoc, exploratory analysis using Study 1's data by including the following interaction terms into the full model: (i) *valence x semantic similarity* to see if semantic leaps (i.e., greater semantic distance) in word pairs provided a "defusing" contrast for low-valenced words and (ii) *concreteness x semantic similarity* to see if semantic distance enhanced the humorous-ness of concrete concepts. The full model contained all the predictors that were previously described in Study 1. The models with each of the interaction terms were then submitted to a stepwise forward and backward search procedure (by eliminating and adding 1 variable at the time) that aimed to minimize AIC by only including the optimal set of predictors in the final model. In both cases, the interaction term was retained.

Table A5. Final logistic regression models from the stepwise search. Panel (a) shows the model with the valence x semantic similarity interaction effect. Panel (b) shows the model with the concreteness x semantic similarity interaction effect.

(a)				(b)			
Predictors	*Odds Ratios*	*CI*	*p*	*Predictors*	*Odds Ratios*	*CI*	*p*
Humor	1.47	1.44 – 1.51	<0.001	Humor	1.47	1.44 – 1.51	<0.001
Valence	0.97	0.93 – 1.00	0.035	Valence	0.97	0.93 – 1.00	0.034
Arousal	1.08	1.05 – 1.10	<0.001	Arousal	1.08	1.05 – 1.10	<0.001
Dominance	0.92	0.89 – 0.95	<0.001	Dominance	0.92	0.89 – 0.95	<0.001
Concreteness	1.18	1.15 – 1.21	<0.001	Concreteness	1.18	1.15 – 1.21	<0.001
No. of letters	1.06	1.00 – 1.12	0.044	No. of letters	1.06	1.00 – 1.12	0.044
No. of phonemes	1.05	0.99 – 1.11	0.077	No. of phonemes	1.05	0.99 – 1.11	0.081
Letter frequency	0.94	0.92 – 0.97	<0.001	Letter frequency	0.94	0.92 – 0.96	<0.001
Word frequency	0.93	0.90 – 0.95	<0.001	Word frequency	0.93	0.90 – 0.95	<0.001
Orthographic distance	0.92	0.88 – 0.96	<0.001	Orthographic distance	0.92	0.88 – 0.96	<0.001
Phonological distance	0.93	0.89 – 0.97	0.002	Phonological distance	0.93	0.89 – 0.98	0.002

Semantic similarity	0.9	0.88 – 0.92	<0.001	Semantic similarity	0.9	0.88 – 0.93	<0.001
Valence x Semantic similarity	**0.97**	**0.95 – 0.99**	**0.013**	**Concreteness x Semantic similarity**	**0.98**	**0.96 – 1.00**	**0.09**

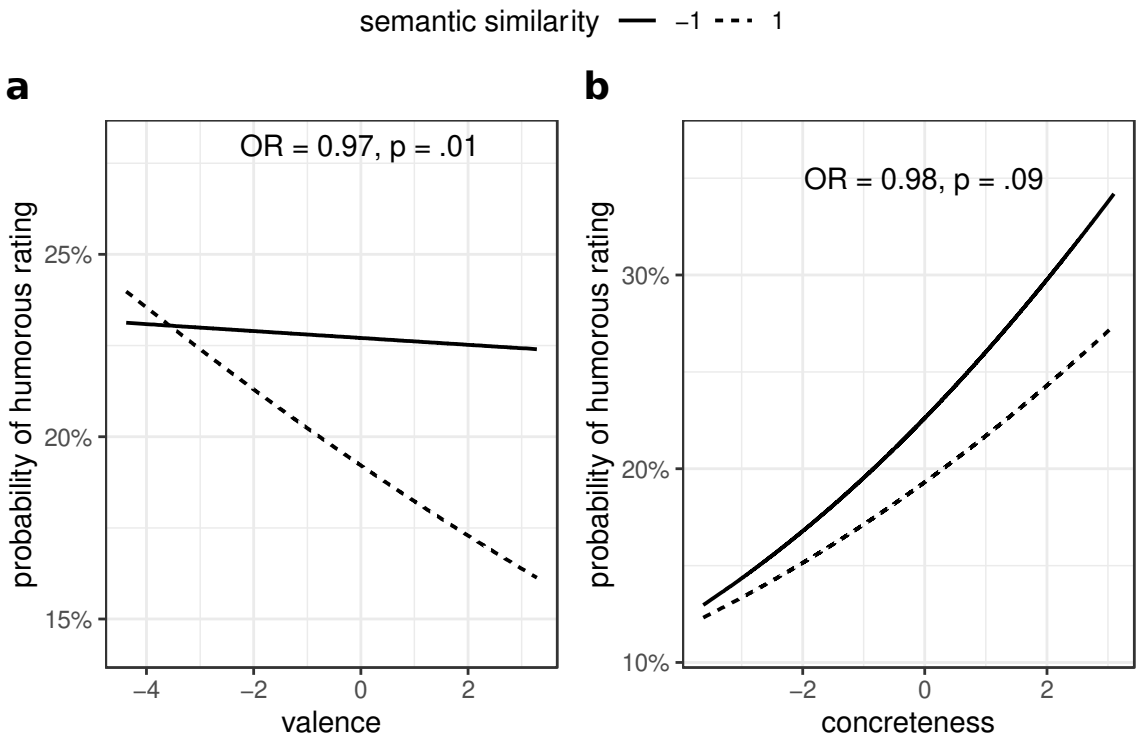

Figure 3. Plots showing the pattern of the interaction effects. Panel a: interaction between valence and semantic similarity. Panel b: interaction between concreteness and semantic similarity. Solid lines indicate semantic similarity less than 1 SD below the mean (i.e., semantically dissimilar); dotted lines indicate semantic similarity more than 1 SD above the mean (i.e., semantically similar).