American Psychological Association

Journal of Experimental Psychology: General

0096-3445/19/$12.00

2019, Vol. 148, No. 1, 97–123
http://dx.doi.org/10.1037/xge0000467

Wriggly, Squiffy, Lummox, and Boobs: What Makes Some Words Funny?

Chris Westbury and Geoff Hollis
University of Alberta

Theories of humor tend to be post hoc descriptions, suffering from insufficient operationalization and a subsequent inability to make predictions about what will be found humorous and to what extent. Here we build on the Engelthaler & Hills' (2017) humor rating norms for 4,997 words, by analyzing the semantic, phonological, orthographic, and frequency factors that play a role in the judgments. We were able to predict the original humor rating norms and ratings for previously unrated words with greater reliability than the split half reliability in the original norms, as estimated from splitting those norms along gender or age lines. Our findings are consistent with several theories of humor, while suggesting that those theories are too narrow. In particular, they are consistent with incongruity theory, which suggests that experienced humor is proportional to the degree to which expectations are violated. We demonstrate that words are judged funnier if they are less common and have an improbable orthographic or phonological structure. We also describe and quantify the semantic attributes of words that are judged funny and show that they are partly compatible with the superiority theory of humor, which focuses on humor as scorn. Several other specific semantic attributes are also associated with humor.

Keywords: semantics, humor, computational model, word2vec

> One nice thing about making jokes is that you don't have to prove them.
>
> —P. J. O'Rourke, *Atlantic Unbound* (August 8, 2002)

Humor is very complex. It is difficult to find any elements in common between a silent movie slapstick comedy routine and a dirty limerick, though we might find both humorous. This is perhaps because humor is a multidimensional construct, with some aspects being more heavily weighted in some manifestations than in others. In this paper, we demonstrate and analyze the complexity underlying an extremely simple form of humor, which is the funniness[1] of single words.

Investigators have pondered the elements of humor for centuries. They have offered a variety of "theories" to explain it, focusing (as Eysenck, 1942 noted) on different aspects such as the cognitive, the affective, and the conative (action directed). Following Keith-Spiegel (1972), we use the scare quotes here to emphasize that these theories are weak. In his review of humor studies, Keith-Spiegel (1972) wrote that:

> . . . the term 'theory' is used to refer to the notions writers have put forth, but this designation is for convenience only and is not to be taken strictly. Many statements are actually descriptions under which humor is experienced rather than attempts to explain humor. Furthermore, many statements involve assumptions or concepts that defy operationalizing thereby precluding empirical testing. Sometimes we find that the explanations offered leave us perched atop 'a black box' (e.g., humor as an instinct). Still others are speculations on the functions humor and laughter perform for the individual or for the group but remain incomplete or unsatisfactory as adequate theory. (p. 5)

Almost no ideas have yet been put forward about how to put theories of humor on a firmer footing by using quantified formal methods to predict and confirm what will be found humorous, and to what degree. The work we outline here is a step in that direction.

For the purposes of this paper, two popular and mutually compatible ideas about humor are of particular relevance: theories of humor as denigration (*superiority theory*) and theories of humor as the improbable intersection of two frames of reference (*incongruity theory*).

The earliest discussion of the nature of humor is in the Platonic dialog *Philebus*, in which Plato focuses on what is now known as the superiority theory of humor. This theory focuses on humor as the scorn we feel for people who make themselves ridiculous through their own ignorance, especially people who take themselves to be something other than what they are. Consider United States President Donald Trump's comments in a May 2017 interview on the origin of the well-known metaphorical extension of a literally meaningful phrase, "priming the pump": "Have you heard that expression used before? Because I haven't heard it. I mean, I

This article was published Online First October 18, 2018.

Chris Westbury and Geoff Hollis, Department of Psychology, University of Alberta.

This work was made possible by a grant from the Canadian Natural Sciences and Engineering Research Council to Westbury.

Correspondence concerning this article should be addressed to Chris Westbury, Department of Psychology, University of Alberta, P220 Biological Sciences Building, Edmonton, AB, Canada, T6G 2E9. E-mail: chrisw@ualberta.ca

[1] In this article, we often use the more colloquial term *funny* (and its morphological variants, most notably *funniness*) as an exact synonym of the less ambiguous but more ponderous word *humor* (and its morphological variants). Our use of *funniness* should not be taken to have any connotations of weirdness (as in "What's that funny smell?") but only of humor.

just . . . I came up with it a couple of days ago and I thought it was good" (The Economist, 2017).

We had to laugh at Trump when he said this, because we could see what he does not: that he is both ignorant of a commonly used expression and vain enough to believe that he must have invented it. Plato would disapprove of our laughter since he believed that such humor is morally reprehensible. In the dialog *Laws* (Plato, trans. 2008), he went so far as to have Socrates suggest that an ideal state would not allow this kind of humor at all:

> A comic poet, or maker of iambic or satirical lyric verse, shall not be permitted to ridicule any of the citizens, either by word or likeness, either in anger or without anger. And if any one [sic] is disobedient, the judges shall either at once expel him from the country, or he shall pay a fine of three minae, which shall be dedicated to the God who presides over the contests. (lines 21,057–21,062; https://www.gutenberg.org/cache/epub/1750/pg1750.txt)

Plato's student Aristotle also emphasized humor as denigration. In his *Ethics* (c. 350BCE/2005), he agreed with Plato about outlawing some forms of humor: "because jesting is a species of scurrility and there are some points of scurrility forbidden by law; it may be certain points of jesting should have been also so forbidden" (lines 4338–4340; https://www.gutenberg.org/cache/epub/8438/pg8438.txt).

Another way of thinking about our laughter at Trump's inane comment is to analyze it as a clash of two contrasting frames of reference. The humor in the situation arises from the incompatibility of his frame of reference ("I just invented the excellent expression 'priming the pump!'") and our own ("You did not, you silly galoot!"). This idea that humor can be analyzed as a clash of frames of reference is known as the incongruity theory of humor. It also has ancient roots. Cicero (trans. 1875) wrote "the most common kind of joke [is when we] expect one thing and another is said; in which case our own disappointed expectation makes us laugh" (line 7689; https://archive.org/stream/ciceroonoratorya00ciceuoft/ciceroonoratorya00ciceuoft_djvu.txt). Later theorists generalized and sharpened this notion. Writing in 1818, Schopenhauer (1818/1907) stated it very clearly:

> The cause of laughter in every case is simply the sudden perception of the incongruity between a concept and the real objects which have been thought through it in some relation, and laughter itself is just the expression of this incongruity. (Book I, sec. 13)

This captures the idea that the violated expectation need not be verbal. We might laugh if our spouse gave us a potato inside an iPhone box for our birthday, due to the clash between the expectation set up by the box and the reality revealed by opening it. Incongruity theory is today, by far, the most dominant theory of humor, expressed in the writings of Francis Hutcheson (1750), Immanuel Kant (1790/1951), Søren Kierkegaard (1846/1941), William Hazlitt (1819/1907), Arthur Koestler (1964), Daniel Berlyne (1971), and Vilayanur Ramachandran (1998), among many others. Many other quasitheoretical notions about humor are closely related to incongruity theory, though they try to distinguish themselves from it. Keith-Spiegel (1972) mentions a dozen scholars who have put forward the idea of humor as surprise, a half dozen who have put forward the idea of humor as ambivalence, and a few who have put forward the idea of a configurational theory of humor (in which "elements originally perceived as unrelated suddenly fall into place" [p. 11]). All of these notions share the key idea that humor is related to uncertainty and its resolution. They are distinguished only by the examples chosen to illustrate that resolution.

Note that incongruity theory and its relations are *post hoc* theories, in the sense that they explain, after the fact, why some things are funny but are unable to predict in advance what will be funny. Incongruity in itself is not sufficient for humor. Most violations of expectation are not funny at all: being pulled over for a speeding ticket, discovering that the milk is sour after pouring it into your morning coffee, receiving a novel e-mail spam, remembering that you didn't brush your teeth once you are in bed.

We end this brief historical overview with a consideration of a special case of incongruity theory that was emphasized by Henri Bergson (1900/2009) in his book, *Laughter*. Bergson (1900/2009) focused specifically on the incongruity that is set up by the juxtaposition of "the illusion of life and the distinct impression of a mechanical arrangement" (line 1361–1363; text is capitalized in the original). We see this juxtaposition theory used for comedic effect in the classic routine of a person slipping on a banana peel, and in modern form in an endless variety of videos at FailBlog.com. Again, we note that although the theory seems compelling for explaining many specific cases of humor after the fact, it has no quantified mechanism for making predictions about which forms of mechanical human action will be experienced as funny or to what degree.

These approaches to humor can be broken down into two broad classes, *semantic theories* and *information theoretic theories*. The superiority and juxtaposition theories are both semantic theories, since they focus on *the content* of what is funny, pointing out that particular topics are inherently funny. Incongruity theory and its many related theories are information theoretic theories, since they focus on the probabilities and structural complexity of humorous stimuli rather than on their contents. There is, of course, no incompatibility between semantic and information theoretic theories, since a thing can be simultaneously unexpected, structurally complex, and belong to a semantic category that is commonly associated (or not) with humor. Public coughing fits are perhaps as improbable as public farting, but children often find it funny when someone farts in public, but not when someone coughs, because, while both are unexpected, the fart belongs to the inherently funny category of excretory functions.[2] In this paper, we separate and quantify the components of semantic and probability theories so we can contrast and predict the extent of their contribution to experienced humor.

Although Bergson (1900/2009) himself did not discuss it, we can see profane humor, focused on excretory and sexual functions, as a special case of his special case. Mere reference to excretion and the biological aspects of sexuality highlights the mechanical nature of our embodied existence. Scatological and sexual humor is very common in modern popular culture, and often relies on

[2] Exceptions to this requirement for inherent funniness are common, perhaps especially in highly improbable situations. For example, in Greg Kohs (2017) film *AlphaGo*, Korean Go masters narrating the computer program AlphaGo's defeat of world human Go champion Lee Sedol started giggling uncontrollably on camera, apparently with genuine humor, as the very unexpected defeat of the human champion became increasingly inevitable.

little more than alluding to these biological demands. As schoolboys of a certain age rediscover repeatedly, there is a sense in which simply uttering the word *fart* is a one-word joke.

In this paper, we examine exactly this form of minimalist humor, the humor inherent in single words. Single words are not the world's worst jokes; they are the world's second-worst jokes. The worst jokes in the world are single *nonwords*, which have been previously studied by Westbury, Shaoul, Moroschan, and Ramscar (2016). They showed that the humor ratings for meaningless nonwords such as *quarban* and *himumma* can be strongly predicted by a very simple measure, which is essentially the average frequency of the component letters (though expressed in the paper in slightly different terms). This finding is consistent with incongruity theory, while allowing, for the first time, for the principled testing of *predictions* about experienced humor. By stripping humor down to almost nothing at all and quantifiably defining what was left, Westbury et al. (2016) were able to statistically test Schopenhauer's explicitly stated theory that experienced humor was a function of the distance between expectation and reality. Here we replicate and extend this attempt to predict humor judgments in advance, while taking into account the added complications that real words have.

The main difference between words and formally plausible nonwords is, of course, that words have *meanings*. It would be impossible (or, at least, very limiting, to a degree that we will quantify below) to analyze the humor of words without taking into account semantics. Here we rely on the fact that semantics can be well modeled using statistics about word co-occurrence. The general insight of these models (operationalizing the earlier philosophical speculations in Wittgenstein, 1953) is that *patterns of word use* capture something about *word meaning*. The fact that we can use word co-occurrence to model semantics computationally was first demonstrated, in slightly different forms, by Lund & Burgess, 1996 and Landauer & Dumais, 1997, and has been demonstrated in many other models since. Words that share a common context are likely to be quantifiably similar in meaning to the extent that their context is similar. We can quantify context similarity by representing co-occurrence as a vector. In their simplest form, these vectors might simply represent a count of how often a target word occurred close to every word in the English language in a large corpus of text, where "close to" might mean within a few words or in the same document.

In practice, merely tallying co-occurrence results in very sparse vectors (vectors with many zero values) because most words do not co-occur with most other words. Since sparse vectors are informationally impoverished, researchers have explored various transformations that increase the information density in co-occurrence vectors. We use vectors from the word2vec model (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Mikolov, Yih, & Zweig, 2013). This model uses a simple three-layer neural network to predict a target word's local context (in our implementation, two words in either direction), adjusting the weights in vectors through back-propagation to minimize the predictive error. This produces very informationally rich vectors, in our implementation (though not necessarily) of length 300. Our word2vec matrix was built from the publicly released Google news corpus, which contains 3 billion words (https://github.com/mmihaltz/word2vec-GoogleNews-vectors). We applied principal components analysis to this matrix, so reported distance measures are based on the matrix with 300 components ordered in terms of how much variance they account for.

We can use the cosine distance between the vectors for two different words as a measure of the words' similarity. As a concrete example, the 10 closest neighbors of the word *humor* in our dictionary of 78,278 words are *wit, irreverence, witty, humorous, levity, satire, comedic, playfulness, whimsy*, and *hilarity*.

For developing our predictions, we rely on humor ratings collected for 4,997 words by Engelthaler and Hills (2017). The distribution of their ratings across all words is shown in Figure 1. The modal rating was 2/5 (average [*SD*]: 2.4 [0.43]) and the distribution is shifted low from a normal distribution (overlain in Figure 1) because most words are judged to be nonfunny. However, at the high end there is a tail of more words than would be expected by chance that are judged funny. Only eight words were judged to be funny at 4/5 or higher: *booty, tit, hooter, booby, moo, waddle*, and *twerp*. We note here a point that will become relevant below: that half of these words contain the phoneme /u/ (*oo* as in *aboot*[3]).

To get a qualitative overview of the semantics of the funniest words, in Figure 2 we have graphed the correlational structure between the word2vec vectors of the 234 words (4.7% of all rated words) that were judged to be most funny, defined as a humor rating >2 $SD = 3.28/5$. There is some clearly discernible semantic structure, although some of it may be a function of the particular words that Engelthaler and Hills (2017) had rated. Most notably, the superiority theory view of humor as denigration is clearly and strongly represented, by a very large and tightly intercorrelated cluster of insult terms, such as *twit, buffoon, nimrod, blockhead, ninny, scoundrel, hussy, douche, windbag, fathead*, and *dunce*, among many others. There are also clearly discernible smaller clusters of animal words (e.g., *hippo, mutt, chimp, baboon, dingo, heifer*, and *varmint*), body-related words focused on sexual and excretory organs (e.g., *pecker, crotch, penis, pubes, sphincter, scrotum, anus, belly*), and several groups of words generally related to having a good time: food-related words (*dumpling, goulash, sausage, strudel, fruitcake*), humor-related words (*chuckle, giggle, fun, antics, joke*), and a few music-related words (*pop, bop, bebop, funk, boogie*, and *polka*).

Some possible phonological structure is also apparent. There is a group of verbs with weakly correlated vectors (indicating a slight semantic relationship) but clear phonological similarity (*waddle, tickle, tingle, nibble, wriggle, jiggle, gobble, squabble*). There is also an apparent overrepresentation of words containing double letters, especially *oo*: for example, *putty, potty, whiff, puss, kisser, buddy, bollocks, smooch, cesspool, yahoo, bloomers, oomph*, and *stooge.* We discuss and quantitatively assess all these apparent regularities in more detail below.

Our goal in this paper was to model these judgment data, and thereby to operationalize, quantify, and compare the two dominant classes of humor theory discussed above: semantic theories, focused on the word meanings of categories associated with humor, and information-theoretic theories, focused on the degree to which a word's form is more or less probable, either because of its own frequency or because of the frequency and/or complexity of its com-

[3] A regional Canadian pronunciation of the word *about*, to rhyme with the word *hoot*.

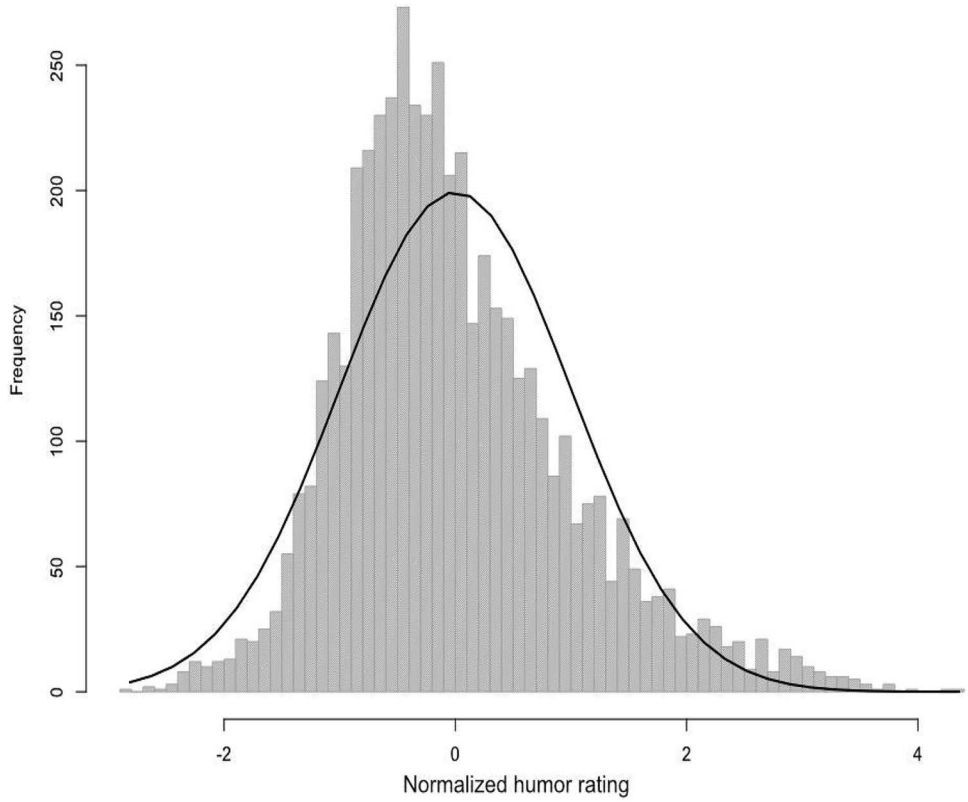


Figure 1. Distribution of humor ratings collected by Engelthaler and Hills (2017), with an idealized normal curve overlain in black.

ponents. In Study 1, we used linear regression to model and quantify the contributions of predictors associated with each of these theories independently, and then combine them to define a full-scale model of single-word humor judgments. In Studies 2 and 3, we present evidence that validates this model. In Study 2, we create a second model using a nonlinear, nonparametric modeling technique, and show that this model produces estimates that are nearly identical with the linear regression estimates. In Study 3, we show the estimates from both models are strongly predictive of new funniness judgments, collected for both a 100-word subset of the Engelthaler and Hills (2017) norms and for 100 previously unjudged words. We are able to model the data about as well as they can model themselves (about as well as split-half correlations along age and gender lines), suggesting that simple models can account for all of the systematic variance in this simple type of humor.

Study 1: Modeling With Linear Regression

Method

As a first step, we used linear regression to model the 4,573 (91.5%) of Engelthaler and Hills' (2017) ratings that had phonological representations in the Carnegie Mellon University (CMU) Pronouncing Dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict; the International Phonetic Alphabet version we used was downloaded from http://people.umass.edu/nconstan/CMU-IPA/), along with all other predictors we consider below. We built models on this full set rather than setting some aside for cross-validation because we planned to cross-validate the models experimentally on a new set of words, as we describe below.

Semantic predictors. We had two types of semantic predictors: general measures known to be components of word meaning, and components that we constructed specifically to focus on the semantics of single-word humor.

Our general measures consisted of computational estimates, from Hollis, Westbury, and Lefsrud (2017), of valence (how good or bad a word's referent is), arousal (how strongly a word's affect is marked), dominance (the extent to which a word's referent is controlling, vs. controlled), and concreteness (the extent to which a word's referent is concrete vs. abstract). We included these measures because they are all well-known and widely accepted factors in the organization of semantics, the first three having been shown by Osgood, Suci, & Tannenbaum (1957/1978) to emerge as principal components in semantic assessments of many different kinds and the latter having been implicated in Hollis & Westbury (2016; see also Hollis et al., 2016). Arousal has been previously implicated in humor (see review in Godkewitsch, 1972). We also included a fifth quasi-semantic measure that has been previously shown to predict judged humor in nonwords (Westbury et al., 2016): a binary rudeness measure that measured the presence of any of 52 phonologically defined rude substrings (rudesubstring). When used with words this would have the potential of boosting the predicted value on actual rude words, since rude words contain rude substrings by definition.

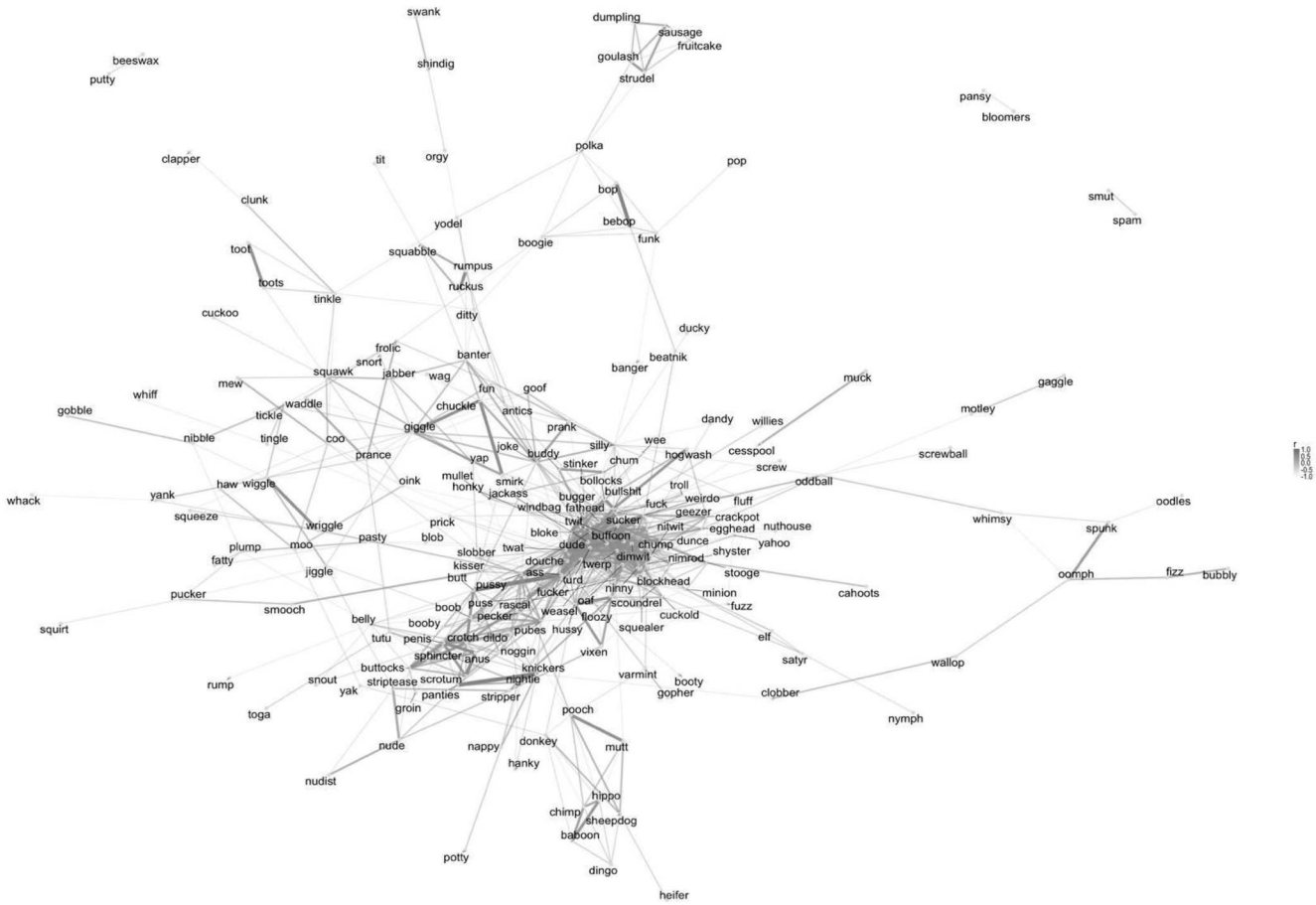


Figure 2. Correlation plot of the word2vec vectors of 197 (84%) words from the 234 words judged funniest (>2 *SD*) that correlated with at least one other word above the threshold of $r = .3$. Line thickness and darkness indicates the strength of the correlation. Distance between unconnected clusters has no interpretation.

To construct semantic measures specific to our focus on the humor of single words, we began by classifying the 200 words that had been judged funniest in the Engelthaler and Hills (2017) norms into one of six categories that we called *insult, sex, party, animal, body function*, and *expletive*. The insult category consisted of 35 insulting terms (e.g., *floozy, buffoon, douche*, and *stooge*). The sex category consisted of 27 words unambiguously related to sex (e.g., *fuck, twat, booty*, and *pecker*). We did not include more ambiguously sexual words such as *thruster, jiggle, tummy*, or *squeeze*, and did not include words such as *toots* and *hussy* that have sexual connotations but had already been classified as insults. The party category consisted of a looser collection of 27 nonsexual terms associated with having a good time. It included terms associated with food and drink (e.g., *dumpling, sausage, goulash, hooch, gobble*), with fun activities (e.g., *antics, boogie, banter*) and with friends and laughter (e.g., *buddy, chuckle, giggle, chum*). The animal category included 19 animal words associated with animals, mainly, but not exclusively, animal names (e.g., *baboon, gopher, hippo*, and *oink*). The body function category consisted of words related to nonsexual bodily functions (e.g., *burp, turd, slobber, whiff*). Finally, the expletive category contained the only four profane words among the 200 classified funniest: *bollocks, bullshit, fuck*, and the more gentle *hogwash*. Together these six categories accounted for 108 (54%) of the funniest 200 words.

In order to use these six word lists to quantify the humor of words, we took advantage of a quality of co-occurrence models, which is that distance from a vector defined by averaging together words from a single semantic category is a good measure of membership in that category. We call such vectors category-defining vectors (CDVs). For example, every one of the closest 100 neighbors to the average vector of the 19 words in the animal vector described is an animal name or a name for a class of animals (e.g., *critter*). We averaged together the words in each of the six categories of interest to define an interim CDV for each category. We then sorted the dictionary by proximity to each of those six interim CDVs and identified the first 100 words from that sorted list that belonged to each category. We used those 100 words to define a final CDV for each category. A word's cosine distance from each of those final CDVs is a measure of the degree to which that word is a member of that category. To allow the reader to assess our own judgment that this method has good face validity, we have included the 50 words closest to the definitive CDV for each category in Appendix A.

Our intent was to use cosine similarities to each of the six CDVs as predictors for identifying funny word candidates. However, similarities of all words in the dictionary from the six CDVs were highly correlated (average [*SD*] correlation: 0.62 [0.20] over 78,278 words; Table 1). This reflects, in part, that there was high overlap between the groups (although there was no overlap in the judged-humorous "seed words" we used to defined each interim CDV). Fifty-nine words appeared in at least two of the categories, including two words (*dude* and *puke*) that appeared in 5/6 categories and seven words (*bitch, cunt, douche, fuck, fucker, shit*, and *twat*) that appeared in 4/6 categories.

We performed principal components analysis to extract a single measure (the first principal component [PC]) that accounted for variance over all six categories. Since that PC was correlated at $r = .99$ ($p < .001$) with the average cosine similarity over the six vectors, we used the conceptually and mathematically simpler average similarity values to the six CDVs as our single semantic predictor of word funniness, which we call average-CDV. Average-CDV values account for about a quarter of the variance in human funniness judgments over the full set of 4,574 judged words, correlating with those judgments at $r = -0.51$ (words closer on average to the six categories are rated funnier; 95% confidence interval [CI]: −0.53 to −0.49, $p < .001$; Figure 3). This demonstrates the unsurprising fact that a large portion of the variance in those judgments is semantic and the perhaps more surprising fact that it is possible to define a single broad category of "funny things."

We conducted a multiple regression with the five general semantic predictors described above, plus average-CDV. The results are shown in Table 2. All six predictors entered into the model with $p < .05$. Together they account for 30.1% of the variance in the human funniness judgments. Valence and rudesubstring are positively associated with increased humor ratings. All other measures have negative associations.

The relationships between humor ratings and valence and arousal are shown graphically in Figure 4. Words with valence below about 1 *SD* from the mean (words not associated with positive emotion) are not judged funny, but there is little effect of valence above that threshold. Increasing arousal is associated with lower judged humor. This is because high arousal (>0 *SD*) is associated with low valence. The words predicted to be most strongly arousing are also predicted to be negatively valenced. We will consider the predictive role of this interaction in when we consider the full linear model.

Word-form predictors. We had four general word-form predictors. Length is the length of the word. Logfreq is the logged word frequency, using frequencies from the UseNet corpus compiled by Shaoul & Westbury, 2006. Logletterfreq is the logged average probability of the letter strings in each word, computed from approximately 4.5 billion characters of English text by Lyons (n.d.). We also computed the phonological equivalent of Logletterfreq, Logphonemefreq, which is the logged average probability of the phonemes in each word, using frequencies computed by Blumeyer, 2012.

Along with these general predictors, we included four predictors specific to the current analysis of humor judgments. We examined the *ratio* of the frequencies of occurrence of all letters and phonemes in funny (≥2 *SD* [>3.26/5] by human judgment) versus unfunny (<2 *SD* by human judgment) words. One phoneme stood out for having an outlying funny/unfunny frequency ratio, 3.28*z* above the mean for all other English phonemes: the phoneme /u/. This phoneme occurs in 17.4% of the words judged most funny, versus 6.3% of the rest of the words, 2.74 times more often in funny than unfunny words. Because of this disparity, we added a binary flag of its existence, Contains-/u/.

Two infrequent letters (*y* and *k*) had funny/unfunny ratios >2 *SD* from the average for all letters, so we also added flags for those letters, Contains-y and Contains-k.

We noted above that words ending in *le* seemed by inspection to be overrepresented among the funny words. In fact, final consonant+*le* does occur much (1.89 times) more often among words with funniness ratings >2 *SD* (6.47% of the time) than among words with funniness ratings <2 *SD* (3.42% of the time). We therefore added a binary marker indicating whether a word ended with a consonant (other than *l*) followed by a final-*le*, Cons+*le*.

We conducted a linear regression analysis entering just these form variables. All contributed with $p < .05$. The final model is shown in Table 3. It accounted for 19.2% of the variance in the human funniness judgments ($p < .001$), about 64% as much variance as was accounted for by the semantic measures alone (see Table 2).

Discussion

The finding that letter *k* is overrepresented in funny words and that its presence in a word is predictive of higher humor judgment is of particular interest since it relates to a famous piece of advice for comedians, brought to popular attention by the movie version of Neil Simon's play, *The Sunshine Boys* (Stark & Ross, 1975). In

Table 1
Correlations and Between Distances to Six Category-Defining Vectors and to Their Average, Over 78,278 Words

Category	Sex	Party	Insult	Profanity	Body function	Animals
Sex	1					
Party	.62 [.61, .62]	1				
Insult	.72 [.71, .72]	.65 [.65, .66]	1			
Profanity	.78 [.78, .78]	.64 [.63, .64]	.92 [.92, .92]	1		
Body function	.92 [.92, .92]	.58 [.58, .59]	.69 [.68, .69]	.77 [.77, .78]	1	
Animals	.49 [.49, .50]	.32 [.31, .32]	.35 [.34, .35]	.29 [.29, .300]	.50 [.50, .50]	1
Average	.92 [.91, .92]	.77 [.77, .77]	.89 [.88, .89]	.90 [.90, .90]	.90 [.90, .90]	.57 [.57, .58]

Note. Values in square brackets are 95% confidence intervals. All *p*s $< .001$.

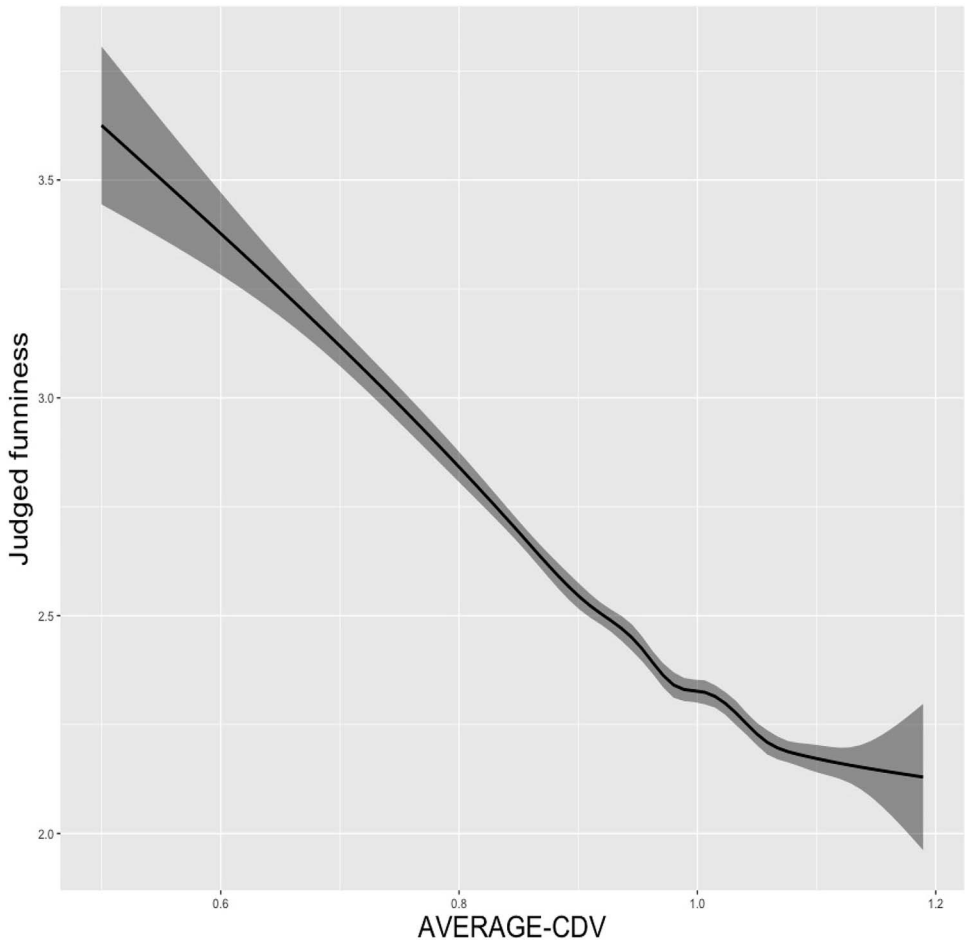


Figure 3. Judged Funniness × Average-CDV for 4,573 words, smoothed with a generalized additive model, with 95% confidence intervals.

that movie, a comedian (ignoring the difference between the phoneme */k/* and the letter *k*) tells his nephew:

> Fifty-seven years in this business, you learn a few things. You know what words are funny and which words are not funny. Alka Seltzer is funny. You say "Alka Seltzer" you get a laugh . . . Words with *k* in them are funny. Casey Stengel, that's a funny name. Robert Taylor is not funny. Cupcake is funny. Tomato is not funny. Cookie is funny. Cucumber is funny. Car keys. Cleveland . . . Cleveland is funny. Maryland is not funny. Then, there's chicken. Chicken is funny. Pickle is funny.

Table 2
Model for Predicting Judged Funniness From Semantic Cues Only

Predictor	Estimate	*SE*	*t*	*p*
(Intercept)	4.98	.12	39.97	<.001
Average-CDV	−2.42	.06	−41.87	<.001
Valence	1.03	.12	8.77	<.001
Dominance	−.83	.19	−4.36	<.001
Arousal	−.49	.10	−4.68	<.001
Concreteness	−.14	.04	−3.55	<.001
Rudesubstring	.07	.02	3.18	.0015

Note. Multiple R^2: .30. *F* statistic: 329.2 on 6 and 4,580 *df*. *p* value < .001. CDV = category-defining vector.

This notion that *k* is inherently funny did not originate with Simon. The American humor writer H. L. Mencken (1948) wrote that:

> *k*, for some occult reason, has always appealed to the oafish risibles of the American plain people, and its presence in the names of many other places has helped to make them joke towns also [like Podunk, the subject of Mencken's article]; for example, Kankakee, Kalamazoo, Hoboken, Hohokus, Yonkers, Squeedunk, Stinktown (the original name of Chicago), and Brooklyn. (p. 80)

We will defer discussion of why words containing the letter *k* might be funny until we have considered the full model below, which provides a clearer overview.

Why are words containing /u/ funny? A dissatisfying circular answer is: because a lot of funny words contain /u/. However, when we examine the 35 words containing /u/ that were rated highly (>2 *SD*) in funniness, we can see that no better answer may be possible. There are many funny /u/ words related to the funny category of sex but most bear no apparent etymological relationship to each other (*nude/nudist, booby/boob, pubes, booty, screw,*

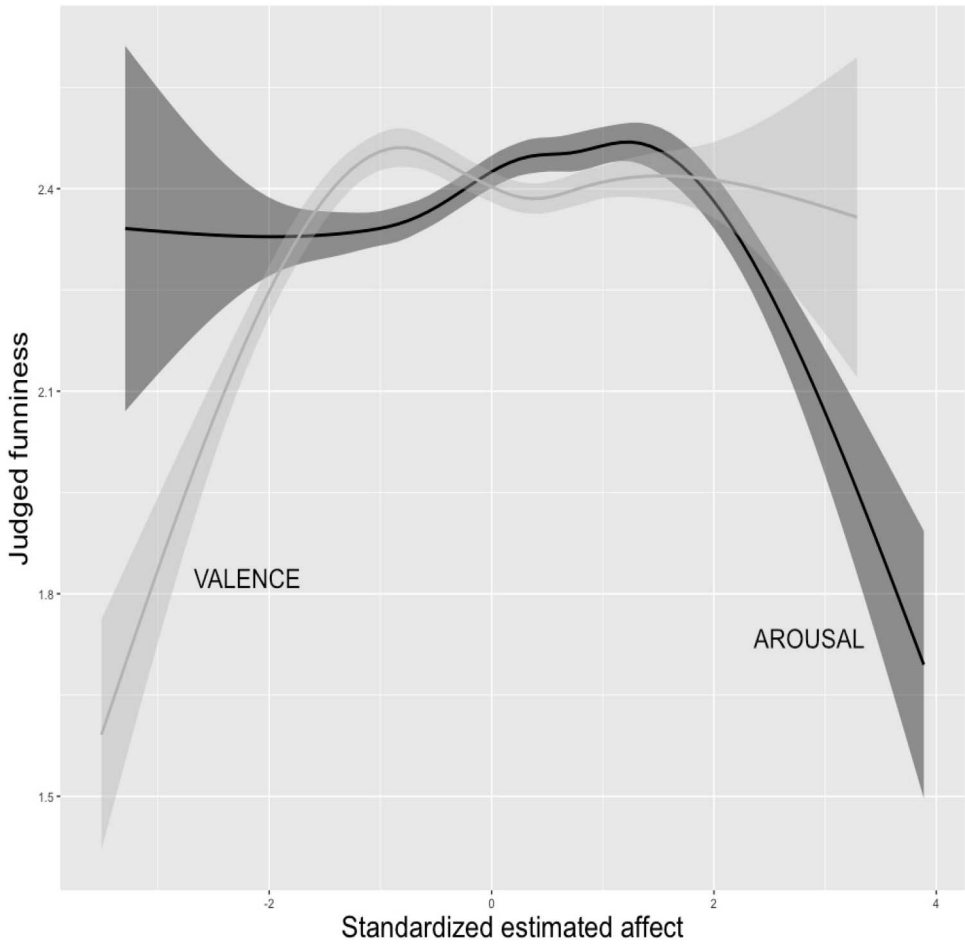


Figure 4. Estimated Funniness × Estimated Valence and Arousal for 4,573 words, smoothed with a generalized additive model, with 95% confidence intervals.

hooter, floozy, bloomers). Similarly, there are many funny /u/ words related to the funny category of insults that also seem etymologically unrelated (*douche, buffoon, stooge, goof, poof, yahoo, screwball*). The 35 funniest /u/ words also contain three animal names (*pooch, baboon, cuckoo*) and three words potentially related to the party category (*fruitcake, goulash*, and *hooch*). This subset of 23 (66%) of the 35 funniest /u/ words thereby hits at least four of the six categories we identified above as associated most strongly with humor. There may, of course, be a deeper (possibly sound-symbolic) reason for why so many words identified as funny contain the phoneme /u/, but it is difficult to imagine what that reason could be, given the wide range of words that are implicated. Some models of language development predict the emergence *by pure chance* of such word/meaning correlations. In introducing their naïve discriminant learning model, Baayen, Milin, Đurđević, Hendrix, and Marelli (2011) wrote that:

> a discriminative learning approach [to language development] predicts that even small local consistencies in the fractionated chaos of local form-meaning correspondences will be reflected in the weights [for shared meaning between linguistic strings], and that they will codetermine lexical processing, however minute these contributions may be. (p. 56)

To quantify the humor-predicting value of the set of 34 funny /u/ words, we created a CDV composed of the average of their vectors. Over our entire dictionary of 45,516 words, distances from that /u/-defined CDV correlate with average-CDV at $r = .93$ (95% CI [0.93, 0.93], $p < .001$), near identity for such a large number of

Table 3
Initial Linear Regression Model for Predicting Judged Funniness From Formal Cues Only

Predictor	Estimate	*SE*	*t*	*p*
(Intercept)	3.2	.05	61.04	<.001
Logletterfreq	−.37	.03	−12.05	<.001
Contains-/u/	.15	.02	6.77	<.001
Cons+*le*	.14	.03	4.67	<.001
Logphonemefreq	−.14	.06	−2.57	.010
Contains-k	.06	.02	3.14	.0017
Logfreq	−.05	.0021	−25.21	.010
Contains-y	.04	.02	1.96	.050

Note. See also Table 4. Multiple R^2: .19. *F* statistic: 155.5 on 7 and 4,565 *df*. *p* value <.001.

values. This should not be surprising, since both values reflect similarity to words that are known to be funny. Over the 4,573 human-rated words with all predictors, similarity to the /u/-defined CDV correlate with humor ratings at −0.54 (95% CI [−0.56, −0.52], $p <$ .001), indicating that the cosine distance to the average vector of 34 funny words containing /u/ accounts in itself for about 28.8% of the variance in humor ratings.

We can undertake a similar analysis for the Cons+*le* words. There are only 10 words in this category that have humor ratings above 2 *SD*: *gaggle, jiggle, tinkle, waddle, wiggle, wriggle*, plus two words related to eating (*gobble* and *nibble*) and two words related to laughter (*giggle* and *chuckle*). The CDV defined by averaging the vectors of these 10 words is correlated over our 45,516 word dictionary with average-CDV at $r = .61$ (95% CI [0.60, 0.61], $p <$ .001).

To understand their relationship, we created a visual depiction of the strength of the correlations between distances from the semantic CDVs, plus their average, and distances from the vectors for each of the 10 Cons+*le* words, plus their average (Figure 5). The figure suggests that the funniness of Cons+*le* may be largely due to the two words *giggle* and *waddle*. Cosine distances from each of these vectors are strongly correlated with each other, with the average Cons+*le* distances, and with the party semantic category ($r > .65$, $p <$ .001 in all cases). Distances from the vector for the word *giggle* are also strongly correlated with distances from the sex category CDV ($r = .67$, 95% CI [0.666, 0.674], $p <$ .001) and the body function category CDV ($r = .66$, 95% CI [0.66, 0.66], $p <$ .001), in keeping with the discussion above of sex and bodily excretion as inherently humorous.

Further explanation for the humorous aspect of CONS-*le* words may be found in the morphological role of the suffix *–le*, which is complex. According to the online Oxford English Dictionary (OED; http://www.oed.com/), the suffix serves at least four roles: as a diminutive (e.g., *bramble, twinkle*), as a marker of repetition (e.g., *crackle, sparkle, paddle*), as a marker of propensity (e.g., *brittle, fickle, nimble*), as an expression of membership in the category of tools (e.g., *thimble, handle, beadle*), and as a mistaken singular form of the nonplural Old English word ending *–els* (as in *riddle*, which the OED suggests was incorrectly derived historically by dropping the final letter of the old English singular form *raedels*). All 10 of the Cons+*le* words that were judged most funny are words that mark repetition, usually with a diminutive aspect. For example, *to nibble* is to take small repeated bites; *to jiggle*, *to wiggle*, and *to wriggle* are all to make small repeated movements; and *to tinkle*, *to giggle*, and *to chuckle* are to make small repeated sounds. We speculate that because repetition and tininess are low-frequency events (and small size is often positively valenced), these words achieve their humorous status through having a morphological marker of incongruity. Some weak evidence for this is that there are several diminutive words that do not end with Cons+*le* that have a very high judged funniness (>2 *SD*): *booby, whimsy, quickie, panties, piggy, dump-*

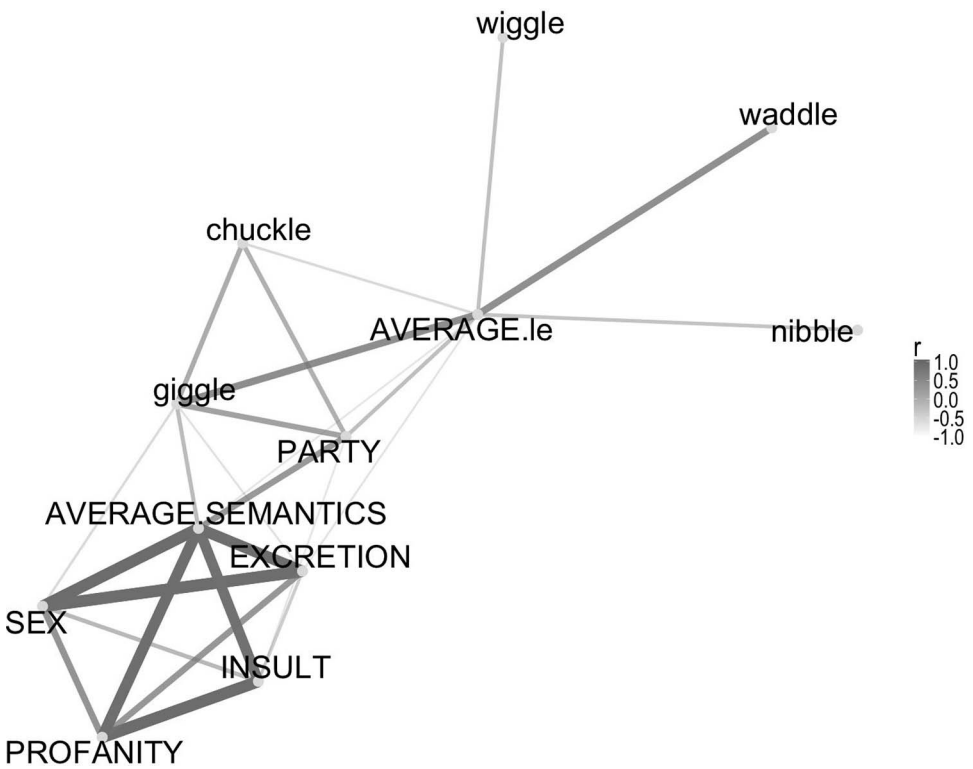


Figure 5. Correlations between distances from the six semantic CDVs, plus their average (AVERAGE.SEMANTICS), and distances from the vectors for each of the 10 funniest (>2 *SD*) CONS+le words, plus their average (AVERAGE.le), over a dictionary of 78,279 words. The figure is thresholded at $r \geq 0.65$ ($p <$ .001); correlations lower than that are not shown.

ling, and *ditty*. Note that, as these examples show, final-*y*, which we associated above with funniness, can sometimes (though it does not always) function as a diminutive. Conversely, there are many words with above average valence that end in Cons+*le*, but have a funniness rating <-0.5 *SD* (e.g., *settle, handle, stable, marble*, and *purple*). None are diminutives.

This discussion suggests that simply marking the orthographic pattern Cons+*le* may not be the best marker of the morpheme's role in humor judgments, since the morpheme marks both funny and nonfunny words. Accordingly, we reran the model in Table 3, but instead of using the binary marker of orthographic Cons+*le*, we used distance from the CDV obtained by averaging the eight of the 10 Cons+*le* words with humor ratings above 2 *SD*, excluding *giggle* and *chuckle* because of their overt semantic association with humor. Distance from this CDV (Eight-*le*-CDV) correlated with human humor judgments at $r = -0.39$ (95% CI [−0.42, −0.37], $p < .001$). Substituting Cons+*le* with Eight-*le*-CDV did result in a substantial improvement to the form model ($R^2 = 0.276$, Akaike information criterion [AIC] = 3,758) compared with the original model in Table 3 ($R^2 = 0.195$, AIC = 4,256). However, note that we have achieved this improvement by redefining a formal marker as a semantic marker. One advantage of this substitution is that other forms of Cons+*le* root words (e.g., *wiggling, wiggly, tiddly, prattled*) will be marked as funny, since they have vectors similar to Eight-*le*-CDV.

The final linear regression model derived from formal markers, with Eight-*le*-CDV substituted for the Cons+*le* flag, is shown in Table 4.

Full Model

To construct a full model, we entered all the predictors from the semantic and form models at once. Two formal predictors (Contains-y, Contains-k) dropped out of the model with $p > .05$, leaving a model with 10 predictors that produced estimates accounting for 39.5% of the variance in the human funniness judgments ($p < .001$; AIC = 2,937). Adding an interaction between average-CDV and Eight-le-CDV (which amounts to putting emphasis on diminutives among words semantically related to funniness, if the analysis above is correct) improved the model ($R^2 = 0.402$; AIC = 2,893). The AIC values suggested that the model with this interaction was millions of times more likely to minimize information loss. Adding a three-way interaction between valence, arousal, and dominance improved the model even more substantially, reducing the AIC to 2,727 (while knocking rudesubstring out of the model) and increasing the amount for variance accounted for to 42.4%. Words that have a very low product of the three affect predictors (e.g., *ailment, stench, neglect, syndrome, illness, strain*) are especially nonfunny. Finally, adding an interaction between Logletterfreq and Logphonofreq improved the model slightly more, to account for 42.6% of the variance (AIC = 2,716, over 200 times more likely to reduce information loss than the same model without this interaction).

The final linear regression model is shown in Table 5. The raw correlations between the individual predictors in the model and the human funniness judgments are shown in Table 6.

Table 4
Final Linear Regression Model for Predicting Judged Funniness From Formal Cues, Plus Eight-le-CDV

Predictor	Estimate	*SE*	*t*	*p*
(Intercept)	4.18	.06	64.54	<.001
Eight-*le*-CDV	−1.23	.05	−23.45	<.001
Logletterfreq	−.25	.03	−8.64	<.001
Logphonemefreq	−.15	.05	−2.88	.0040
Contains-/u/	.15	.02	7.02	<.001
Contains-y	.06	.02	2.96	.0031
Logfreq	−.04	.002	−22.34	<.001
Contains-k	.04	.02	2.46	.01

Note. See also Table 3. Multiple R^2: .276. *F* statistic: 248.5 on 7 and 4,565 *df*. *p* value <.001. CDV = category-defining vector.

Results

As shown in Table 6, by far the strongest predictors of funniness are the semantic predictors, average-CDV, $r = -0.52$, $p < .001$, Eight-*le*-CDV, $r = -0.39$, $p < .001$, and their interaction, $r = -0.53$, $p < .001$. Words are funny simply because they are associated with (operationally, closer to the CDVs of) the categories we have discussed above as being funny.

The next strongest individual predictor in Table 6 is Logfreq. As shown in Figure 6, less frequent words are judged funnier, $r = -0.36$, $p < .001$. This is consistent with incongruity theory, which suggests that more improbable elements should be perceived as funnier. Note that the beta weight assigned in the full model in Table 5 is much lower, at −0.04, suggesting that some of the variance attributable to logged frequency is shared with other predictors in the model.

Word frequency is one measure of a word's improbability. Another is the probability of its components. Both Logletterfreq and Logphonemefreq were strong predictors of funniness (Figure 7). Consistent with the findings of Westbury et al. (2016) and with incongruity theory, strings with less common letters or phonemes are judged funnier than strings with more common letters. Figure 7 shows that the effect is stronger across the range of the predictors for Logletterfreq than it is for Logphonemefreq. Figure 8 breaks this effect down. It shows that there are negative correlations between the difference in how much more often a letter or phoneme occurs in funny (>2 *SD*) versus unfunny words, and the global frequency of those letters or phonemes. In other words, the more likely it is that a letter or phoneme occurs more often in funny words than in unfunny words, the less probable that letter or phoneme is (for phonemes: $r = -0.59$, 95% CI [−0.76, −0.33], $p < .0001$; for letters: $r = -0.71$, 95% CI [−0.859, −0.441], $p < .0001$). This is the third time that a negative correlation between *symbolization strength* and *symbol frequency* has been demonstrated. As we have mentioned above, Westbury et al. (2016) showed that nonword funniness judgments were well predicted by the frequency of the letters in the judged string. Westbury, Hollis, Sidhu, and Pexman (2018) extended this to show an inverse relationship between both average letter and phoneme frequency and the probability that a nonword string would be judged as a good word for an entity in a range of different categories.

Table 6 also suggests that words that have higher concreteness, $r = .11$, $p < .001$ and valence, $r = .07$, $p < .001$ are judged funnier than words with lower values. The association with concreteness probably reflects the strong semantic emphasis on em-

Table 5
Final Linear Regression Model for Predicting Judged Funniness From Formal and Semantic Cues

Predictor	Estimate	*SE*	*t*	*p*
(Intercept)	13.37	1.28	10.44	<.001
Dominance	−7.5	2.38	−3.15	.0017
Valence	−5.81	2.36	−2.47	.010
Average-CDV	−4.59	.44	−10.41	<.001
Eight-*le*-CDV	−3.44	.45	−7.71	<.001
Arousal	−11.84	2.33	−5.09	<.001
Logphonemefreq	−1.28	.32	−3.95	<.001
Logletterfreq	−.59	.1	−5.98	<.001
Concreteness	−.21	.04	−5.55	<.001
Contains-/u/	.11	.02	5.7	<.001
Logfreq	−.040	.0019	−18.86	<.001
Valence × Arousal	16.79	4.7	3.58	<.001
Arousal × Dominance	17.48	4.77	3.67	<.001
Valence × Dominance	8.79	3.85	2.28	.020
Valence × Arousal × Dominance	−23.48	7.67	−3.06	.0022
Average-CDV × Eight-*le*-CDV	3.05	.46	6.62	<.001
Logletterfreq × Logphonemefreq	.65	.18	3.62	<.001

Note. Multiple R^2: .4186. *F* statistic: 198 on 16 and 4,401 *df*. *p* value < .001. CDV = category-defining vector.

bodiment (sex, bodily functions, silly people) as funny. The somewhat weak association of humor judgments with valence captures the unsurprising fact that sad things are not funny. Figure 4 suggests that the effect is nonlinear, with judged funniness being roughly constant for words with valence >0.4 (−1.6 *SD*), but dropping rapidly for words with valence lower than that. The least funny word in the list by both human judgment and by the model was the extremely low valence (−2.84 *SD*) word *rape*.

Table 6 includes two other predictors that have statistically significant but rather weak associations with humor ratings: Contains-/u/, $r = .11$, $p < .001$, and arousal, $r = .05$, $p = .0014$. Dominance is a nonsignificant predictor of humor in itself in itself, $r = .01$, $p = .45$, but interacts with valence and arousal, both individually and together.

One predictor that does not appear in Table 6 despite being a significant predictor of funniness judgments is word length. Shorter words are judged funnier than longer words ($r = -0.079$, 95% CI [−0.11, −0.05], $p < .001$). Collapsed across the lengths in the full range of the judged words (length 4–12), the magnitude of the correlation between the eight word lengths and the average funniness judgments for those eight lengths was large ($r = -0.91$, 95% CI [−0.64, −0.98], $p < .001$). Linear regression revealed that the humor cost of each additional letter was about −0.03 of a rating point, summing to about a quarter of a rating point over the whole range. Length probably does not appear in the final model because it is significantly correlated ($p < .001$) with many predictors that do appear in that model, including average-CDV, Eight-*le*-CDV, and the average letter and phoneme frequencies.

We are able to use the regression model in Table 5 to estimate funniness ratings for 45,516 words for which we have the necessary measures. Across this set, the 10 words estimated to be funniest are (in descending order): *upchuck, bubby, boff, wriggly, yaps, giggle, cooch, guffaw, puffball*, and *jiggly*. The 200 words estimated to be funniest are listed in Appendix B. The word estimated to be least funny was *harassment*.

The correlations between the vectors of the 500 words [top 1.1%] estimated funniest are shown in Figure 9. Several categories seen in the original human-rated words are apparent in that figure: insult terms (e.g., *floozy, bozo, honky, schmuck*), words related to body excretions (e.g., *puke, pooping, fart, drool*), words related to sexuality (e.g., *dick, titty, pussy, blowjob*), words related to general good times (e.g., *schmoozing, nattering, frolicking, cavorting*), words related to laughter (e.g., *giggle, snicker, chortle, guffaw*), and animal-related words (e.g., *porker, pooch, grackle, oink*).

Constraints imposed by the model explain some of the apparent oddities in Figure 9. One is the large number of words ending in *y* (e.g., *floozy, hussy, bunny, boozy*). The letter *y* is a low frequency letter, so words that contain it get rated as funny due to the weight on average letter frequency, especially if they are also low frequency words or contain other low frequency letters. Moreover, as mentioned above, final *y* can serve as a diminutive marker. We have suggested above that diminutive terms are likely to be judged as funny.

We used the same method to build individual models for estimating the male judgments and female judgments provided by Engelthaler and Hills (2017). Although there were some small differences in the models (perhaps most notably, in light of the earlier discussion of the comedic value of the letter *k*, that the binary marker Contains-k entered into the male but not the female model), across the full set of words, both male and female estimates correlated with the global estimates at $r > .98$ ($p < .001$). They correlated with each other at $r = .98$ (95% CI [0.97, 0.98], $p < .001$). Because they are so similar to the model in Table 5, we have not reproduced the models here. We also attempted to model the difference between male and female judgments. The attempt was not successful, with the best model accounting for just 1% of the variance.

Although their very high correlations with each other and with the global estimates suggest that the model estimates for each

Table 6
Correlations Between the Individual Predictors From the Model in Table 5 and 4,573 Human Funniness Judgments, in Descending Order of Correlation Magnitude

Predictor	*r*	95% CI *LL*	95% CI *UL*	*p*
Average-CDV: Eight-*le*-CDV	−.53	−.54	−.50	<.001
Average-CDV	−.52	−.54	−.50	<.001
Eight-*le*-CDV	−.39	−.42	−.37	<.001
Logfreq	−.36	−.39	−.34	<.001
Logletterfreq	−.24	−.27	−.21	<.001
Logletterfreq: Logphonemefreq	−.20	−.23	−.17	<.001
Logphonemefreq	−.14	−.17	−.11	<.001
Valence × Arousal	.12	.10	.15	<.001
Contains-/u/	.11	.08	.14	<.001
Concreteness	.11	.080	.14	<.001
Valence × Arousal × Dominance	.088	.054	.11	<.001
Arousal × Dominance	.086	.054	.11	<.001
Valence	.071	.037	.094	<.001
Arousal	.046	.018	.076	.0014
Valence × Dominance	.036	.0020	.059	.038
Dominance	.016	−.018	.040	.45

Note. CI = confidence interval; *LL* = lower limit; *UL* = upper limit; CDV = category-defining vector.

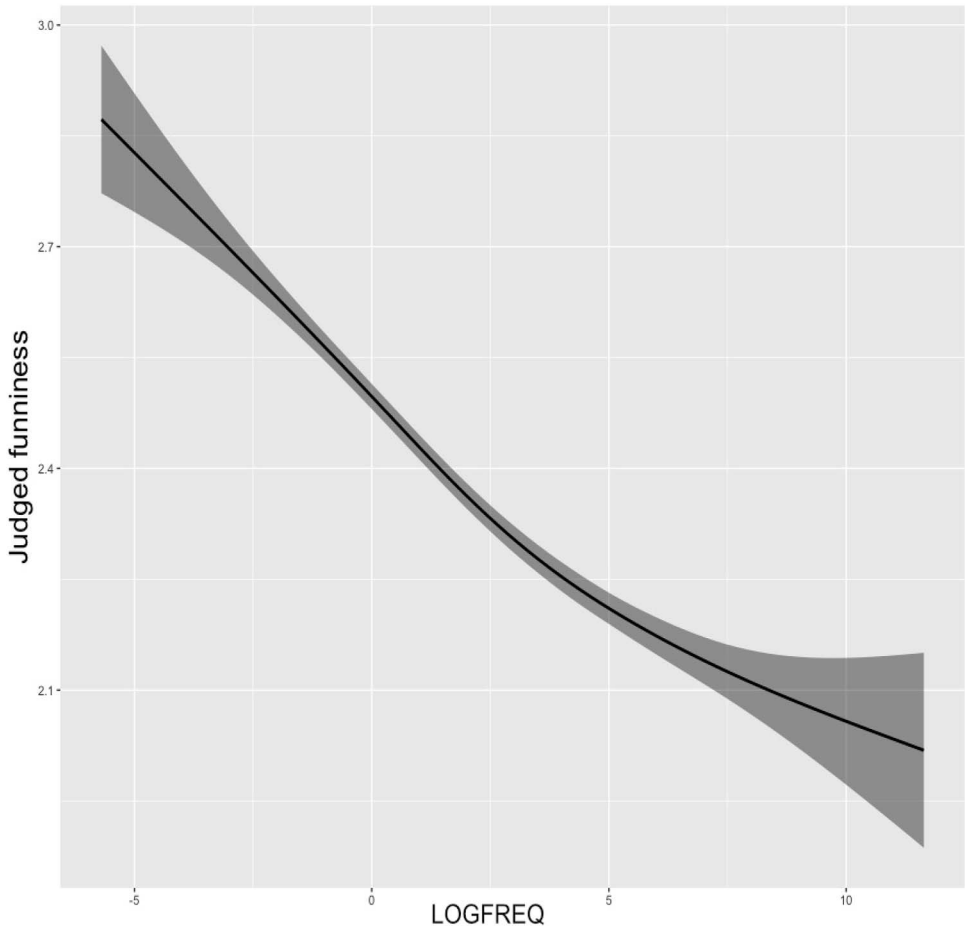


Figure 6. Estimated Funniness × Average $\log_2$ word frequency [LOGFREQ] for 4,573 words, smoothed with a generalized additive model, with 95% confidence intervals.

gender are nearly identical, we attempted to characterize the gender differences qualitatively by examining the words that were assigned the largest differences by the models over the 45,516 words for which we have all measures. Confining ourselves only to the 1,746 words in that set that had mean humor estimates ≥2 *SD*, there were 39 words with normalized male estimate-normalized female estimate differences >0.5 *SD* (words estimated funnier for males), including, for example, *douche, puke, cooch, fucker*, and *wank*. Only nine words showed the opposite difference with normalized male-normalized female estimate differences <−0.5 *SD* (words estimated funnier for females): *saunter*, *pizzazz/pizazz*,[4] *impassion*, *purr*, *pussyfoot*, *frolic*, *snazzy*, and *placidly*. By inspection, this difference suggests that males find more humor than females in negatively valenced, high arousal words. Figure 10 lends support to this idea, showing that the normalized male-normalized female estimates are maximal for words estimated to have low valence, dominance, and arousal. Although that graph shows effects over the full range of estimated funniness, we note that normalized male-normalized female estimates effects above 0.5 *SD* (roughly speaking, effects large enough to matter) are limited only to the extreme ends of this range.

Engelthaler and Hills (2017) also broke down their norms by age, splitting their participants at Age 32. As with the judgments by gender, we modeled the young and old judgments separately. Across all judged words, both old and young estimates correlated with the global estimates at $r > .95$ ($p < .001$). They correlated with each other at $r = .93$ (95% CI [0.93, 0.94], $p < .001$). As above, due to the close similarity of their estimates to those produced by the model in Table 2, we have not reproduced the models here.

We again attempted to characterize the gender differences qualitatively by examining the words that were assigned the largest differences by the models over the full set of 45,468 words. Among the 1,746 words in that set that had mean humor estimates ≥2 *SD*, there were 135 words that were estimated at least 0.5 *SD* funnier by the normalized young model than the normalized old model. The words with the largest normalized young minus normalized old funniness differences were *pusillanimous, pock, unfortunates, slouching*, and

[4] The OED lists the word with the first spelling, but also uses the second, without comment, in its examples, including as the first known use. Both spellings appear to be acceptable.

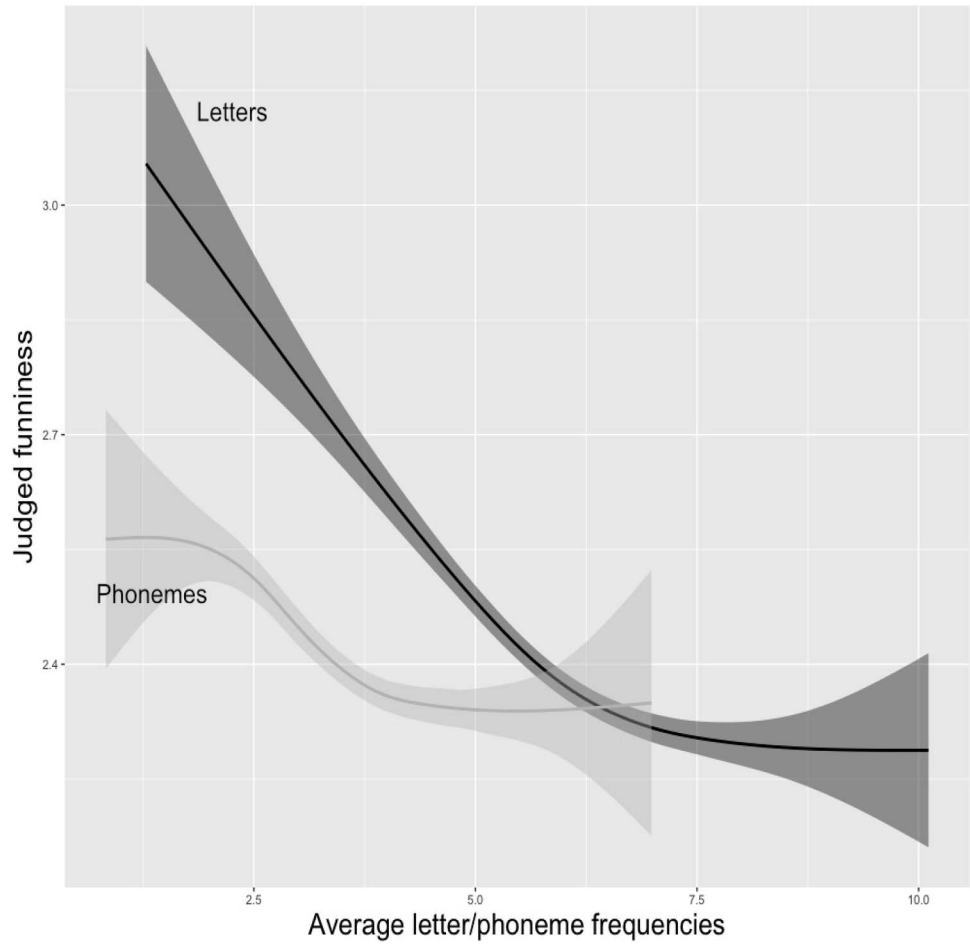


Figure 7. Judged Funniness × Average Letter Frequency (dark) and average phoneme frequency (light) for 4,573 words, fitted with generalized additive models, with 95% confidence intervals.

pissant. One hundred ninety-six words were estimated at least 0.5 *SD* funnier by the old model than the young model. The words showing the largest estimated difference in that direction were *playful, yummy, bubbly, joyfully*, and *elbow room*. Inspection of these exemplar words suggests the hypothesis that words estimated to be judged funnier by older judges tend to be high valence words. Figure 11 clarifies this possibility, suggesting that older adults tend to assign larger humor ratings to words that are extreme in either direction (very high or very low) in estimated dominance and valence. This effect is modulated by arousal at the positive end only. Words that have higher arousal values are judged more humorous by the older judges when those words also have high valence and dominance values.

Although the predictor Contains-k did not enter into the final model, the letter *k* is nevertheless highly overrepresented among the words estimated to be more than 2 *SD*s from the mean in estimated funniness. It appears in 16.7% of those 1,746 funniest words, compared with just 6.7% of the remaining 43,770 words, a disparity that is extremely unlikely to occur by chance ($\chi^2(1) = 252.04$, $p < .001$). The same is true for the letter *y*, which also was not flagged in the model, but occurred in 23.9% of the words estimated to have humor >2 *SD*, as compared with appearing in 11.8% of the remaining words ($\chi^2(1) = 228.70$, $p < .001$). Since the model that assigned the estimates did not include an explicit marker of these letters' occurrences, this disparity must occur due to some other factor. One factor might be that the letters are both rare. Letter *k* is the fifth least-common letter after *j*, *x*, and *z*, and the similar-sounding *q*. Letter *y* is the seventh least-common letter.

There may also be some semantic reasons for the letter *k*'s role in humor, as comedian lore has long suggested. Figure 12 shows the similarity (correlation) between the vectors of the words containing *k* that fell more than 2 *SD* above the mean in estimated funniness. It shows that there is some clear semantic structure among those words, notably a large cluster of words related to humor (e.g., *chuckle, smirk, wisecrack, snicker*), a cluster of words related to sex (e.g., *wank, dick, pecker, fuck*) as well as words related to insults (e.g., *honky, sucker, jerk, jackass*), and to animals (e.g., *cluck, porker, squawk, grackle*). We note that this observed coherence may be an *effect* rather than a *cause* of funniness, since these k-containing words were specifically selected to be funny.

Across the 45,516 words, word length correlated with estimated funniness at $r = -0.15$ (95% CI [−0.16, −0.14], $p < .001$), again suggesting a small negative effect of word length on humor judgments.

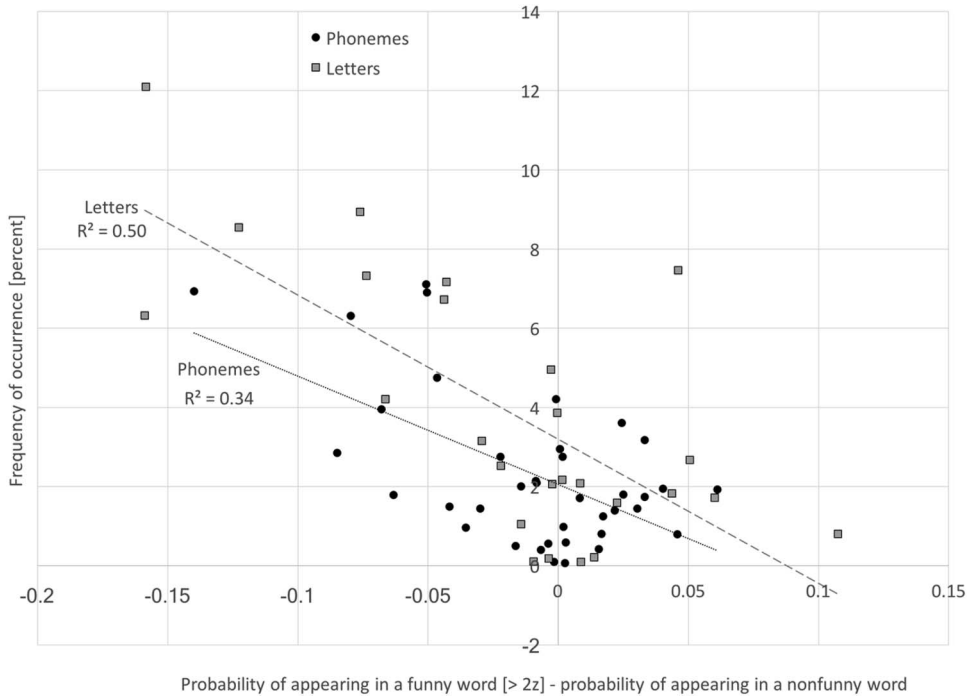


Figure 8. Relationship between the differential probability of a letter or phoneme appearing in a funny (>2 *SD*) versus an unfunny word (*x*-axis) and the probability of that letter or phoneme (*y*-axis).

Study 2: Nonparametric Modeling

Linear regression has many limitations. The most obvious limitation, of course, is that linear regression models are limited to documenting linear relations. Evidence considered above (see Figures 4, 7, and 11) suggests that some of the relations in which we are interested are not linear. A second limitation is that linear regression models are arbitrarily parameterized, since the modeler must make a decision about what criterion to use to allow a predictor into the model. The models would be different if we had set different criteria for entry into the models. A third limitation is that linear regression models make assumptions (of normality, heteroscedasticity, and mutual independence) about the distribution of and relationships between the predictors in the model. We know that these are not met by some of our predictors: our binary predictors are, of course, not normally distributed; Logletterfreq has a high kurtosis of 3.05, reflecting the fact that is bounded on the high end and has much variance on the low end; and some of our predictors are correlated (e.g., average-CDV and EIGHT-*le*-CDV are correlated at $r = .53$, 95% CI [0.521, 0.534], $p < .001$; arousal and dominance are correlated at $r = -0.37$, 95% CI [−0.376, −0.360], $p < .001$). Finally, although linear regression models provide an explicit mathematical description of the relationship between predictors and the dependent measure, that description can be difficult for us to grasp when there are many predictors. These weaknesses make it desirable to replicate the linear modeling presented above using a nonparametric method that does not make any assumptions about the shape of the relationship between predictors and the dependent measure, in the hope that it may provide a simpler and/or more useful model of the phenomenon of interest. At a minimum, it may increase our confidence in the model have developed above using linear regression, by replicating it using a different methodology.

To do this, we used a relatively little-used but powerful computational method, *genetic programming*. Genetic programming harnesses the power of natural selection to evolve equations that maximize a given fitness function, which, in our case, is the correlation between the evolved equation and the human funniness judgments. Computationally, the method is quite simple, and can be conceived of as an automated method for selective breeding of mathematical equations. To begin the process, the genetic programming algorithm constructs a large set of random equations formed by legal conjunctions of specified mathematical operators and predictors. It uses the fitness function to rank order these equations on a randomly selected subset of the data in terms of their fitness. It then discards all but the fittest equations and repopulates the set by forming new "offspring" equations constructed by randomly conjoining elements from the equations it has determined are most fit. By repeating this process across many generations, the method converges on a maximally fit model. The method is stochastic, so to satisfy ourselves that we have the best possible description of the data, we can conduct many independent runs. A useful feature of genetic programming for modeling is that it is possible to specify a length penalty, which reduces the fitness of equations as a function of their length, preventing the otherwise-likely evolution of extremely rococo solutions that may be optimally predictive, but very difficult for humans to understand. An example of the functional details of genetic programming is provided in Appendix D.

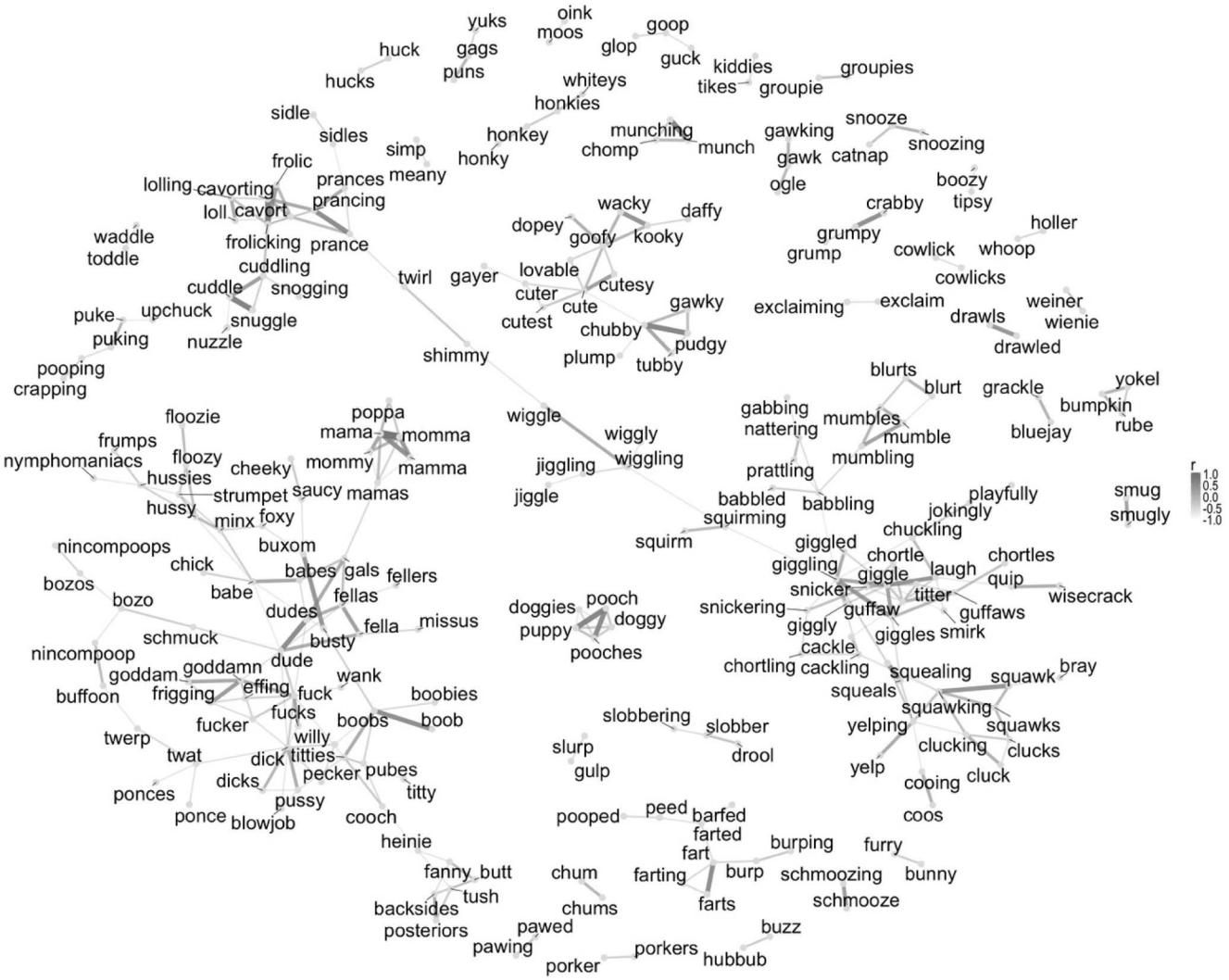


Figure 9. Correlations between the vectors of 241 of the 500 words estimated funniest by the final model in Table 5. Words not shown did not correlate with any other words above the threshold of $r = .5$. Line thickness and darkness both indicate the strength of the correlation. Distance between unconnected clusters has no interpretation.

Method

We used the Naturalistic University of Alberta Nonlinear Correlation Explorer (Hollis & Westbury, 2006; Hollis, Westbury, & Peterson, 2006), a software application designed explicitly for the purposes of evolving highest-correlation (maximal R^2) models, given a dependent measure and a set of predictors. Our predictors were all the predictors in the final model. The software automatically allows access to random real number constants. The function set consisted of the four basic mathematical operators (addition, subtraction, multiplication, and division) as well as a logarithm function, functions for squaring and cubing, and functions for finding square and cube roots. The population size was 2,000. Each equation was scored by correlating it with a random one third of the data, which means that evolved equations are cross-validated as part of the algorithm which generates them. Each run consisted of 1,000 generations, unless the run converged on the same equation for 50 generations in a row, in which case the run was terminated. Our fitness function was the linear correlation of the funniness judgments with the evolved equation. It used "greedy overselection" (Koza, 1992), which replaces the worst-performing 80% of the population in the previous generation with children produced from the best-performing 20%, with the remainder of the population filled by children of the worst-performing 80% from the previous generation. We allowed for a mutation rate of 0.5%, which means that 0.5% of nodes (operators or arguments) were randomly changed during offspring creation. We imposed a parsimony penalty to penalize very long solutions. It was set to 0.05, which means that an equation's fitness was reduced by 0.05 × its numbers of nodes. We ran 100 runs. Since many may be unfamiliar with this little-used methodology, we note that these settings (a

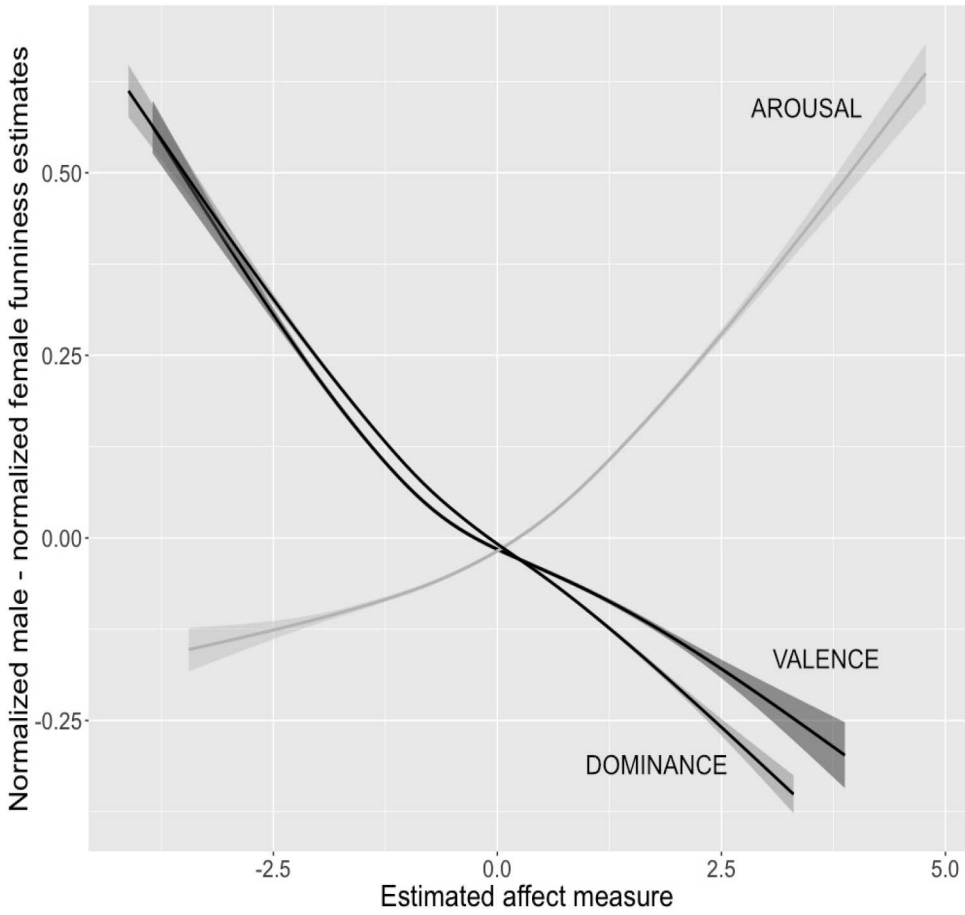


Figure 10. Difference between normalized funniness estimates by Gender (*y*-axis) × Normalized Estimated Affect Measures (valence, arousal, and dominance; *y*-axis) for 45,516 words, fitted with generalized additive models and showing 95% confidence intervals.

fairly large population coupled with a large number of potentially long runs) constitute a very thorough search of the space of possible equations and are designed to disallow extremely long solutions.

Results

The best-performing equation arose in Generation 73 of Run 56. In the prefix notation in which it was evolved, it was:

```
(ln (- (* (cube (+ (/(+ (* AVERAGE-CDV (sqrt
LogAveLetterFreq)) (× AVERAGE-CDV (/AROUSAL
DOMINANCE))) 0.307) EIGHT-1e-CDV)) EIGHT-1e-
CDV) (- (cbrt EIGHT-1e-CDV) (* AVERAGE-CDV (+
(* AVERAGE-CDV (+ (ln AVERAGE-CDV) (square (/(-
LOGFREQ (- VALENCE (* (sqrt LogAveLetterFreq)
(+ (* EIGHT-1e-CDV (+ DOMINANCE (square (+
AVERAGE-CDV (- (/(cbrt AVERAGE-CDV) (/0.354 (+
(cube AROUSAL) (* (sqrt (sqrt LogAveLetter-
Freq)) 0.704)))) (/VALENCE 0.354)))))) (ln (*
AVERAGE-CDV (sqrt (cbrt AVERAGE-CDV)))))))
DOMINANCE)))) (cbrt EIGHT-1e-CDV))))))
```

We hand-simplified this and were able to slightly shorten it. We have not reproduced that simplified equation here since it was nearly as complex as the one above, offering no additional insight.

The evolved equation includes seven predictors: the two semantic measures, average-CDV and Eight-*le*-CDV; two frequency measures, Logfreq and Logletterfreq; and the three affective measures valence, arousal, and dominance. It achieved a correlation with the set of human funniness judgments of $r = .65$ (95% CI [0.63, 0.67], $p < .001$). We tested whether the residuals from this correlation were normally distributed and found they were not (by Shapiro-Wilks test: $W = 0.99$, $p < .001$). As shown in Figure 13 the residuals reflect some of the nonnormal qualities of the underlying judgments (see Figure 1), with more items near and just below the average than in a perfectly normal distribution. This suggests that the model has left some systematic variance unaccounted for, though perhaps mainly in the least-interesting mid to low range of the spectrum of funniness.

By Fisher's *r*-to-*z* test, the correlation of the estimates with human judgments is not different from that achieved by the linear equation considered above ($z = 0.33$, $p = .74$). However, the two are not directly comparable at this stage, since the nonlinear equation is cross-validated (in virtue of being derived and repeatedly tested across random thirds of the data) and the

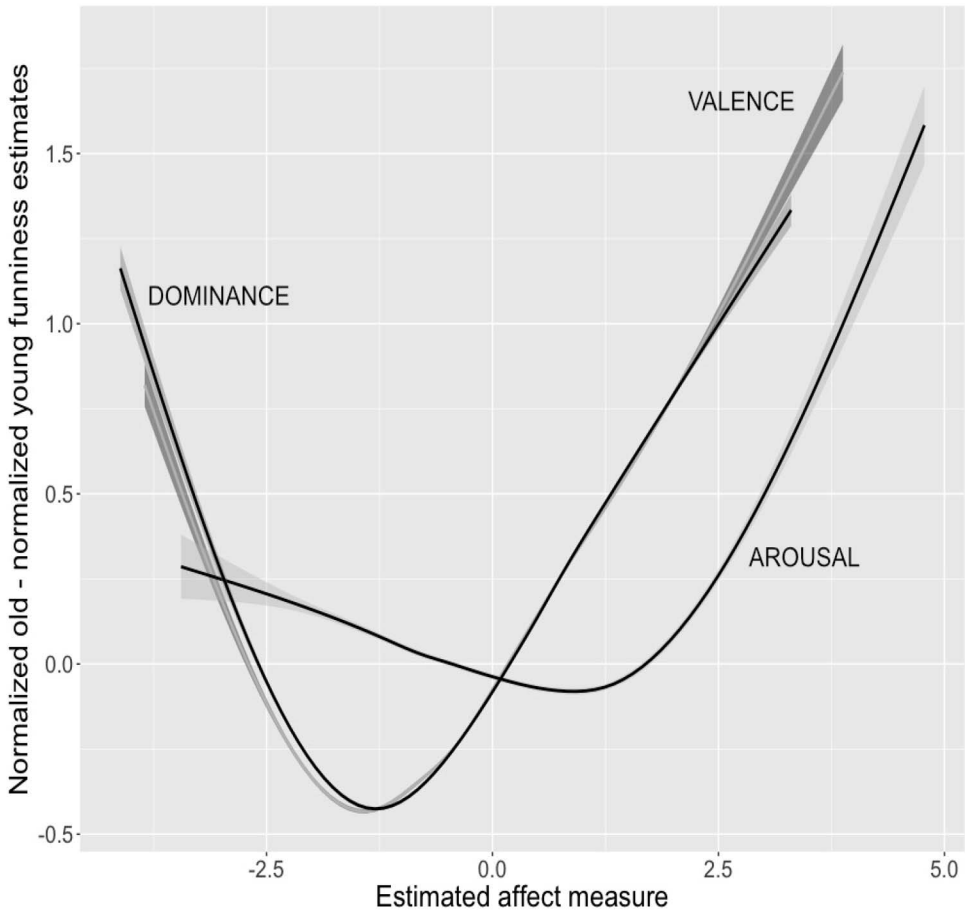


Figure 11. Difference between normalized funniness estimates from models derived from old and young judges (*x*-axis) by estimated affect measures (valence, arousal, and dominance; *y*-axis) for 45,516 words, fitted with generalized additive models and showing 95% confidence intervals.

linear equation is not. How the two equations perform on the same new dataset will be examined in the next section. Across all words for which we have estimates, the nonlinear equation estimates correlated with the linear regression estimates at $r = .94$ (95% CI [0.94, 0.94], $p < .001$), suggesting that access to nonlinearity does not add to prediction of humor.

When applied to the full dictionary, the 10 words predicted funniest by this equation are *slobbering, puking, fuzz, floozy, cackling, humping, fellated, bawl, mangy*, and *puss*. The funniest 200 words are provided in Appendix C. As marked in Appendices B and C, 95 (47.5%) of the 200 words estimated funniest by this model also appeared in the top 200 words estimated funniest by the linear model. The word predicted to be least funny was *disease*.

Discussion

This nonlinear equation serves as a double check on the linear equation, since it does not make any assumptions about the distributions or relationship between the predictors. The very high correlation between the estimates of the two models derived using quite different methods, and the high overlap of their 200 funniest words, serves as a validation of their estimates. Because that correlation is so high, we will not discuss the results here in detail, as we have already done above for the linear model.

Study 3: Model Validation

To validate the two models, we collected new estimates for 100 words from the Engelthaler and Hills (2017) norms, and 100 words that were not among those norms, both selected to systematically cover the range of estimated funniness.

Method

Best-worst scaling. Data were collected using a response format called best-worst scaling (Louviere, Flynn, & Marley, 2015). In best-worst scaling, participants are presented with a set of items and have to choose the superior (best) and inferior (worst) item in the set, along some described latent dimension. In this particular case, the latent dimension participants were instructed to make judgments over was humor. They were asked to pick the most and least humorous word. Four words were shown per trial. For example, participants might be shown the words *bullywug*, *orange*, *snark*, and *noodle* and instructed to choose the most and

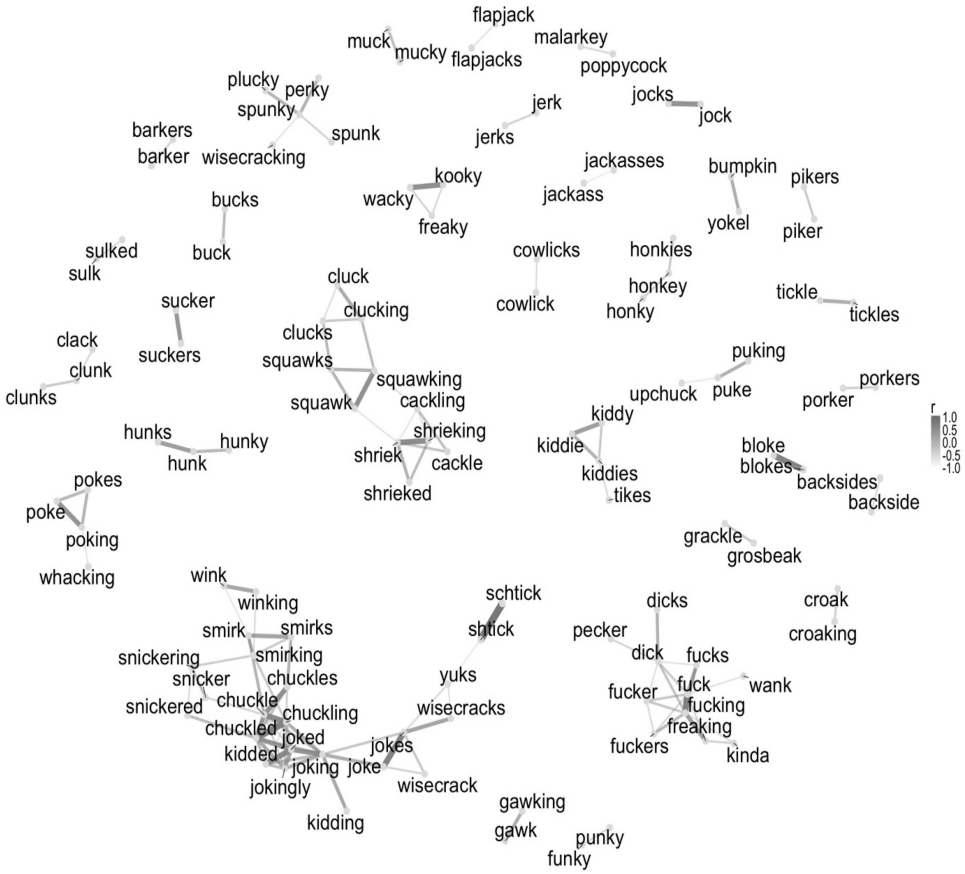


Figure 12. Correlation plot of the word2vec vectors of 116 words (39.8%) from the 291 words containing the letter *k* that were estimated funniest (>2 *SD*) using the model in Table 5 and that correlated with at least one other word above the threshold of $r = .5$. Line thickness and darkness indicates the strength of the correlation. Distance between unconnected clusters has no interpretation.

least humorous word in the set. Best-worst trials are constructed so that each word appears with a maximum number of other possible words across all trials (Louviere et al., 2015; Hollis, 2018), in order to maximize the amount of ordinal information available from participant judgments. Various algorithms are available for using this ordinal information to assign each word a point along an interval scale that approximates the proportional distances of words along the underlying latent dimension of interest (Hollis, 2018).

Norming efforts are typically carried out with rating scales, including the funniness norms collected by Engelthaler and Hills (2017). However, recent work has demonstrated that best-worst scaling has higher measurement reliability than rating scales for collecting semantic norms (Kiritchenko & Mohammad, 2017), that the format requires few trials to collect estimates that converge on true values (Hollis, 2018), and that best-worst scaling produces norms with predictive validity of lexical processing higher than or equivalent to rating scales and numeric estimation (Hollis & Westbury, 2018).

Latent values for individual words have typically been estimated from best-worst scaling by calculating the number of times a word is chosen as highest on the judged dimension, minus the number of times it is chosen lowest on the judged dimension, across all data points (e.g., Kiritchenko & Mohammad, 2016, 2017; Louviere et al., 2015). However, there are other ways of deriving estimates of latent values, most of which regularly require less data for estimates to converge on true values (Hollis, 2018). We use one such method, referred to as *value scoring*.

Suppose on a trial containing the four words, *bullywug*, *orange*, *snark*, and *noodle*, a person chooses *bullywug* as the funniest word and *orange* as the least funny word. From these two decisions, ordinal information is available for five of six possible pairs in the set of four items: *bullywug* is funnier than *orange*, *snark*, and *noodle*. *Snark* and *noodle* are funnier than *orange*. The only pair for which no information is available is *snark* and *noodle*. The value scoring method assigns an expected value to each word, informed by the other words to which it is superior or inferior. Each time a word is identified as superior to another word, its expected value is incremented toward a value of 1. Each time a word is identified as inferior to another word, its expected value is decremented toward a value of 0. In the above example, the expected value of *bullywug* would be incremented toward 1 three times, the expected value of *orange* would be decremented toward 0 three times, and the expected values of *snark* and *noodle* would

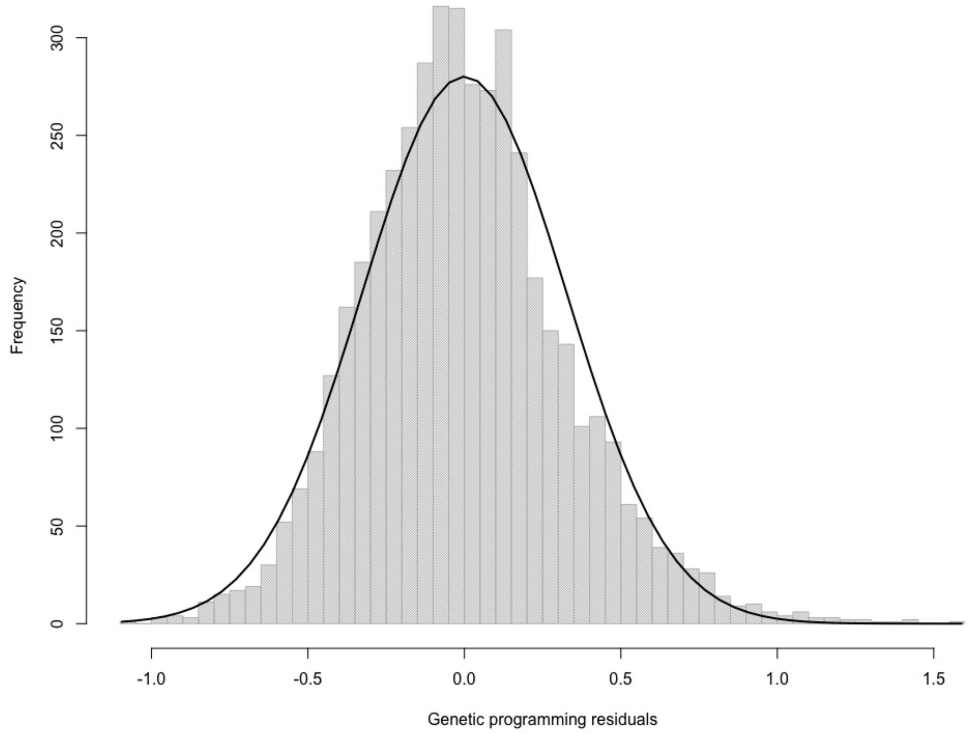


Figure 13. Distribution of residuals in predicting humor judgments using the genetic programming model.

be incremented toward 1 once and decremented toward 0 once. The size of each increment or decrement depends on how unexpected the results are, which is determined from expected values learned from previous trials. Readers are referred to Hollis (2018) for a complete introduction to best-worst scaling as it pertains to the collection of semantic norms, and the various ways by which latent values can be estimated from the response format.

Procedure. Participants were recruited via the University of Alberta undergraduate psychology research pool and compensated for their time in the form of 2% course credit toward an introductory psychology course. Participants were presented with 50 best-worst trials where they had to make humor judgments between four words. On each trial, participants were instructed to choose the most humorous and least humorous word out of the set of four. Each participant saw each word once across all 50 trials. Trial order was randomized. Trials were constructed according to best practice rules determined by Hollis (2018): each word occurred an equal number of times across all trials and the number of times that any two pairs of words occurred together was minimized.

Data were collected using three Apple G4 Macintosh Minis with 17.1-in. monitors. Screen resolutions were set to 1,280 × 1,024 and all text was presented in 36-point font. Words were ordered horizontally on the screen and centered. Selections of most and least humorous were made using two rows of radio buttons (one for best, one for worst) below the trial words, with each button centered under one of the four corresponding words. Responses were not timed, although participants were instructed to work at a rapid pace and, in the event of uncertainty, to respond with their first impression.

Participants. Data were collected from 74 participants (57 female). The mean [*SD*] age of female participants was 20.23 [2.81] and for male participants was 20.29 [2.03]. 91.89% of participants self-identified as speaking English as their first language. The remainder self-identified as speaking English for five or more years.

Participant compliance. In previous norming efforts using best-worst scaling, we have noticed that some participants produce noncompliant behavior by choosing items randomly or according to some other criterion that was inconsistent with the behavior of other participants. Hollis (2018) describes a procedure for identifying noncompliant participants based on their responses, compared with what would be expected from the responses of other participants. We applied this procedure to the current data. We found no evidence of noncompliant behavior from any of the participants.

Results

The results are presented graphically in Figure 14 (linear model) and Figure 15 (nonlinear model). The new human judgments were strongly correlated with the linear regression model predictions for both the 100 words from the Engelthaler and Hills (2017) norms ($r = .82$, 95% CI [0.75, 0.88], $p < .001$) and for the 100 words that had not been previously judged ($r = .84$, 95% CI [0.77, 0.89], $p < .001$). By Fisher's *r*-to-*z* test, these two correlations are not reliably different ($z = 0.45$, $p = .65$). The correlations from the nonlinear model were slightly lower: for the previously normed judged words, $r = .77$ (95% CI [0.67, 0.84], $p < 2.8 \times 10^{-11}$), and for the new words, $r = .76$ (95% CI [0.66, 0.83], $p < .001$). By Fisher's *r*-to-*z* test, these two correlations are also not reliably different ($z = 0.17$, $p = .86$).

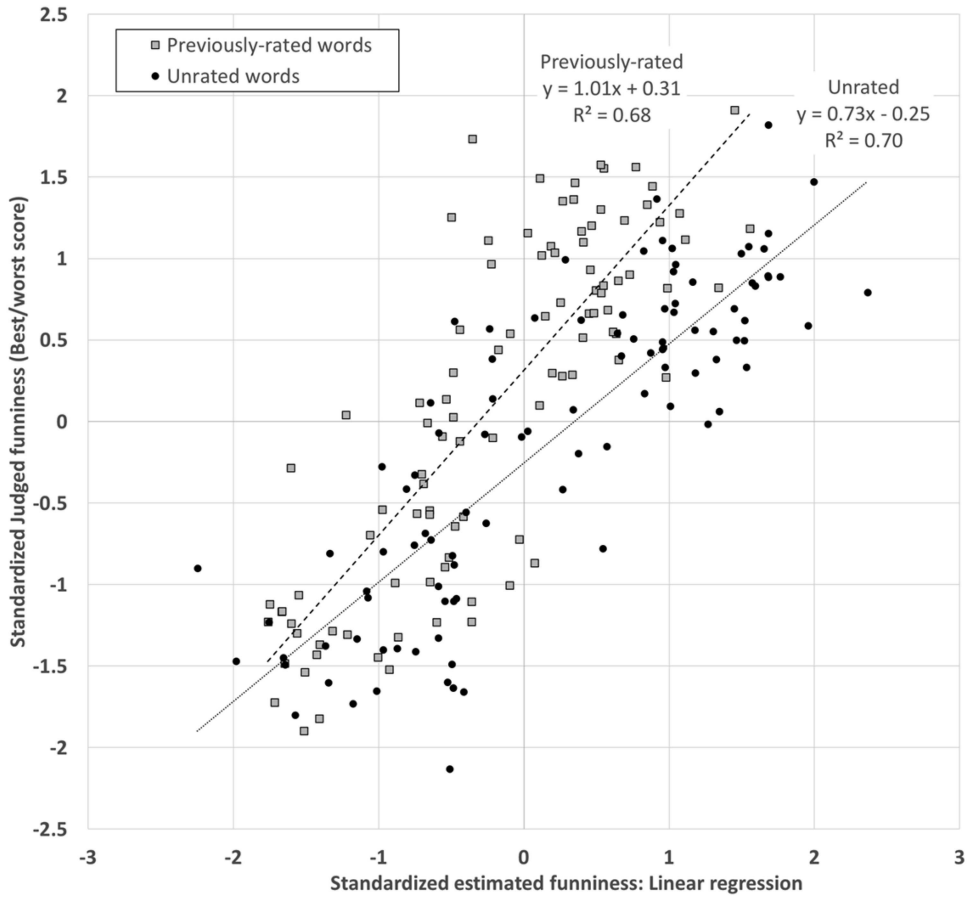


Figure 14. Estimates from the final linear regression model in Table 5, graphed against values derived from human best/worst judgments, for old and new stimuli.

Discussion

Since the gender and age splits showed only a small difference, we can use Engelthaler and Hills' (2017) male/female and young/old judgments to estimate their sample's split-half reliability. The 4,997 male/female judgments were correlated at 0.61, and the young/old judgments were correlated at 0.63. To directly compare with our experimental results, we estimated the equivalent correlations for the linear model over 100 pairs by taking 1,000 random samples of 100 pairs from each of the male/female judgments and the old/young judgments. The average (*SD*) correlation for 100 random items was $r = .60$ (0.07) for the male/female judgments and $r = .63$ (0.07) for the old/young judgments. The higher of these two estimates is reliably worse than the $r = .84$ obtained by our regression model (by Fisher's *r*-to-*z* test, $z = 3.34$, $p = .0008$), suggesting (if we accept the old/young and male/female splits as reasonable estimates of human split-half reliability) that our model is a more reliable judge than Engelthaler and Hills' (2017) human judges. This is perhaps because the model has access to very specific values for the relevant predictors, whereas humans presumably have to work with estimates of those values that are prone to error.

Conclusion

We have made the predictors and humor estimates for 45,416 words (along with the best/worst judgments for the 200 words judged) available at the American Psychological Association repository hosted by the Center for Open Science at https://osf.io/nphsd/.

We have been able to statistically model and extend judgment norms for word funniness with a quantifiably high degree of success. Our results paint a clear picture of the ideal funny word: it is a short, infrequent word composed of uncommon letters that is likely to include one or more specific letter or phonemes (/u/, consonant+*le*, *y*, and/or *k*), with an animate (or animacy-related) referent, especially one that is human and insulting, profane, diminutive, and/or related to good times.

One potential complication to our results is that the humor norms were collected in Britain, while our validation data were collected in North America. Although our models cross-validated well, words may have had their humor judgments misestimated due to large frequency differences between geographical regions. Since UseNet is international, the UseNet frequencies we used probably do not reflect strong regional biases, but thereby equally are not the best reflection of the British frequencies to which the

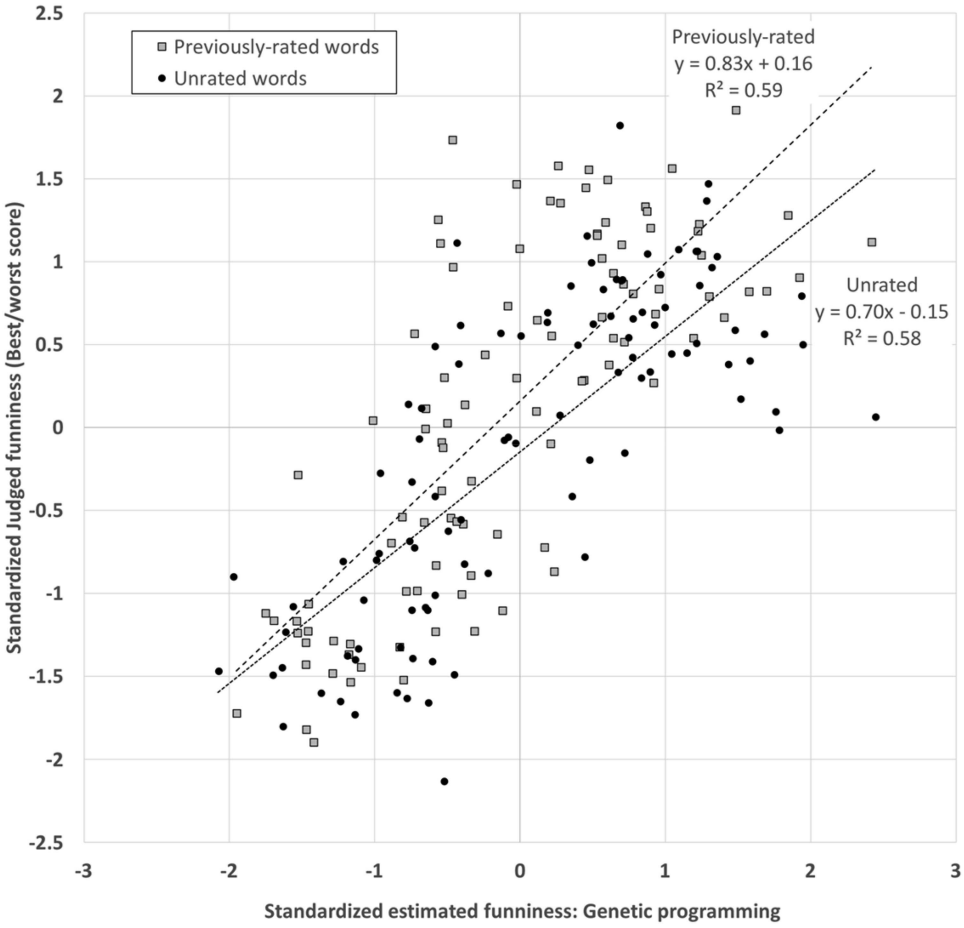


Figure 15. Estimates from the final nonlinear regression model, graphed against values derived from human best/worst judgments, for old and new stimuli.

original raters were presumably exposed. If we compare with the UseNet frequencies to United Kingdom Facebook frequencies (Herdağdelen and Marelli, 2017), which can be reasonably expected to be a better measure of British frequencies, we find there are many words that have very different frequencies. For example, words that occur over 100 times more frequently in the United Kingdom FaceBook frequencies include *granddad, tactically, brill* [short for *brilliant*, used by the British in roughly way North Americans use the word *cool*, to express general positive sentiment], *festive*, and *vodkas*. Although the differences were smaller among the words judged funniest (>2 *SD*), many of those words occurred more than 10 times more often in the United Kingdom FaceBook frequencies (e.g., *banter, fizz, bloke, sausage*, and *twat*) or more than 10 times as often in the UseNet frequencies (*nibble, gopher, hogwash, spam*, and *yahoo*). To assess the effects of these frequency differences, we reran the full linear model in Table 5 using the United Kingdom FaceBook frequencies instead of the UseNet frequencies, for the 4,418 judged words which appeared in the FaceBook frequency list. The new model was nearly identical to the model developed above (and therefore we have not reproduced it here): the same predictors and interactions all entered in, their 16 beta weights correlated at 0.99 (95% CI [0.99, 0.99], $p < .001$), and the estimates produced correlated with the estimates from the linear model produced using the UseNet frequencies at 0.98 (95% CI [0.98, 0.98], $p < .001$). It is predictable that words that are much more common in the UseNet frequencies than the United Kingdom ratings (e.g., *gopher, hogwash, oddball, swank*, and *gusher*) will have their funniness *under*-estimated by our original (see Table 5) model, since the ratings are from the United Kingdom and less common words are funnier. By the same reasoning, words that are much more common in the United Kingdom Facebook frequencies than the Usenet ratings (e.g., *bugger, knickers, bollocks, twat*, and *bloke*) are likely to have their funniness *over*-estimated by that model, since the UseNet frequencies assign lower frequencies than the British judges experienced. However, this must necessarily be a small effect according to the model. With a beta weight of just −0.04 in the linear model, logged frequencies have an effect bounded between about −0.46 and 0.23 on funniness estimates that range, across all words for which we have regression estimates, between 2.6 and 7.9.

In our introduction, we noted that theories of humor are scientifically dissatisfying because they generally serve as post hoc descriptions of observed humorous phenomena, rather than being quantified theories that allow for prediction of experienced funniness. In the narrow domain of single-word humor to which we have confined ourselves, we have achieved consid-

erable success in predicting which words would be judged funny, presenting suggestive evidence that we have accounted for as much variance as can be accounted for by human judgments.

This predictive success helps to make clear why a single simple 'theory of humor' is unlikely to be achieved, except in broadest terms. Our model of an extremely simple form of humor is simultaneously consistent with several major theories of humor (superiority theory; incongruity theory and its various relations; and Bergson's juxtaposition theory) and reflects a high degree of complexity, synthesizing and weighting information from multiple independent sources: affective measures in interaction, lexical semantics, word and subword frequency measures, word length, phonological form, and perhaps individual human characteristics such as gender and age.

Though we have not done so (since it would amount to no more than changing the units on the beta weights) it would be easy to convert all our predictors to probability measures. For example, distance from AVERAGE-CDV could be expressed as 'probability of having a funny meaning' by simply converting it to a percentile, and so on for all our measures. We believe that this is a reasonable way to conceive of the predictors since it reduces humor to a single, though very broad, measure: predictability. However, thinking of all our predictors as probability measures does not clarify how we might systematically extend the prediction to more complex stimuli. Even something as simple as a one-word pun introduces a complex set of additional probabilities that are not relevant to single-word humor. How likely are each of the words, both in their form and their probability of occurrence? How likely is it that the confusion between the two words would be made? How likely is it that the two domains to which the two words refer would be united into a single context? How likely is that the two words would occur in close proximity in ordinary language use? More complex jokes introduce even more complex probabilities. How often *do* a priest, a rabbi, and a Buddhist monk walk into a bar, anyway?

Notwithstanding these complexities, it is easy to envisage a small step forward to quantitative study of what may be the world's *third* least funny jokes, *word pairs*. With humor estimates for the individual words now computed, it is possible to start studying how those values interact with each other and with other measures to affect humor judgments of potentially humorous word pairs such as *toothy weasel*, *muzzy muffin*, and *fizzy turd*. These pairs introduce a small number of manageable complications to the study of simplified humor.

Despite the challenges we face in extending this work into realistically complex domains of humor, we believe that the step we have made here toward full quantification of the elements of single-word humor is an important one. Analysis of highly simplified models have almost always been the first step in scientific progress. In analyzing the humor of single words, we have built very directly on (and replicated) earlier work by Westbury et al. (2016) that looked at even simpler "jokes", the humor of nonword strings. It was only the success of that work that encouraged us to extend the methodology a little further into real words. We hope the next steps may build on this work to make quantified predictions about still funnier jokes.

Context Paragraph

The work we describe in this paper builds directly on earlier work (Westbury et al., 2016) that showed that humor ratings in nonwords could be well predicted from the statistical structure of the nonword strings. We were inspired to undertake that earlier work because we noticed that experimental participants sometimes laughed at nonword strings used in our lexical decision experiments, and often at the same strings. We work primarily on studying semantics, with a particular interest in what analysis of the structure of co-occurrence models can tell us about the basic dimensions of semantic processing. Replicating and extending earlier work that we admire very much by Osgood et al., 1957/1978, our work on deconstructing lexical semantics (Westbury, 2014; Hollis & Westbury, 2016) has shown a strong role for affect in lexical semantics. Our interests in statistical and affective processing in lexical access and in quantitative models of semantics intersect on humor.

References

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review, 118,* 438–481. http://dx.doi.org/10.1037/a0023851

Bergson, H. (2009). *Laughter: An essay on the meaning of the comic* (C. Bereton & F. Rothwell, trans.). Retrieved from https://www.gutenberg.org/cache/epub/4352/pg4352.txt (Original work published 1900)

Berlyne, D. E. (1971). *Aesthetics and psychobiology*. New York, NY: Appleton-Century-Crofts.

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

The Economist. (2017, May 11). Interview with Donald Trump. *The Economist Group Limited*. Retrieved from https://www.economist.com/Trumptranscript

Engelthaler, T., & Hills, T. (2017). Humor norms for 4,997 English words. *Behavior Research Methods, 50,* 1116–1124.

Eysenck, H. J. (1942). The appreciation of humour: An experimental and theoretical study. *British Journal of Psychology, 32,* 295–309.

Godkewitsch, M. (1972). The relationship between arousal potential and funniness of jokes. In J. H. Goldstein & P. E. McGhee (Eds.), *The psychology of humor: Theoretical perspectives and empirical issues* (pp. 143–158). New York, NY: Academic Press. http://dx.doi.org/10.1016/B978-0-12-288950-9.50013-4

Hazlitt, W. (1907). *Lectures on the English comic writers*. New York, NY: Oxford University Press. (Original work published 1819)

Herdağdelen, A., & Marelli, M. (2017). Social media and language processing: How Facebook and Twitter provide the best frequency estimates for studying word recognition. *Cognitive Science, 41,* 976–995. http://dx.doi.org/10.1111/cogs.12392

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

Hollis, G., & Westbury, C. (2006). NUANCE: Naturalistic University of Alberta nonlinear correlation explorer. *Behavior Research Methods, 38,* 8–23.

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review, 23,* 1744–1756. http://dx.doi.org/10.3758/s13423-016-1053-2

Hollis, G., & Westbury, C. (2018). When is best-worst best? A comparison of best-worst scaling, numeric estimation, and rating scales for collec-

tion of semantic norms. *Behavior Research Methods, 50,* 115–133. http://dx.doi.org/10.3758/s13428-017-1009-0

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,* 1603–1619. http://dx.doi.org/10.1080/17470218.2016.1195417

Hollis, G., Westbury, C. F., & Peterson, J. B. (2006). NUANCE 3.0: Using genetic programming to model variable relationships. *Behavior Research Methods, 38,* 218–228.

Hutcheson, F. (1750). *Reflections upon laughter, and remarks on the fable of the bees*. Glasgow, Scotland: R. Urie. Retrieved from https://books.google.ca/books?id=xuAtAAAAYAAJ

Kant, E. (1951). *Critique of judgment*. New York, NY: Hafner. (Original work published 1790)

Keith-Spiegel, P. (1972). Early conceptions of humor: Varieties and issues. In J. H. Goldstein & P. E. McGhee (Eds.), *The psychology of humor: Theoretical perspectives and empirical issues* (pp. 4–39). New York, NY: Academic Press.

Kierkegaard, S. (1941). *Concluding unscientific postscript* (D. Swenson & W. Lowrie, Trans.). Princeton, NJ: Princeton University Press. (Original work published 1846)

Kiritchenko, S., & Mohammad, S. M. (2016). Capturing reliable fine-grained sentiment associations by crowdsourcing and best-worst scaling. In K. Erk & N. A. Smith (Eds.), In *Proceedings of the 15th annual conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 811–817). Vancouver, British Columbia, Canada: Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/N16-1095

Kiritchenko, S., & Mohammad, S. M. (2017). Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In R. Barzilay & M. Kan (Eds.), In *Proceedings of the annual meeting of the association for computational linguistics* (pp. 465–470). Vancouver, British Columbia, Canada: Association for Computational Linguistics.

Koestler, A. (1964). *The act of creation*. London, UK: Hutchinson.

Kohs, G. (Director). (2017). *AlphaGo* [Motion picture]. New York, NY: Reel As Dirt & Moxie Pictures.

Koza, J. R. (1992). *Genetic programming II: Automatic discovery of reusable subprograms*. Cambridge, MA: MIT Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211–240. http://dx.doi.org/10.1037/0033-295X.104.2.211

Louviere, J. J., Flynn, T. N., & Marley, A. A. J. (2015). *Best-worst scaling: Theory, methods and applications*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781107337855

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers, 28,* 203–208. http://dx.doi.org/10.3758/BF03204766

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

Mencken, H. L. (1948, September 25). The Podunk mystery. *The New Yorker* (pp. 75–81). Retrieved from http://archives.newyorker.com/?i=1948-09-25#folio=074

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *The Computing Research Repository*. Advance online publication. Retrieved from https://dblp.uni-trier.de/db/journals/corr/corr1301.html

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems: Neural information processing systems conference* (pp. 3111–3119). Lake Tahoe, Nevada: Neural Information Processing Systems (NIPS). Retrieved from http://www.proceedings.com/21521.html

Mikolov, T., Yih, W. T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In P. Fung & M. Poesio (Eds.), In *Proceedings of the annual conference of the North American chapter of the association for computational linguistics: Human language technologies* (Vol. 13, pp. 746–751). Retrieved from http://www.aclweb.org/anthology/N13-1090

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1978). *The measurement of meaning*. Urbana: University of Illinois Press. (Original work published 1957)

Ramachandran, V. S. (1998). The neurology and evolution of humor, laughter, and smiling: The false alarm theory. *Medical Hypotheses, 51,* 351–354. http://dx.doi.org/10.1016/S0306-9877(98)90061-5

Schopenhauer, A. (1907). *The world as will and idea* (6th ed., R. B. Haldane & J. Kemp, Trans.). London, UK: Routledge & Kegan Paul. (Original work published 1818)

Shaoul, C., & Westbury, C. (2006). *USENET orthographic frequencies for 1,618,598 types (2005–2006)*. Edmonton, AB: University of Alberta. Retrieved from http://www.psych.ualberta.ca/~westburylab/downloads/wlallfreq.download.html

Stark, R. (Producer), & Ross, H. (Director). (1975). *The sunshine boys* [Motion picture]. Los Angeles, CA: Metro-Goldwyn-Mayer.

Westbury, C. (2014). You can't drink a word: Lexical and individual emotionality affect subjective familiarity judgments. *Journal of Psycholinguistic Research, 43,* 631–649. http://dx.doi.org/10.1007/s10936-013-9266-2

Westbury, C., Hollis, G., Sidhu, D. M., & Pexman, P. M. (2018). Weighing up the evidence for sound symbolism: Distributional properties predict cue strength. *Journal of Memory and Language, 99,* 122–150. http://dx.doi.org/10.1016/j.jml.2017.09.006

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. http://dx.doi.org/10.1016/j.jml.2015.09.001

Wittgenstein, L. (1953). *Philosophical investigations*. London, UK: Blackwell.

(*Appendices follow*)

Appendix A

The 50 Words With Most Similar Vectors to the Category-Defining Vectors for Six Categories Associated With Judged Funniness, in Decreasing Order of Similarity

Party	Sex	Insults	Profanity	Body function	Animals
banter	pubes	moron	shit	puke	critter
joke	boob	idiot	fuck	puss	puppy
giggle	pussy	cretin	moron	pussy	reptile
chuckle	crotch	buffoon	damn	pubes	feline
laughter	pecker	jackass	twat	butt	pooch
antics	cunt	twat	fucker	crotch	raccoon
chitchat	dildo	bastard	bitch	slobber	bobcat
shindig	penis	weirdo	bullshit	pecker	lizard
dinner	panties	dude	dude	cunt	tortoise
ditty	twat	douche	bastard	snot	possum
soiree	knickers	slob	douche	vomit	collie
bash	hussy	fucker	idiot	twat	terrier
frolic	butt	bitch	piss	penis	monkey
lark	fuck	chump	cunt	shit	piglet
booze	douche	whore	stupid	turd	chimp
supper	cleavage	ninny	jackass	fucker	otter
buddy	fucker	blockhead	puke	anus	giraffe
brunch	babe	cunt	turd	douche	hippo
toast	puss	stupid	bollocks	scrotum	cheetah
goof	booby	jerk	whore	tummy	rooster
grub	slut	turd	bugger	burp	toad
puke	tummy	shit	cretin	fuck	mutt
smirk	dude	thug	pussy	boob	snake
rowdy	chick	nimrod	weirdo	backside	poodle
clown	bitch	rascal	dumb	spit	goat
fondue	shit	fuck	slut	dildo	mammal
jester	whore	bigot	bleep	belly	panther
patter	puke	bullshit	commie	pimple	zebra
riff	panty	dumb	jerk	mouth	parrot
floozy	floozy	commie	nimrod	buttocks	spaniel
boogie	sweetie	floozy	snot	bitch	tiger
prank	kisser	coward	chump	oink	baboon
trivia	pants	slut	slob	cretin	bulldog
glee	belly	liar	bloke	piss	turtle
romp	hubby	shyster	darn	snout	bird
gumbo	prude	groupie	honky	dude	tomcat
holler	buttocks	hick	pubes	panties	rhino
polka	backside	hussy	joke	bleep	frog
silly	fetish	bloke	heck	lick	python
suds	garter	kidder	ninny	cranium	leopard
sweetie	kiss	lackey	scum	stink	falcon
chatter	tampon	loony	buffoon	forehead	dingo
threesome	nude	weasel	wont	pants	skunk
groupie	vixen	grouch	drivel	gullet	wolf
buffoon	scrotum	bugger	butt	ninny	boar
chat	blonde	egghead	hussy	snort	buzzard
dance	groupie	snob	blockhead	toilet	beaver
dude	cock	fool	sucker	prick	panda
jest	tights	chap	silly	nostril	penguin
grin	anus	puke	chick	damn	canine

(*Appendices continue*)

Appendix B

Two Hundred Words Estimated to Be Funniest by the Final Linear Regression Model in Table 5, in Descending Order

upchuck*	puke*+	jiggling*	slaphappy	mumbo*	gawky
bubby*	bawl*	huffy*	farted*	momma*+	pixy
boff*	chucky*	simp*	how're	coxcomb	gabby
wriggly	cluck*	guff	weeny*	doggies	trow
yaps	yack	jollying	squawk*+	accusingly	ducky+
giggle*+	booby*+	punkin	schmooze	sylph	cutty
cooch*	frowzy	cuddle+	chubby*	bitsy	booger*
guffaw*	giggling*	titties*	bogart*	guck	nincompoop*
puffball	backslap	barfed*	heinie*	pooped	clinger*
jiggly	wienie	fuck+	kiddies	yummy	farts*
squiffy	cavort	grizzle	bozo*	mamma	scamp
cooky	boobs*	whoop*	pudgy*	unglue	hubby+
poop*	slobbering*	waggling	floozy*+	puking*	buffy
flappy*	fellating*	wiggling*	clucking	guppy*	plumy
bucko*	tike	frisky*	cute	giggles*	shaggily
twirly	prancing*	guffaws	boobies*	huck	puss*+
lubber	snots	huffs	floozie	snicker	muff+
frumps	dingle	fanny*	hirsute	farting*	fellers
waddle*+	burp*+	cheerios	cooing	hucks	cavorting
humping*	goofy	tush	foolery*	porky	puff
wiggly	fuzz*+	drool*	licker*	buzz*	holler+
humph*	jitterbugging	pooping*	shimmy	widdle*	weiner*
hussies*	boob*+	braw	hussy*+	bluejay+	squealing*
lummox	pubes*+	gulp*+	grump	tubby*	jowls
yuks	cowlicks	foxy*	hoot	gabbing	squish
bulgy	ponces	clucks	loll	wham*	
poppa	prance+	pukes*	nuzzle	dumpy*	
poops*	buxom*	cackling*	diddle*	yobbo*	
tiddly	whoopee	schmuck*	youngs	pecker*+	
slobber*+	blowzy	teethe	titter	shucks	
giggly*	titty*	muzzy*	simper	cuddling	
fellate*	wank*	groupies*	goddamn*	chortles*	
crumby	chortle*	blowjob*	jiggle*+	nilly*	
bunghole*	goos	girlishly	pretties	mangy*	
				skulk	
				prances	
				wags*	
				bally	
				fluff+	

Note. The 95 (47.5%) words marked with an asterisk are common to the funniest 200 lists of both the linear and nonlinear models (see Appendix C). The 26 (13%) words marked with a "+" were part of the original Engelthaler & Hills (2017) set rated by humans.

(*Appendices continue*)

Appendix C

Two Hundred Words Estimated to Be Funniest by the Best Nonlinear Regression Model, in Descending Order

slobbering*	weiner*	groupies*	honky+
puking*	burp*+	minx	quip
fuzz*+	buzz*	jumpy	weirdo+
floozy*+	squawking	snickering	boobies*
cackling*	goddamn*	hijinks	crawly
humping*	letch	wiggle+	shagging
fellated	schmuck*	booby*+	masturbator
bawl*	crapping	foolery*	yobbo*
mangy*	pubes*+	boff*	chump+
puss*+	guffaw*	clinger*	bleep+
pukes*	boob*+	masturbators	fornicated
meany	boobs*	barfed*	willy
pooping*	foxy*	wiggling*	cunts
nymphomaniacs	farted*	tipsy	frigging
bunghole*	boner	huffy*	giggly*
titties*	blowjob*	rotter	prats
titty*	jiggling*	poops*	squirming
strumpet	puke*+	nincompoop*	crybaby
chucky*	farting*	winker	dicks
upchuck*	fellate*	buxom*	catcall
licker*	cackle	jiggle*+	ponce
twerp	chubby*	effing	ogress
cootie	blurt	scarry	flabby
wank*	prancing*	fellating*	squawk*+
booger*	pudgy*	heinie*	mumbo*
giggle*+	finks	groupie+	what'd
giggling*	dumpy*	squealing*	killjoy+
slobber*+	fucker+	chick+	mouthy
ninny+	biff	chomp	jackass+
conniption	belcher	douche+	weeny*
hussy*+	pecker*+	farts*	smirking
snogging	poop*	nilly*	bubby*
cooch*	fanny*	douches	peed
simp*	honkey	hussies*	bucko*
humph*	weirdos	fornicate	flamer
dippy	smarty	diddle*	smirks
lout	tyrannosaurus	waddle*+	fathead
giggles*	rumpus	guppy*	drool*
flatulent	nymphomaniac	bellyache	duper
gulp*+	bacchanalia	prissy	malaprop
jollies	klutz	gummy	momma*+
chortle*	wags*	bozo*	crabby
twats	mamba	bogart*	dopey
whoop*	muzzy*	legless	squirm+
blurts	coitus	flappy*	busty
biddies	oink+	twat+	shriek+
gawk	frisky*	widdle*	chomping
cluck*	bray	how'd	wham*
pouty	pygmy	juju+	tubby*
chortles*	boozy	shamus	jock

Note. The 95 (47.5%) words marked with an asterisk are common to the funniest 200 lists of both the linear and nonlinear models (see Appendix B). The 34 (17%) words marked with a "+" were part of the original Engelthaler & Hills (2017) set rated by humans.

(*Appendices continue*)

Appendix D

Genetic Programming Explained

Genetic programming (GP) is a method of function approximation that works using principles of natural selection. Natural selection operates over a population when three features are present: variation among the population, inheritance of varying trait(s), and a selection criterion. For example, natural selection operates over height because there is variation in height, height is heritable to some degree, and there are social and environmental factors that make individuals of some heights more likely to survive and reproduce than individuals of other heights. The basic GP algorithm is defined by the following steps:

1. Seed an initial population

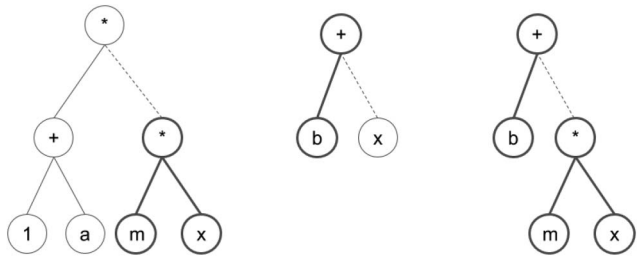
While termination condition has not been met, repeat the following steps:

2. Evaluate population members against selection criterion
3. Remove members from the population that do not meet selection criterion
4. Produce offspring based on remaining members until population is replenished
5. Check if termination condition has been met

When using genetic programming for function approximation, the selection criterion is often the degree to which a member of the population (which is just a function) minimizes mean squared error over a set of data or, alternatively, maximizes *R*-squared over the data. Below (or above) some threshold, a population member is selected for survival and reproduction. The algorithm terminates once a population member is produced that approximates the desired function with sufficient accuracy, or once a specified number of iterations pass without producing one such population member. Population members are represented as trees where terminal nodes are constants and variables, and nonterminal nodes are mathematical operators. Offspring are produced by combining branches from each of their parents, sometimes introducing a mutation to introduce new variability into the population.

Consider the case of attempting to approximate a functional relationship between some response variable, *y*, and an explanatory variable, *x*. Two constants, *m* and *x*, are thought to be relevant. Suppose we know that the functional form of this relationship is $y = mx + b$. GP would approach this problem by first creating a population of random functions out of the supplied constants, explanatory variables, and mathematical operations the user thought might be relevant for solving the problem. Each population member would be evaluated based on how well it approximated the function specified by *y* by, e.g., correlating output of the population member with *y*. Members of the population would then be selected, based on how well they approximate the function specified by *y*. Selected members would then be mated to produce offspring—typically, by randomly selecting subbranches of each parents' tree and recombining those branches. Offspring would then be introduced into the selected population and the process would repeat until one member was found that sufficiently approximated the desired function.

An example of mating between two parents is provided in the figure below. Subbranches from two parent trees are randomly chosen (bolded). They are then combined to form a new tree along the vertex which each subbranch was pruned (dotted lines).



Received September 12, 2017
Revision received May 2, 2018
Accepted May 4, 2018 ■