Journal of Memory and Language 86 (2016) 141–156

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Telling the world's least funny jokes: On the quantification of humor as entropy

Chris Westbury [a,*], Cyrus Shaoul [b], Gail Moroschan [a], Michael Ramscar [b]

[a] Department of Psychology, University of Alberta, P577 Biological Sciences Building, T6G 2E9 Edmonton, AB, Canada
[b] Department of Linguistics, University of Tübingen, Wilhelmstrasse 19-2, 72074 Tübingen, Germany

ARTICLE INFO

Article history:
Received 2 March 2015
revision received 18 July 2015
Available online 6 October 2015

Keywords:
Humor
Non-words
Information theory
Lexical access
Shannon
Schopenhauer

ABSTRACT

In assessing aphasics or conducting experiments using a lexical decision task, we have observed informally that some non-words (NWs) reliably make people laugh. In this paper, we describe a set of studies aimed at illuminating what underlies this effect, performing the first quantitative test of a 200 year old theory of humor proposed by Schopenhauer (1818). We begin with a brief overview of the history of humor theories. Schopenhauer's theory is formulated in terms of detection/violation of patterns of co-occurrence and thereby suggests a method to quantify NW humor using Shannon entropy. A survey study demonstrates that there is much more consistency than could be expected by chance in human judgments of which NWs are funny. Analysis of that survey data and two experiments all demonstrate that Shannon entropy does indeed correctly predict human judgments of NW funniness, demonstrating as well that the perceived humor is a quantifiable function of how far the NWs are from being words.

Introduction

In the course of rejecting both extreme sensationalism and extreme cognitivism, William James (1890/1950) wrote that: "We ought to say a feeling of *and*, a feeling of *if*, a feeling of *but*, and a feeling of *by*, quite as readily as we say a feeling of blue or a feeling of cold" (pp. 245–246). Using a word is as much a matter of *feeling* as it is of *thinking*. Words don't just have different semantic and syntactic properties; they also *feel* different. Perhaps James chose to use as examples only closed class words, with 'bleached semantics', to emphasize that this 'feeling system' might have a particularly clear role to play when cognitive semantics does not. Faced with diminished competition from semantics, the feeling evoked by a letter string might be freed to play a stronger role. In these studies we take this idea to its limit, by focusing on the feeling evoked by *non-word strings* (NWs), which have even more bleached semantics than closed class words (for evidence that NWs do sometimes have *some* semantics, see Reilly, Westbury, Kean, & Peele, 2012; Westbury, 2005). We present evidence showing that some NWs do reliably evoke feelings of humor in readers. Based on extant models of humor, reviewed in the next section, we are able to use Shannon entropy to manipulate and predict the amount of humor evoked by novel meaningless strings. The results have implications for understanding both humor and language processing.

How does humor function?

Although there are many theories of humor and no final consensus on what makes something funny, one clear

* Corresponding author. Fax: +1 780 492 1768.
E-mail address: chrisw@ualberta.ca (C. Westbury).

http://dx.doi.org/10.1016/j.jml.2015.09.001
0749-596X/

thread that informs many post-Platonic theories of humor is of particular relevance here.[1] This is the idea that humor involves the recognition of some specific forms of incongruity, the improbable and therefore surprising co-occurrence of two or more ideas and/or events.

Although alluded to by Aristotle, this idea has its modern roots in Francis Hutcheson's (1725/1973) essays *Reflections on Laughter*, which were first printed in *The Dublin Journal* in 1725. Hutcheson argued that humor was based on "the perception of an incongruity between something dignified and something mean" (Telfer, 1995, p. 360). He would have appreciated the nattily dressed hobo made famous by Charlie Chaplin, whose juxtaposition of dignified dress and undignified behavior exactly conforms to what Hutcheson had in mind as humorous.

In his (1865/2004) discussion of this idea of 'ludicrous incongruity', the psychologist Alexander Bain pointed out the deficiencies in Hutcheson's simplistic view, noting in a much cited passage that:

> "There are many incongruities that may produce anything but a laugh. A decrepit man under a heavy burden, fives loaves and two fishes among a multitude, and all unfitness and gross disproportion; an instrument out of tune, a fly in ointment, snow in May, Archimedes studying geometry in a siege, and all discordant things; a wolf in sheep's clothing, a breach of bargain, and falsehood in general; the multitude taking the law into their own hands, and everything of the nature of disorder; a corpse at a feast, parental cruelty, filial ingratitude, and whatever is unnatural; the entire catalogue of vanities given by Solomon,— are all incongruous, but they cause feelings of pain, anger, sadness, loathing, rather than mirth." (pp. 256–257).

Bain was apparently not familiar with the philosopher Arthur Schopenhauer's (1818/1883) *The World As Will and Representation* (which was not translated into English until the 1880s). In the course of writing about the limitations of reason and deliberation, Schopenhauer sharpened the notion of exactly what kind of detected incongruity would be found humorous. He stated that "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*" (1818/1883, p. 76, emphasis added). Schopenhauer thereby proposed that it was not incongruity *per se*, but only incongruity that plays into an *a priori* conceptual expectation that is funny: i.e. an unexpected dissociation between *an event* and *an idea about that event*. This simple rider eliminates most of the counter-examples listed by Bain, since most of them are incongruous but not *expectation violating*. We may not often encounter instruments out of tune or a fly in our ointment, but we would not violate any conceptual expectation if we did.

Perhaps because his writing on the matter is somewhat turgid, it is not always fully appreciated (see, e.g. Martin, 1983, who appears to have stopped reading Schopenhauer at the sentence above) that Schopenhauer goes on to further specify that there are two particular related forms of perceived incongruity that are funny. One form involves a *conceptual bifurcation* (our own term): the realization that a concept that had been seen as belonging to a single category actually belongs to two categories simultaneously. The other involves the opposite realization, of *conceptual subsumption*: the realization that two apparently different concepts can be subsumed under a single category. Schopenhauer wrote:

> "It often occurs in this way: two or more real objects are thought through *one* concept, and the identity of the concept is transferred to the objects; it then becomes strikingly apparent from the entire difference of the objects in other respects, that the concept was only applicable to them from a one-sided point of view. It occurs just as often, however, that the incongruity between a single real object and the concept under which, from one point of view, it has rightly been subsumed, is suddenly felt. Now the more correct the subsumption of such objects under a concept may be from one point of view, and *the greater and more glaring their incongruity with it [...] the greater is the ludicrous effect which [sic] is produced by this contrast*. All laughter then is occasioned by [...] *unexpected subsumption*." (pp. 76–77, emphasis added)

There are two important aspects to Schopenhauer's theory, one of which we have already emphasized: that Schopenhauer defines humor in terms of *patterned co-occurrence*. It is not simply the low frequency of an event that is humorous; it is the probability of co-occurrence of an event with a pre-existing inconsistent expectation. The second important claim made by Schopenhauer is that jokes are funnier the more they are they are incongruous.

Puns and other word play provide a most obvious example of Schopenhauer's point. The pun *When the clock is hungry it goes back four seconds* is funny because of the unexpected dual meanings of *four* [*for*] and *seconds*. Schopenhauer's theory predicts what many of us would agree with: that it would be less humorous (because it is simply confounding and therefore mildly annoying) to make a very similar statement that does not so cleanly violate a *specific* conceptual expectation (although it may still violate our expectations of lexical co-occurrence): e.g. "When the clock is hungry, it rides a horse". Schopenhauer is surely right in suggesting that the resolution inherent in recognizing *which specific expectation has been violated* by a joke is an important element in it being 'a good joke'.[2] Consider, for example, the generally unpleasant feeling of 'not getting a joke'. When we recognize the anomaly in knowing that a joke has been made, but fail to identify the precise expectation that has been violated, we experience the

[1] Plato (as well as Aristotle and, later, Thomas Hobbes) speculated on the origins of humor but they focused largely on derisive humor, in which a person takes pleasure in the perceived deficiencies of another person. Hutcheson pointed out that many instances of disparity between two people were not funny, and that many funny things did not involve any perceptible personal disparity.

[2] This may be one reason why the NW string 'jokes' that we consider in this paper are not very funny.

unpleasant emotions of bewilderment and embarrassment rather than the pleasant feeling of humor. We have more to say about the possible relationship between a sudden resolution of negative feelings and the experience of humor in the final discussion.

Schopenhauer's specification that an unexpected conceptual bifurcation or subsumption relation be present in any funny incongruity rules out the last few of Bain's contending counter-examples to the congruity theory, and many other counter examples listed elsewhere (e.g. Clark, 1970). While we certainly do not expect to encounter a corpse at a dinner party, if we did, we would be unlikely to be struck by the idea that the dinner party and the corpse had been subsumed by a single concept. A corpse bound hand and foot wearing a tie-dye shirt: that would be funny.

Versions of expectation violation theory have been proposed several times in the last few decades (e.g. Attardo & Raskin, 1991; Suls, 1972). It was recently sharpened again in Hurley, Dennett, and Adams (2011), who further emphasized the need for what they call "epistemic conflict" (i.e. a need to resolve a belief that is inconsistent with another, consistent with the *conceptual unexpectedness* that had been emphasized by Schopenhauer) in the course of arguing that one possible evolutionary function of humor might be to reward an organism for checking the epistemological consistency of the contents of mind. This notion is (perhaps) somewhat in conflict with that put forward by Chafe (2007), who in contrast emphasized that the non-seriousness that underlies humor appreciation is "a reaction to situations *that it would be counter-productive to take seriously*, with the result that they are rejected as candidates for inclusion in one's repertoire of knowledge" (p. 13; emphasis added). Perhaps Hurley, Dennett, and Adams would accept that belief checking only feels *amusing* when a belief is recognized as something that would be counter-productive to take seriously. After all, a belief that is worth taking seriously (i.e. because it is has some practical utility) is presumably its own reward.

The expectation violation theory is not the only theory of humor,[3] and the forms specified by Schopenhauer may not be the only kind of funny incongruities (for a detailed discussion of the perhaps surprising range of humor that does fall under the theory, see Hurley et al., 2011). However, as well as being a dominant theory, the expectation violation theory of humor as stated by Schopenhauer has a particular attraction for our current investigation into the slight humor of NWs, which is that *the qualities relevant to humor in NW strings are mathematically well-defined*. This allows us, uniquely, to check Schopenhauer's claim that the greater the incongruity, the greater is the ludicrous effect it produces, without having to worry about the difficult-to-quantify effects of semantics or of the real world co-occurrences in jokes. It is after all hard to accurately estimate just how often a priest, a rabbi, and Buddhist monk *do* walk into a bar.

[3] Though it is quite consistent with a theory not otherwise discussed here, Freud's (1905/1960) theory of humor as a release (which was first proposed, without the psychoanalytic context, by Bain (1865/2004)).

Concretely, the idea of the humorous as expectation violation naturally fits in with the information theoretic formalization that was defined 130 years after Schopenhauer's book was published. Shannon (1948) formalized surprise as an inverse measure of probability by defining the measure now known as *self-information* or *Shannon surprisal*, expressed (here, in bits) for any event as $-\log_2(p(\text{event}))$. In a series of n events, *the Shannon entropy* of that series is defined by summing the surprisal of each event multiplied by the probability of occurrence of that event:

$$-\sum p(event_k)\log_2(p(event_k))$$

Unlikely events have a lower informational entropy than more likely events. To make this concrete, consider the simplest example of information, a coin flip, which defines two events, a head (H) and a tail (T). If the coin is fair, then the Shannon entropy of any series of coin flips using that coin can be computed using the definition above as:

$$\begin{aligned}&-[(p(\mathrm{H})\log_2 p(\mathrm{H})) + (p(\mathrm{T})\log_2 p(\mathrm{T}))]\\&= -[(0.5^* - 1) + (0.5^* - 1)]\\&= -[-0.5 + -0.5]\\&= 1\end{aligned}$$

If our coin were biased so it flipped heads 99 times out of 100, then the Shannon entropy of any series generated by flipping that coin can be computed as:

$$\begin{aligned}&-[(p(\mathrm{H})\log_2 p(\mathrm{H})) + (p(\mathrm{T})\log_2 p(\mathrm{T}))]\\&= -[(0.99^* - 0.0145) + (0.01^* - 6.64)]\\&= -[-0.0143 + -0.0664]\\&= 0.0807\end{aligned}$$

The total entropy is lower in the second case precisely because that case is much more predictable (less surprising), in the sense that you can make a very good guess (a guess that will be correct 99% of the time) that any particular flip will comes up heads. In the first case, you can't guess better than chance (50/50). Note in particular the much larger contribution made to the total entropy by a coin with a high 0.5 probability of heads (0.5) than the contribution made to total entropy by a coin with a low 0.01 probability of heads (0.0664). Less probable individual event components make a smaller contribution to an event's total entropy than more probable event components.

Our goal in this series of studies was to directly test Schopenhauer's hypothesis that greater incongruity between expectation and events produces a stronger feeling of humor, as operationalized by testing the hypothesis that funny NWs will have lower summed entropy values than non-funny NWs. It is not our intention to claim that our measure is the only explanation for humor even within the very limited domain of funny NWs. As we note in the general discussion at the end of this paper, there are various other forms of formally-definable incongruity that might be included that are not captured by the simple computation on which we focus here. Moreover, although we limit ourselves for pragmatic reasons to discussing

what are quite possibly the least funny jokes ever told, we also discuss how an information-theoretic approach to humor studies could generalize to more realistic, funnier jokes.

Study 1: Are NWs funny?

The first issue we wanted to address was to discover whether the funny NW effect is common and reliable enough to warrant focused investigation.

Method

To address this issue, we conducted a web survey that asked participants to rate a large set of NWs for their humor level on a seven-point Likert scale ranging from 1 (labeled 'Not at all funny') to 7 (labeled 'Definitely funny'). We asked undergraduate students at the University of Alberta to complete our web survey as part of a group of optional surveys that were completed over the course of approximately 1 h. Each student was randomly assigned to one of 60 groups, and each group was asked to rate a set of 100 NWs in a web-based survey. Due to restrictions imposed by the administrators of the web survey, the order of the presentation of the words within each group was identical. No restrictions were place on the timing or location of the experiment. Participants were requested to use a reliable computer and network connection and to find a quiet location where they would not be disturbed during their participation. Participants who did not complete the survey within 24 h of beginning were not included in the study.

Participants

Ethical approval for this experiment was obtained from the University of Alberta REB. All participants gave their informed consent before participating in the study. They received partial course credit in return for participating.

A total of 968 students participated in the survey. Although participants were permitted to submit the survey without entering ratings for all items on the survey, almost all of the participants provided ratings for all the items presented to them.

Stimuli

The NWs rated were generated using a program called Language-Independent Neighborhood Generator of the University of Alberta (LINGUA; Westbury, Hollis, & Shaoul, 2007). LINGUA generates NWs by Markov-chaining *n*-grams (strings of *n* letters, where *n* is a user-specified integer) from a frequency dictionary of a specified language. The process of using a Markov chain to create NWs from a dictionary ensures that every *n*-gram that appears in a NW appears in at least one real word, and that the statistical distribution of *n*-grams among the NWs mirrors the statistical distribution of *n*-grams among real words in the source dictionary. The software automatically precludes from the output any strings that appear in the source dictionary. We used a dictionary of 111,627 word types derived from a USENET corpus of 7,781,959,860 word tokens (Shaoul & Westbury, 2006). NW strings were generated using 3-grams (i.e. with the restriction that every three consecutive letters in a NW appear in a real word), a length that (as Shannon (1948), showed) produces highly word-like, usually pronounceable NWs such as 'inquenter' and 'artorts'. We generated 1200 strings of each length from 5 to 9 characters. We deleted pseudohomophones, and any recognizable English words that were not in our dictionary, ending up with 5928 randomly generated strings for rating. Dividing this set up into 60 groups of approximately 100 strings meant that each string was rated approximately 16 times.

Results

Our primary interest was in determining whether there was any consistency in the average funniness ratings, which ranged from 1.1 (for the strings 'exthe' and 'suppect') to 5.92 (for the string 'whong', whose sexual connotation is discussed further below). Note that this range is itself suggestive of consistency, since such extreme high and low values could only be obtained with consistent ratings for these stimuli at the extremes, at least.

In order to definitively test the hypothesis that the ratings were more consistent than could be expected by chance, we could have simply correlated split half judgments for all 5928 NWs. However, this test sets the bar very low, since extremely weak correlations will be significant with such a large set of stimuli (i.e. with $r = .03$, accounting for just 0.09% of the variance, $p = .01$ one-tailed). Since we are interested not in such weak population-level effects but rather in understanding whether the effects are strong enough *to be detected by individuals*, we conducted a more stringent analysis using

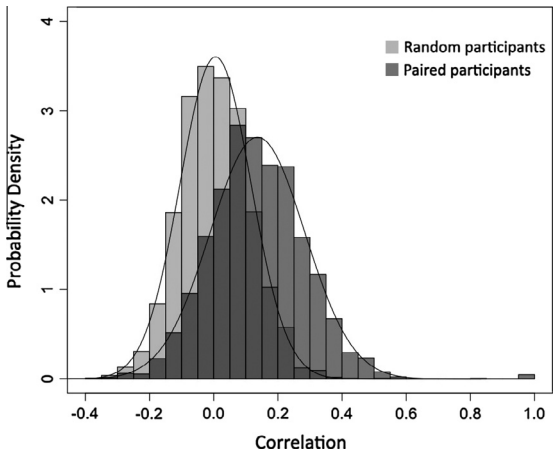

Fig. 1. Probability density distributions of 5000 random correlations between 100 NW pairs, drawn from participants who rated the same NWs ('Paired participants') or different NWs ('Random participants'). The average correlation for the paired judgments was 0.136 ($p = .04$ for 100 item pairs). The average correlation for the randomly paired judgments was 0.006 ($p = .24$ for 100 item pairs). The dark central region shows the region of overlap between the two distributions.

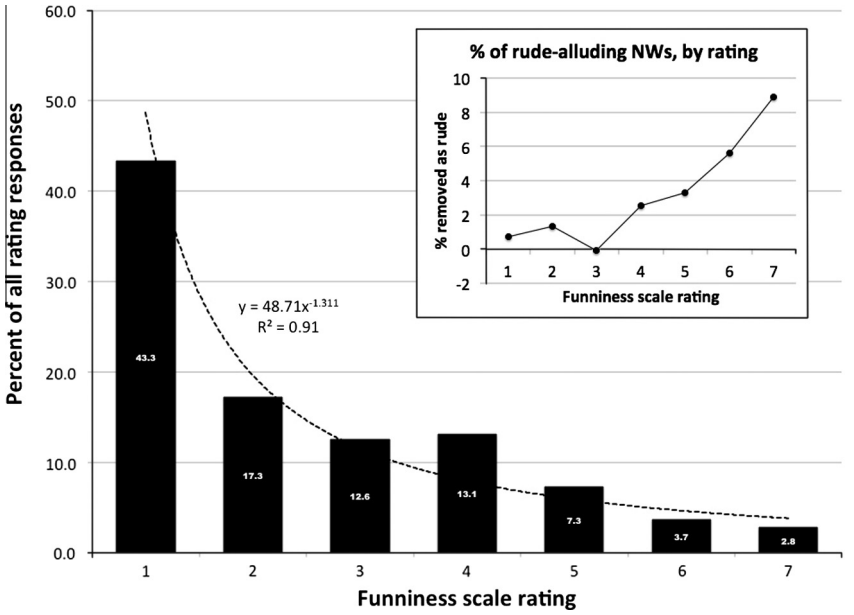

Fig. 2. Total number of ratings obtained for each value of the rating scale, from '1' ('Not at all funny') to '7' ('Definitely funny'). 43.3% of the rated NWs received the lowest rating. The inset graph shows the percent of words removed because they contained a string that alluded to a rude word of English, as a function of rating (see text for details). More words were removed from funnier ratings, reflecting the fact that rude-alluding NWs tend to be judged funny.

a Monte Carlo analysis. We randomly sampled 5000 unique subject pairs who had seen the same 100 stimuli and compared the correlations between their judgments to the correlations of a random sample of 5000 unique subject pairs who had seen a different 100 stimuli. The average rating correlation for the 5000 within-group subject pairs of 100 NW ratings was 0.136, which is a reliable correlation for 100 items ($p = .04$). The average rating correlation for the 5000 between-group subject pairs of 100 NW ratings was a non-reliable $r = .006$ ($p = .24$). As shown in Fig. 1, these two samples were reliably different ($t(9275.39) = 50.08$, $p < .0001$; Cohen's $d = 0.89$).

The correlations in both these samples were reliably higher than zero by one-sample t-test (Within-group: ($t(4999) = 65.3$, $p < .0001$); Between-group: $t(4999) = 3.61$, $p = .0002$). The fact that random between-group correlations (in the between-group case) exceed zero probably reflects the distribution of the judgments. As shown in Fig. 2, there were many more ratings of '1' ('Not at all funny'; 43.3% of all ratings) than there were of any other value, and the number of ratings decreased as the rating increased, in a power law relationship ($r = .91$). Because of this, strings that are randomly drawn are likely to contain low numbers, reflecting the unsurprising fact that most randomly generated NWs are not perceived to be very funny. This sample bias would allow the average random correlation to be reliably higher than zero.

Although agreement on non-funniness is *relevant* agreement, and despite the fact that the average random correlation obtained was not reliably different from zero for 100 pairs of random NW ratings, we wanted to be sure that the correlations we obtained did not simply reflect this 'default' judgment that most NWs are not funny. We therefore repeated the Monte Carlo analysis with the constraint that it only use judgment vectors that contained at least 25 ratings greater than '1'. The average rating correlation for the 5000 within-group subject pairs was 0.142, which was reliably different than zero ($t(4999) = 68.3$, $p < .0001$) as well as being a reliable correlation for 100 item pairs ($p = .04$). The average rating correlation for the 5000 between-group NW pairs was 0.002 ($p = .49$ for 100 items), which was not reliably different than zero ($t(4999) = 1.11$, $p = .13$). The two samples were again markedly different ($t(9389.054) = 53.40$, $p < .0001$; Cohen's $d = 0.94$). The correlation between judgments of matched words reflects more than just agreement that many NWs are 'Not at all funny'.

These analyses provide evidence that random NWs do have a reliably consistent humor rating. Subject judgments for sets of the same 100 NWs are reliably correlated, and much higher than average correlations of subject judgments for a different 100 NWs, which are close to or indistinguishable from zero.

Before we could begin our attempt to model what made the strings funny, we needed to address a semantic complication that makes some funny NWs funnier than any purely formal incongruity measure could suggest: the problem of dirty words. Recognizing a rude word in a non-rude setting is notoriously funny. It was immediately obvious that some of the funniest NWs were funny because they were alluding to rude concepts, because the funniest-rated six NWs among the 5928 that were rated included the strings (which we present here without further comment) 'whong', 'dongl', 'shart', 'focky' and 'clunt'.[4]

[4] The sixth string among the top six, rated third-funniest, was 'blablesoc'. We note that several of the listed words are recognized by some as slang words with rude meanings, though they did not appear in the dictionary we used to filter words, and thus did not appear in 7.8 million word token UseNet corpus that we used to build that dictionary.

At the risk of using statistics to prove the obvious, and because we wished to eliminate all these 'semantically parasitical' NWs from consideration, we quantified the relationship between rude allusions and humor judgment in NWs. To identify rudeness-alluding NWs, we wrote down every string we could think of that sounded like, looked like, or actually was a rude English word. We came up with 52 such strings, which we have re-printed in Appendix A. We searched the list of rated NWs for occurrences of the rude strings. 287 (4.92%) of the NWs contained at least one of the 52 substrings. To demonstrate empirically that their existence was reliably related to their rated humor level, we ordered our NWs strings by rating, from funniest to least funny. We then counted the rude-alluding strings in each half, and conducted a chi-square test (with Yates' correction) to test if their distribution among the two halves varied beyond chance expectation. Of the 287 rude strings, 207 (72%) occurred in the first (funniest) half, giving a chi-squared value of 58.1 ($p < .0001$). NWs alluding to rude words are indeed more likely than chance would predict to be rated among the funniest NWs (see also the inset graph in Fig. 2, which shows the percentage of rude-alluding words as a function of rated humorousness). Note that this finding addresses the potential criticisms that subjects either (i) misunderstood what we were asking and rated words as 'funny peculiar' instead of 'funny ha-ha' or (ii) understood what we were asking but, upon failing to find *any* NWs humorous, rated words as 'funny peculiar'. If subjects were choosing *peculiar* rather than *humorous* strings, it is vanishingly improbable that we would see rude-alluding strings clustering consistently at the top of their ratings, especially since these strings are, in virtue of being noticeably close to real words, *less* peculiar than the majority of strings which did not obviously allude to English words.

It is possible that the measured reliability of funniness judgments we reported above is due entirely to subject agreement about these rude-alluding NWs, i.e. that NWs are only humorous if they allude to something rude. In order to test if this was the case, we repeated the Monte Carlo analysis described above, but first removed all the rude-alluding NWs. The average rating correlation for the 5000 within-group subject pairs was 0.125 when all pairs were included, which was a just marginally unreliable correlation for the average of 95 pairs that remained after removing the rude-alluding NWs, with $p = .056$. The average rating correlation for the 5000 between-group subject pairs was 0.006, which was not a reliable correlation between 95 pairs ($p = .48$). The two distributions were again clearly drawn from populations that were very different ($t(9435.9) = 46.6$, $p < .0001$; Cohen's $d = 0.87$). We have not included a graph of the density distribution, as it is very similar to the distribution shown in Fig. 1.

The *average* correlation of the 100 matched NW pairs in this sample was marginally lower than the value reliable at $p = .05$, suggesting (as indeed did the just-marginally-reliable correlation reported above, for strings that included the rude-alluding NWs) that the funny NW effect may be a delicate effect when examined in random strings. However, the reliability of the average correlation is not the best measure of the strength of the effect in this case. Since there were slightly more reliable correlations at $p < .05$ (2510/5000 or 50.2%) than unreliable correlations in this sample, another way of stating the finding that the average correlation is close to the reliable correlation is that randomly-selected judgment vector pairs agree fully 50.2% of the time at a level of probability that is predicted by chance to occur only 2.5% of the time ($p < .05$, one-sided). The chance probability of seeing a distribution so strongly skewed toward the improbable (at least 2510 reliable correlations when chance predicts 125) is less than 0.0001.

We know that this mis-estimates the true chance probability in our observed distribution, since the over-representation of low judgments in our dataset makes chance correlations more likely and the observations are not independent (there may be effects due to drawing from the same subjects or from re-sampling the same vectors). We can obtain a more realistic estimate of the chance probability of a reliable correlation between random sets of 100 NW pairs in this particular dataset by using the observed Monte Carlo probability of a reliable correlation in the unmatched strings as our estimate of the probability of obtaining a reliable correlation by chance. Six hundred and twenty-four (12.5%) of the strings in the unmatched set correlated at a level higher than predicted by chance, $p < .05$ one-sided. If we set the probability of obtaining a reliable positive correlation by chance to this empirically-observed value of 12.5%, the chance probability of obtaining the 2510/5000 reliable correlations is still less than 0.0001.

Based on these considerations, we concluded that the judgment data justified an attempt to isolate the source of these extremely improbable effects. We removed all the rude-alluding NWs from further consideration, and conducted all further analyses with the remaining 5641 strings.

Experimental tasks: forced-choice & humor judgment

To test the Schopenhauer-inspired expectation-violation hypothesis, operationalized as a test of whether manipulating Shannon entropy can reliably predict NW humor, we conducted two experiments. One was a forced choice experiment, which required subjects to select which one of two strings was more humorous. The second was a rating task, in which subjects had to rate strings for humor on an unmarked Likert scale. As the same participants were used in both tasks, using closely-matched counter-balanced stimuli sets, we begin with a description of the stimuli, methods, and participants that were common to both experiments.

Stimuli

Shannon entropy in letter strings can be defined over different linguistic elements. For example, we could use single letters, letter pairs (bigrams) or letter triplets (trigrams) as a unit for the computation. We computed the entropy of the n-grams in our NWs, with $n = 1–3$, dividing by the number of n-grams to adjust for string

length. Across all 5641 non-rude-alluding strings (which, recall, includes 43% non-funny NWs), the correlation of length-adjusted entropy with the human judgments was −0.24 for the single letters, −0.18 for the bigrams, and −0.10 for the trigrams ($p < .00001$ in all cases). After regressing out the effect of the one-grams, the bigrams accounted for just 0.002% of additional variance in the humor ratings and the trigrams did not account for any additional variance. Given the very small additional contribution of the bigrams, we used only the simplest length-adjusted unigram entropy as our predictor of NW funniness. This simple measure is essentially a measure of the summed probabilities of the individual letters in each string.

To select our experimental stimuli, we generated 6000 new NWs, none of which appeared in the original list of rated strings. In order to avoid any complications of length effects, we fixed the length at 7 characters. We ordered these 6000 NWs in increasing order of unigram entropy. We then selected approximately every 60th string, plus the closest following string, subject to the following three conditions:

- The selected strings contained no rude-alluding substrings, according to the algorithm described above.
- The strings were judged to have an unambiguous pronunciation.
- The strings were judged not to violate English spelling rules.

If the 60th string violated any of these conditions, the next closest acceptable string was chosen as the first string. In this way we selected 200 NWs without using human judgment, in two sets of 100 that varied systematically across the full range of unigram entropy and that were extremely closely matched on that measure (Average [*SD*] unigram entropy in both sets: 1.68 [0.19]; Correlation between the two sets: 0.999). The 100 matched stimuli pairs and their entropy values are included in Appendix B.

Participants

Fifty-six fluent English speakers (23 males, 33 females) participated in these experiments in return for partial course credit. All participants self-reported that they had learned English as a first language. They had an average [*SD*] age of 19.9 [3.3] years and an average [*SD*] of 13.6 [1.9] years of formal education. All subjects gave written consent to participate in the experiment, which was conducted in accordance with the regulations of the University of Alberta Research Ethics Board.

Methods

Data were collected in a laboratory setting using ACTUATE software (Westbury, 2007) running under OS 10.6 on G4 Mac Minis connected to 17″ LCD monitors, in one of three rooms constructed to reduce outside noise.

All participants completed two experiments, using the two different sets of NWs. Experiments were run in a counter-balanced order, and the two forms were counter-balanced within each experiment. It was explained to participants that from previous research we knew that people are able to make reliable decisions about whether NWs evoke humor and that we were currently testing a model of that effect. In explaining this, we used a written text which was read aloud to each subject by a research assistant. The instructions for both experiments began with the following paragraph (with Canadian spelling intact, and the single difference between the experiments in square brackets):

> "In this experiment we ask you to undertake an offbeat task, which is to [compare/rate] how humorous nonwords are. We have previously shown that people have consistent intuitions about which nonwords are humorous and which are not. This has implications for understanding how people process nonwords and how they process humour. Based on previous ratings of many NWs, we developed a successful statistical model that predicts how humorous people will find a nonword string. In this experiment we are seeing if we can improve that model: we want to see if people's ratings accord on average with what we predict."

We deliberately used the word 'humorous' rather than the more natural 'funny' throughout, to ensure that there was no possible misunderstanding by participants about whether we meant 'funny ha-ha' or 'funny peculiar'. After describing the requirements of the specific task, both sets of instructions also included this sentence: "This is not a test; we are just interested in your gut feeling".

The data in each experiment were analyzed with R (R Core Team, 2013), using generalized linear mixed-effect regression models (binomial models for the percent data), fitted by Laplace approximation (Baayen, 2008). Stimulus order and participant were entered into each model as random effects.

Experiment 1: forced choice

Methods

In the forced choice experiment, subjects were shown two strings on the screen, side by side, and asked to decide as quickly as possible which one was "more humorous". Strings were counter-balanced so that each was seen equally often on the right and the left side of the screen. The strings were presented in random order in lower case 60-point black Times font against a white background. Participants were asked to press the 'x' key if they thought the string on the left was more humorous, and the 'c' key if they thought the word on the right was more humorous. In case they thought neither was more humorous, they were told to choose randomly.

The 50 string pairs in each form were arranged so that the string with the highest entropy was paired with the string with the lowest entropy, the string with the second lowest entropy was paired with the string with the second highest entropy, and so on. In this way the pairs varied systematically across the range of possible entropy difference, with some string pairs being as different as possible on their entropy and some being nearly identical. This allowed us to test the hypothesis that human

responses would be more accurately predicted for NW pairs with large entropy differences (i.e. when the differences in predicted funniness was also large) than for pairs with small entropy differences (when the difference in predicted funniness was small). Since easier decisions are usually made more quickly, we also predicted that we would find faster RTs for NW pairs with large entropy differences than for NW pairs with small entropy differences.

Results

Prior to analysis, we deleted all stimuli with RTs < 400 ms (112 stimuli; 3.9% of all stimuli), on the grounds that such quick judgments about two NWs reflected response errors or inattentive judgments. We then computed the average and *SD* RT for the remaining data and removed all stimuli with RTs > 3 SDs from the mean (>7908 ms, 50 stimuli, 1.7% of all stimuli). As the stimuli removed by these means included 23 stimuli from a single male participant (i.e. nearly half of his data), we removed the rest of that participant's data from all further consideration, leaving 55 participants.

To predict string choice, we considered models that used the entropy distance between the two strings and the side on which the target (lowest entropy string) was presented, along with demographic variables: gender, age, and education. No demographic predictors contributed reliably to the model, but the entropy distance did. Adding random slopes did not improve the model (see Table 1). The best model is presented graphically in Fig. 3. As hypothesized, larger differences in entropy more strongly predict human decisions than smaller differences in entropy.

According to exact binomial probability, 22 subjects (40% of all subjects) responded reliably ($p < .05$) in the direction predicted, with the most accurate subject choosing 76% of the time as predicted (an occurrence with an exact binomial $p < .0001$). Considering only the stimuli in the top 50% of entropy difference (i.e. those predicted to differ most clearly), 31 participants (56.4%) responded reliably in the direction predicted, with the most accurate subject choosing 92% of the time in the predicted direction (exact binomial $p < .00001$). In contrast, only a single subject (1.8%, no different from what could be expected by chance) responded reliably in the predicted direction for the stimuli pairs that had an entropy difference in the bottom 50%.

The best model for predicting RT (for prediction-consistent items only) included just the entropy distance between the two strings, with none of the demographic variables contributing reliably (Table 2) and without random slopes. As shown in Fig. 4, participants made faster decisions (reflecting an easier choice) when the entropy difference was large than when it was small. However, although the effect of entropy difference on LNRT was statistically reliable ($t = 3.11$, $p = .002$ as estimated using the Kenward–Rogers approximation for degrees-of-freedom implemented in R-package *afex*, Singmann & Bolker, 2014), the effect was very small. To compare the models, we used the Akaike Information Criterion (AIC, Akaike, 1974), a measure that allows us to estimate the relative information lost by each model. The AIC difference between the base and fitted model was just 3, suggesting that the better model was only about 4.5 times more likely to minimize information loss than the base model.

Table 1
Regression model analysis for predicting forced-choice decisions about NWs humorousness. 'Improvement?' is the improvement of the model's ability to minimize information loss compared to the base model, computed by comparison of AIC values. The best model is shown in bold text.

Model	AIC	Improvement?
M1: Random effects only	3579	[BASE]
M2: M1 + age	3581	No
M3: M1 + education	3581	No
M4: M1 + gender	3579	No
M5: M1 + entropy difference	**3511**	**>1,000,000*x***
M6: M5 + subject random slopes	3512	No
M7: M5 + order random slopes	3511	No

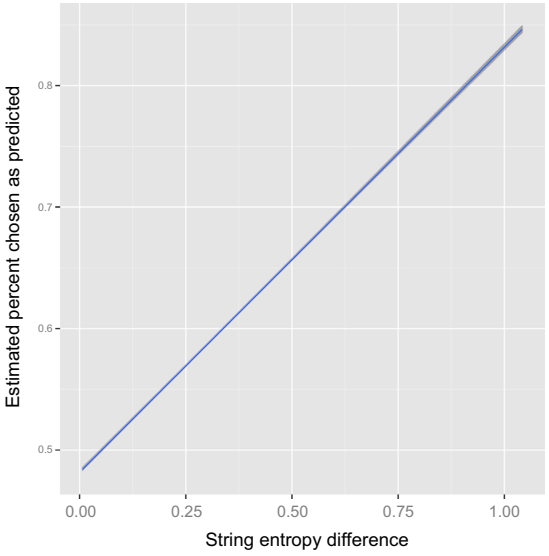

Fig. 3. Estimated probability (using the model in Table 1) of choosing the string with lower entropy as more humorous in the forced choice humor judgment task, as a function of entropic distance between 100 string pairs. The 95% confidence interval is graphed but hardly visible.

Table 2
Regression model analysis for RTs of correctly-predicted forced-choice decisions about NWs humorousness. 'Improvement?' is the improvement of the model's ability to minimize information loss compared to the base model, computed by comparison of AIC values. The best model is shown in bold text.

Model	AIC	Improvement?
M1: random effects only	2707	[BASE]
M2: M1 + age	2714	No
M3: M1 + education	2712	No
M4: M1 + gender	2710	No
M5: M1 + entropy difference	**2704**	**4.5*x***
M6: M5 + subject random slopes	2706	No
M7: M5 + order random slopes	2708	No

Discussion

The results from this forced-choice humorousness decision task lend straightforward and consistent support to

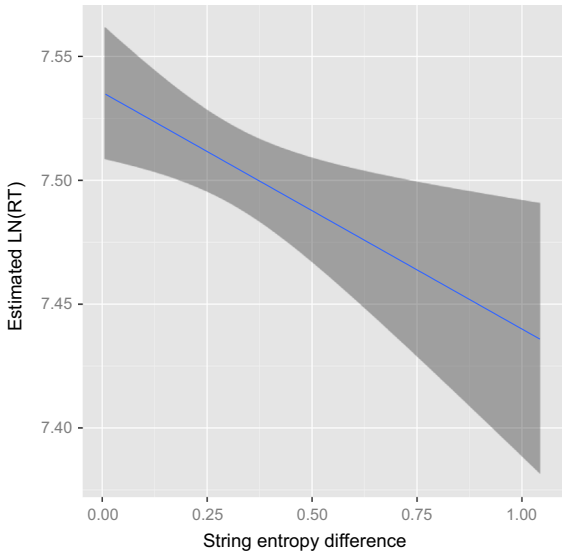

Fig. 4. Estimated RT (using the model in Table 2) for choosing the string with lower entropy as more humorous in the forced choice humor judgment task, as a function of entropic distance between 100 string pairs, with 95% confidence intervals. Participants make the predicted decision marginally more quickly when the entropic distance between two strings is larger.

the hypothesis that the subjective funniness of NWs is a continuous function of the likelihood of those strings, as measured by unigram entropy. Strings with low unigram entropy are reliably chosen as more funny than paired strings with higher unigram entropy, and the likelihood of this choice increases with the entropy distance of the two strings. Weaker evidence suggests that participants make the predicted decision more quickly for strings that have larger entropy differences than for strings that have smaller entropy differences.

Our stimuli were paired in such a way that the pairs with the least difference was the pair that fell in the middle of the range of distance. The decision to structure the experiment in this way meant that we did not compare words that were both high or both low on entropy. The second experiment sidesteps this methodological limitation by presenting words one at a time.

Experiment 2: rating task

Our second experiment simply showed subjects the 100 strings they had not seen in the first experiment (which alternated between subjects), and asked them to rate them using an unmarked Likert scale.

Methods

Strings were presented in black 40 point Times font at the top of an unmarked Likert scale labeled 'Least humorous' on the left and 'Most humorous' on the right. Participants were asked to use the computer mouse to drag a marker to the location on the Likert scale that corresponded to their judgment of how humorous the string was. The scale score was recorded as an integer from 1 (Least humorous) to 100 (Most humorous).

Results

None of the demographic variables were reliably predictive of the subjective judgments of funniness. The best model for predicting subjective judgments of funniness included just the unigram entropy of the string, without random slopes (Table 3). As shown in Fig. 5 and as predicted, strings with lower entropy were judged to be funnier than strings with higher entropy, with high reliability ($t = 15.6$, $p < .00001$ as estimated using the Kenward–Rogers approximation for degrees-of-freedom). Across all 200 strings, the correlation between the model estimates and the entropy was -0.83 ($p < .00001$).

Discussion

The results of Experiment 2 add strength to the two earlier main conclusions: that there are consistencies between participants in judgments of NW funniness, and that the magnitude of those judgments is reliably predictable across the entire range of the unigram entropy of the NW strings.

General discussion

Our investigation into the very slight humor of NWs has allowed for a useful methodological innovation in humor studies, namely, the formal quantification of the 'ludicrous effect' in a way that is not affected by the complexities of real-world probabilities or of lexical semantics. To our knowledge, this paper presents the first quantitative test of Schopenhauer's nearly 200-year old explicit claim that the expectation-violating categorical anomaly underlying humor provokes a sense of funniness precisely to the extent that it departs from the expected category, as well as the first attempt to explicitly formulate and test that hypothesis in information-theoretic terms. We found empirical support for Schopenhauer's claim in three different studies, first showing that it fits as a post-hoc description of funniness ratings, then showing two experiment results: first, that the pair-wise 'categorical anomaly' (non-wordness) distance between two NWs predicts the probability that one will be chosen funnier, and second, that the same measure of non-wordness correlates with funniness ratings for strings presented individually.

Although we formulated the idea independently, the idea that NWs are humorous because they contradict our expectations that what we will read is a meaningful word was first explicitly stated by the German philosopher Theodor Lipps in his (1898) *Komik und Humor: Eine Psychologisch-Ästhetische Untersuchung* [*Comedy and Humor: A Psychological and Aesthetic Investigation*].[5] His analysis of the humor of NWs is quite consistent with Schopenhauer's conceptual expectation violation framework for analyzing humor (although elsewhere in his book Lipps dismisses that

[5] As this work has not been translated into English, we rely here on a commissioned translation of the relevant passages from the Project Gutenburg German text.

Table 3
Regression model analysis for Likert scale ratings of the humorousness of 200 NW strings. 'Improvement?' is the improvement of the model's ability to minimize information loss compared to the base model, computed by comparison of AIC values. The best model is shown in bold text.

Model	AIC	Improvement?
M1: random effects only	51,648	[BASE]
M2: M1 + age	51,649	No
M3: M1 + education	51,647	No
M4: M1 + gender	51,645	No
M5: M1 + entropy	**51,409**	**>1,000,000*x***
M6: M5 + subject random slopes	51,412	No
M7: M5 + order random slopes	51,413	No

framework as tautological, using a rather weak argument that need not concern us here). Lipps argues that the humor of words we do not understand rests on the fact that we expect them to denote something, but they violate this expectation because they do not denote anything:

> "One knows the trend of young people to change or reverse words such that they stop being meaningful linguistic signs, but because of the similarity to the original can still be understood. The humor of these 'funny word reversals', as well as the funniness of words in general, *is based on the contrast of meaninglessness and understandable meaning*. [...] The possibility of "funny pseudo-terms" rests on the habit of associating words with a meaning. For example, if someone asks me what this or that is, I answer the question with a word that does not exist and that has no meaning for anybody; simply based on the belief of the listener that it must be possible, *given that he only hears words*, to think something. The humor arises for those who let themselves be astonished and who are tricked for one moment, but then realize that they were duped. [...] *The precondition is that prior language experience makes these forms appear meaningful despite their meaninglessness*." [no page numbers; Emphasis by italics was added by us].

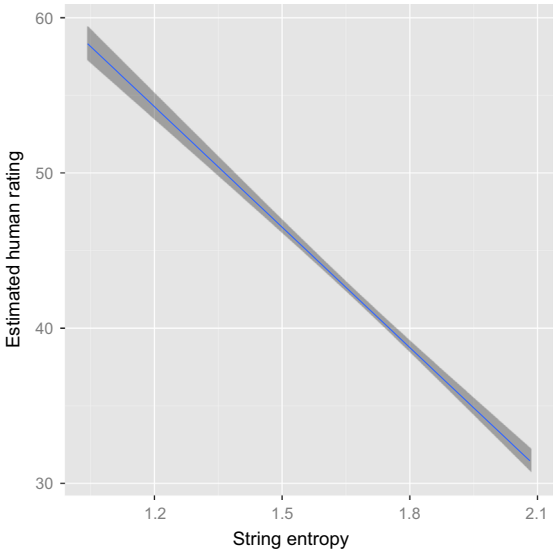

Fig. 5. Estimated humor rating (using the model in Table 3) as a function of unigram entropy for 200 NWs, with 95% confidence intervals. Participants rate strings as being more humorous when the unigram entropy of the string is lower.

This conception is closely consistent with what we have tested here.

We did not deliberately design our NWs to be funny. However, others have done so, at least intuitively. The children's writer Dr. Seuss (Theodor Geisel) is famous for his use of funny nonsense words such as 'wumbus', 'skritz' and 'yuzz-a-ma-tuzz'. As a validation check on our results, we computed the length-adjusted entropy of 65 nonce words from Dr. Seuss books, and compared them to the length-adjusted entropy of a full dictionary (111,625 words) of real English words, to test the prediction that Dr. Seuss's funny strings would have low entropy. Dr. Seuss's nonce words are reliably lower in entropy than ordinary English words ($t(64.03) = 7.52$, $p < .00001$), as shown in Fig. 6.

Although our results do provide strong support for the hypothesis that greater one-gram entropy distance is associated with more strongly experienced humor, we do not believe this is the only relevant measure. There are other measures of how unusual a string is, including notably how many doubled letters it includes (Dr. Seuss's strings often include double letters, such as 'zz', 'oo' and 'kk'), how similar the NWs are to real English words, and how unusual the string's phonology is. There may exist other forms of 'semantic parasitism' than the single class we considered, of strings that resemble rude words. Some of our NWs undoubtedly resembled real words to a greater degree than others, and it is likely that some NW allusions to real words may be more humorous than others. Quantifiable incongruity based on letter entropy is almost certainly just one element of a larger set of cues to humor that subjects experience and synthesize.

A related question to this question of whether our entropy measure is the 'proper' one is the question: Why did we cast out explanation of humor in NWs in terms of *information theory* instead of *raw frequency*, given that our 1-gram information measure is closely related to and strongly correlated with simple letter frequency?

One answer to this question is simply that the theory we were testing—Schopenhauer's 1818 expectation violation theory—is inherently information-theoretic in its very formulation. Although of course Schopenhauer had never heard of Information Theory (which was not known until Shannon published his paper a hundred and thirty years later), the main point that Schopenhauer was trying to make was that the humor of a thing depends upon its conceptual context ("the sudden perception of the incongruity between a concept and the real objects which have been thought through it in some relation"). Schopenhauer's contribution was to contradict Bain's (1856) theory by specifically emphasizing that it is *not* the raw frequency of a thing that makes that thing funny. It is rather *the context of conceptual expectation is which that thing is embedded*, a relational measure between two or more things rather than a simple frequency count.

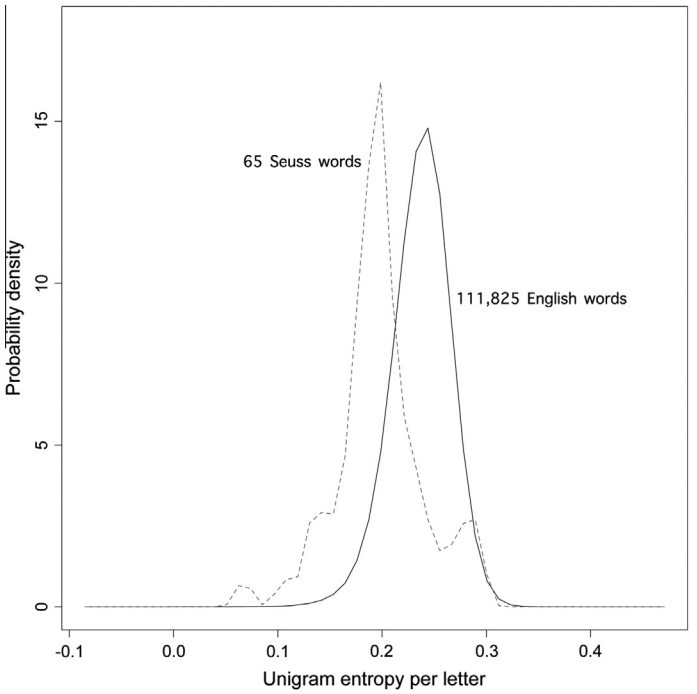

Fig. 6. Comparison of the length-normalized unigram entropy of 65 NWs invented by Dr. Seuss, to 111,825 English words.

A second answer to the question is that we consider frequency to be a problematic construct. Although frequency is often used as an explanatory construct in accounting for lexical processing, many authors (e.g. Adelman, Brown, & Quesada, 2006; Baayen, 2010; McDonald & Shillcock, 2001) have pointed out that frequency measures beg numerous questions, most notably: what exactly should one count when measuring frequency? This may seem obvious, but it is not. A clear example of why not is word frequency, in which the obvious thing to count is, of course, occurrences of the word in which we are interested. However, in addition to these word frequencies, the variety of relational contexts in which a word is situated are also a critical component of the so-called word frequency effect in lexical access tasks. This point was made forcefully by Baayen (2010), who argues against the existence of *any* word frequency effects in lexical access by showing that all the variance attributable to word frequency can be accounted for by the relational measures. Similar arguments apply equally to *n*-gram frequencies, as Baayen and his colleagues made clear in implementing their Naïve Discriminant Learning model of word acquisition, which ignores word boundaries entirely by acquiring language from the co-occurrence probabilities of letter trigrams that may happen to include spaces (Baayen, Milin, Filipovic Durdevic, Hendrix, & Marelli, 2011).

If we ignore contextual relationships and assume that processing is influenced only by raw frequency counts, we are forced to make the implausible assumption that the mind faithfully tracks the frequency of occurrence of lexical items at every level of description, prior to integrating and transforming these raw counts at the time of processing. This seems implausible, since there is scant evidence for mental counters at any level of linguistic description and of course undertaking such complex probability calculations in real time would be a Herculean task.

There is a plausible alternative approach to assuming that human beings are super-counting probability computers who enumerate everything they can before synthesizing the co-occurrence probabilities on demand in scant microseconds. The alternative is to assume that relevant lexical items and their relevant contexts are directly learned by standard discriminative learning principles (e.g. Rescorla & Wagner, 1972) *as they are encountered*. Under this assumption, the many frequency measures at various levels of description that have been shown to correlate with lexical processing in behavioral tasks do so not because anything has actually been enumerated, but rather because the frequencies are *implicitly* captured in the system of weights that develop in psychologically plausible learning models (see Baayen, 2010, 2012; Baayen, Hendrix, & Ramscar, 2013; Baayen & Ramscar, 2015; Baayen et al., 2011; Mulder, Dijkstra, Schreuder, & Baayen, 2014; Ramscar, Hendrix, Shaoul, Milin, & Baayen, 2014 for further discussion of this approach, along with numerous empirical demonstrations of its efficacy, and see Rescorla 1988, for evidence that discriminative learning learns about *the informational structure of the environment*, rather than simply the frequency of occurrence of stimulus–stimulus pairings).

Not only does the discriminative learning approach obviate the need to assume that our minds count the frequency of events at every level of description, but these same principles are foundational to much contemporary research in other domains involving learning and decision-making (Ramscar, Dye, & McCauley, 2013),

allowing results in lexical tasks to be related by a common explanatory framework to these other linguistic and cognitive domains. For example, Latent Semantic Analysis (LSA), perhaps the most well-known distributional model of lexical semantics, subjects the raw co-occurrence data between words and documents to a series of transformations in which local term frequency is first log-transformed, and then divided by the term's entropy over documents. Landauer and Dumais (1997) are explicit in noting a connection to learning and information, remarking that this inverse entropy measure "accomplishes much the same thing as conditioning rules such as Rescorla and Wagner (1972) in that it makes the primary association better represent the informative relation between the entities rather than the mere fact that they occurred together" (p. 216).

Likewise, tf–idf, a common metric used in term weighting in vector space models, can also be understood in terms of information. As Aizawa (2003) concludes, tf–idf, is simply a measure of the amount of information in a term weighted by its occurrence probability (see Blei, Ng, and Jordan (2003) for discussion of the relation between tf–idf and latent semantic indexing). Jones, Johns, and Recchia (2012) incorporate a principled variant on tf–idf in their computational model of lexical access, which "adjusts its encoding strength for a word relative to the information redundancy between the current memorial representation of the word and the current linguistic context in which the word is experienced" (p. 116).

That so many different models capitalize on these broadly similar measures, suggests that information theoretical measures does indeed capture something important about how humans cognize their environment.[6] Indeed, as Gallistel (2003) notes, the principles that are enshrined in discrimination learning models (and other neural network architectures) are, in fact, always information theoretic. Gallistel (2003) elucidates a number of parallels between the basic accepted principles of learning and information theory. On Shannon's view, as outlined in the introduction, the information conveyed by a signal is quantified in terms of the amount of uncertainty it reduces. Similarly, learning seems finely attuned to the informational relation between many different environmental contingencies. Gallistel proposes that conditioning is a function of information content, and that like any other signal, a cue is informative to the extent that it reduces uncertainty about the upcoming outcome (see also Balsam & Gallistel, 2009).[7]

We note, finally, that adopting an information-theoretic framework makes our results more directly applicable to other forms of humor. If we cast of our results in terms of raw letter frequency, we would leave open the questions: What thing(s) should I count in other humorous situations? What entity plays the role that letters plays in NWs in a pun or a riddle? Since information theory focuses on relevant co-occurrence frequencies rather than absolute counts, it provides a common and mathematically well-defined metric for thinking about humor. Following Schopenhauer, our claim is not that anything that occurs less frequently is funnier than anything that occurs more frequently, but rather than humor is a function of the unexpected *occurring in a context of prior expectation*.

That said, it is not obvious how to extend our results in a straightforward way to all real jokes, which may have incongruities that depend on real world co-occurrence probabilities that might be hard to estimate. This problem is exactly what made NWs attractive as experimental stimuli for studying humor. However, an extension beyond NWs is possible, since we can easily compute the word-entropy of phrases. For example, the Schopenhauerian humor-as-entropy theory predicts, *ceteris paribus*, that people should find the high entropy phrase 'existential llama', which Google counts as occurring less than one hundred times (almost all on the same website), more humorous than the lower entropy phrase 'angry llama' which has over 13,000 occurrences on Google. This is an easily testable claim, although we feel about it as Köhler (1929) did about the sound symbolic effect of his nonwords *maluma* and *takete*: "if the reader is asked to choose. . .he will probably be able to decide with ease" (p. 242).

Our findings do not necessarily vindicate the expectation violation theory as the only explanation for humor, even in the very narrow context to which we have confined ourselves. One limitation of the work we have presented here is that we did not make an attempt to understand the subjective motives behind the judgments made by the participants in our studies. We suspect that people may be sensitive to multiple reasons for finding a string humorous, and that often they would not have sufficient conscious insight to be able to explain why (consider how often we hear 'I dunno, it's just *funny*' when we ask people to explain an experience of humor that we don't share).

The question we would really like to answer is: *Why* do people find NWs humorous? Our work addresses this question at an abstract level. Insofar as it shows that NW strings act exactly as the expectation violation theory of humor suggests they should, this work suggests that one answer to this question is: NW strings are funny because they violate our expectations of what a word is.

However, such an explanation does not answer the deeper and more general question: *Why* are violations of expectation experienced as *funny*? As mentioned in the introduction, Hurley et al. (2011) offered one possible answer to this question, by suggesting that the rewarding experience of humor is an evolutionary adaptation, later co-opted for multiple purposes. According to their theory, humor originally evolved to increase the probability of heuristic cognitive search (with its attendant downside

[6] Recchia and Jones (2009) found similarly strong performance for positive PMI. This represents a key advance for computational models of lexical semantics, as there are marked drawbacks to employing SVD for dimensionality reduction. For one, calculating SVD is computationally expensive, making it difficult to scale to realistically sized corpora. For another, its calculation requires that semantic representations be computed in batch over the entire co-occurrence matrix, rather than updated incrementally in response to new input. Given that human semantic memory is dynamic, rather than static, models that rely on SVD are limited in their claims to cognitive plausibility. By contrast, PMI is a simple, scalable algorithm that learns incrementally.

[7] The upper bound on uncertainty is given by a Poisson random rate process, in which the occurrence of one event provides no information about subsequent occurrences. Therefore, it is deviations from Poisson that are informative, and that establish the 'signal value' of the cue.

increase in the risk of making a mistake), both by 'bribing the brain with pleasure' (making us feel good when we go off the beaten cognitive path) and communicating that pleasure to our conspecifics (i.e. via the public behavior of laughter; see Jung (2003) for another theory that focuses on the evolutionary value of publicly communicating humor). Humor thereby makes us more intelligent, by encouraging cognitive exploration, 'cognitive debugging', and the resultant discovery of useful new ideas, as well as the public signaling of how pleasurable these exploratory activities can be. While acknowledging Schopenhauer as an inspiration, Hurley et al. (2011) differentiate their theory from Schopenhauer's expectation violation theory by arguing that Schopenhauer's theory relied too heavily on a distinction between perception and cognition (p. 113), which Hurley et al. (2011) reject. We consider this to be more of a 'friendly amendment' to Schopenhauer's theory than a contradiction of it.

Though it suffers, like many psychological evolutionary theories, from being post-hoc and hard to prove scientifically, Hurley, Dennett, and Adams' theory has the dual attractions of being plausible and offering a possible explanation as to why humor has proven so hard to pin down scientifically. The co-option of humor for multiple disparate purposes, including communication and our own entertainment, suggests that humor detection may not be a basic functional unit, but rather a complex, multi-determined meta-function that connects other psychological functions in complex ways (for a similar conclusion, see Gervais & Wilson, 2005; Polimeni & Reiss, 2006). Insofar as this is true, a processing model of humor *tout simple* will probably prove elusive, for much the same reason that a search for a processing model of 'understanding' is bound to fail (see Wittgenstein, 1953, on the variety of very different ways one can 'understand' a thing).

Functional imaging studies of humor detection support this idea that humor detection is probably not a single monolithic psychological function. Goel and Dolan (2001) presented participants in the scanner with auditory jokes that were either phonological (puns, e.g. *Why did the golfer wear two sets of pants? Because he got a hole in one*) or semantic (e.g. *What do engineers use for birth control? Their personalities*). After subtracting out activation attributable to non-humorous semantic processing or phonological processing alone, they found dissociable patterns of activation for each type of joke that were related to the patterns seen for processing each type of information (bilateral temporal lobe activity for semantic jokes; and left posterior inferior temporal gyrus and inferior frontal gyrus activity for phonological jokes). Overall, joke funniness ratings were associated with increased medial ventral prefrontal cortex activity, reflecting its rewarding affective component. Other studies (Mobbs, Greicius, Abdel-Azim, Menon, & Reiss, 2003; Moran, Wig, Adams, Janata, & Kelley, 2004) have found humor-related activation in other regions, including the right cerebellum, the left temporal–occipital gyrus, the left dorsal anterior cingulate, and (consistent with the cognitive debugging theory of Hurley, Dennett, and Adams, with the medial ventral prefrontal cortex activity associated with funniness by Goel and Dolan, and with the subjective experience of humor as pleasurable) a set of subcortical structures that are associated with the mesolimbic dopaminergic reward system: ventral striatum, the nucleus accumbens, the ventral tegmental area, and the amygdala.

The experimental results here are complemented by other evidence that people can and do naturally detect and react to small probabilistic differences in the formal structure of words. One relevant example is work on the Phonological-Distributional Coherence Hypothesis (Monaghan, Christiansen, Farmer, & Fitneva, 2010) which provided behavioral evidence that people are sensitive to subtle formal (in this case, phonological) differences between nouns and verbs that have been documented in many different languages (English, Dutch, French, and Japanese; Monaghan, Christiansen, & Chater, 2007). Binder, Medler, Westbury, Liebenthal, and Buchanan (2006) put the orthographic analogue of this phenomenon in a neurological context by demonstrating that the fusiform gyrus shows stepped activity at an early stage of word processing to different degrees of orthographic word-likeness in NW strings, even in a task that required participants to make no judgment about the word-likeness of those strings. The fact that the fusiform gyrus has reciprocal functional connectivity to the amygdala (Herrington, Taylor, Grupe, Curby, & Schultz, 2011) provides a pathway for connecting experienced emotion to perceived letter probability patterns. One plausible (albeit speculative) mechanism for the experienced humor of visual recognition—such as the humor we have documented here or the humor that makes us laugh involuntarily after realizing that the man we thought we glimpsed lurking in our backyard is just the neighbor's cat—may lie in a rapid de-activation of the amygdala very soon after it is activated. The laughter that is often engendered by relief of this sort may serve a useful and plausibly adaptive communicative role by sending message to others that *I thought there was something dangerous out there [and you might think so too] but now I know that there is not*. We may therefore find NWs humorous because it has proven adaptive across evolutionary time for us to be structured in a way that makes us involuntarily let conspecifics know about anomalies that we have recognized are not at all dangerous, since anomalies are generally experienced as frightening (for an extensive discussion, see Hirsh, Mar, & Peterson, 2012; Peterson, 1999).

This work we have presented is also is consistent with other work showing that the properties of NW strings can induce consistent semantic intuitions. Perhaps the most famous examples of the phenomenon are the sound symbolic observations of Köhler (1929/1947) and Sapir (1929). As mentioned above, Köhler pointed out (without initially even feeling, in 1929, that evidence was required) that people were more likely to associate the terms 'baluma' (1929) or 'maluma' (1947) with a curvy shape and the term 'takete' with a spiky shape, a claim that latter found much empirical support (by 1947, the wording in Köhler's book implies that he had actually gathered some data; more rigorous support for Köhler's claim has been collected by Davis (1961), Holland and Wertheimer (1964), and Ramachandran and Hubbard (2001)). Sapir (1929) reported that 80% of several hundred subjects

associated the large-voweled string 'mal' (contrasted with the small-voweled 'mil') with a large table. Reilly et al. (2012) showed that people had reliably-consistent intuitions based on form about whether a NW referred to something concrete or abstract.

The reliability with which NW humor judgments can be predicted suggests the possibility that it might find practical applications, especially with more carefully designed strings, perhaps as a means of detecting alterations to humor appreciation following brain damage or pathological mood impairment. The effect may also have practical effects in product naming, if it can be shown that the computable funniness in a name is a relevant factor in consumer behavior. We predict that consumers will strongly prefer (funny NWs) 'whook' or 'mamessa' to (unfunny NWs) 'turth' or 'suppect' for a new product name.

We began this paper by citing William James' observation that words are not only about semantics, but also about *feeling*. The emotionality of humor detection depends on the categorical expectation violation inherent in being presented with a NW in place of (the expected default category) a word. Since words themselves do not violate our expectations of reading a word, they may not be found humorous at all. However, since there is a wide variety of typicality in words that can be captured using the same calculations we have outlined in this paper (that is, since words themselves vary in their entropy), it may be worth looking into the possibility that something of what James was thinking about is amenable to the same kind of quantification. In particular, we speculate that people might prefer a low-entropy word over a synonymous higher-entropy word in contexts in which the intention is to evoke a light or happy emotion, and vice versa in contexts intended to evoke a weighty or serious emotion.

Acknowledgments

This curiosity-based work was unfunded. Thanks to Tom Spalding and Pete Dixon for methodological advice; to Alex Porthukaran for compiling the Seuss words; to several RAs who helped collect the experimental data; to Oliver Schweickart for German translation; and to James Adelman, Matt Hurley, and seven anonymous reviewers for feedback on earlier versions of this manuscript.

Appendix A. Sub-strings used to identify rude-alluding NWs

ass
az
bag
bum
but
coc
cok
cum
cun
dic
dik
fuc
fuk
grab
hoal
hole
horn
hyn
ipple
jam
kun
lic
lik
loo
nip
nut
ock
oin
pee
perc
perk
poo
pork
pu
shat
shit
slit
spit
suc
suk
teet
tit
tush
uct
wana
wang
wild
womb
dong
snot
unt
hump

Appendix B. 100 stimuli pairs used in the experiments in this paper, ordered by increasing unigram entropy (=decreasing predicted humorousness)

NW1	Entropy	NW2	Entropy
subvick	1.04	suppopp	1.11
supplic	1.24	bollyze	1.24
howaymb	1.28	prousup	1.28
proffic	1.32	qualrev	1.32
quarban	1.34	himumma	1.34
quingel	1.36	suprovi	1.36
finglam	1.38	probble	1.38

Appendix B (*continued*)

NW1	Entropy	NW2	Entropy
grappor	1.40	fityrud	1.40
proffin	1.41	morying	1.41
webruic	1.42	dolsimp	1.42
pranomp	1.43	prompan	1.43
digiday	1.44	belysty	1.44
witypro	1.45	rembrob	1.46
currimi	1.46	partify	1.46
bomysta	1.47	advical	1.47
prounds	1.48	rundsop	1.48
prothly	1.49	pachang	1.49
disingu	1.50	proundi	1.50
chanywa	1.51	devivar	1.51
dockles	1.51	filisma	1.52
wiligna	1.52	commari	1.51
compide	1.53	pervica	1.53
engally	1.53	suprega	1.53
locidev	1.54	forings	1.54
expisho	1.55	commedi	1.55
wilysta	1.55	sanybon	1.55
gremsev	1.56	moutlem	1.56
dakinge	1.56	larlize	1.57
micaust	1.57	becadev	1.57
mervirs	1.58	togriva	1.58
throvic	1.59	jusessa	1.59
devervi	1.59	vervide	1.59
betwerv	1.60	posiver	1.60
shaptur	1.60	specity	1.60
clopect	1.61	limanol	1.61
allysti	1.61	gramile	1.61
articlu	1.62	clarral	1.62
strompa	1.62	thregul	1.62
uranged	1.63	suption	1.63
abysted	1.64	montsim	1.64
allopti	1.64	groctsi	1.64
mempise	1.64	varketa	1.64
defasom	1.65	throdyn	1.65
iningla	1.66	shoureg	1.66
larshic	1.66	raddreg	1.66
menclar	1.66	shouges	1.66
maricts	1.67	servica	1.67
numersi	1.68	howinge	1.68
dispont	1.68	grashin	1.68
comorti	1.68	benervi	1.69
tessigg	1.69	fainges	1.69
smadian	1.70	delpith	1.70
messago	1.70	thervil	1.70
bengsta	1.71	pargest	1.71
scrunne	1.71	anathly	1.71
persica	1.72	lifinta	1.72
asplati	1.72	domeave	1.72
forsith	1.73	ausenus	1.73
hignoth	1.73	throuss	1.73
walthel	1.73	premser	1.73
pariono	1.74	nothrog	1.74
crivere	1.75	montems	1.75
agallea	1.75	avelian	1.75

Appendix B (*continued*)

NW1	Entropy	NW2	Entropy
loarlai	1.75	systina	1.75
spanoth	1.76	vernion	1.76
caltsio	1.77	exturea	1.77
thramon	1.77	deffent	1.77
wessish	1.78	sounano	1.78
dessaga	1.78	relysta	1.78
plareed	1.78	telcout	1.79
sersimi	1.79	worinta	1.79
ostallo	1.79	howseri	1.79
mesimie	1.80	somersi	1.80
cortsio	1.81	arthrol	1.81
systess	1.81	incleas	1.81
mesterf	1.82	englita	1.82
kentsia	1.82	strewri	1.83
splenne	1.83	refends	1.83
sertsim	1.84	slanoth	1.84
materal	1.84	thinatu	1.84
angessa	1.85	engstin	1.85
prensta	1.85	mestins	1.86
perecti	1.86	pertice	1.86
afrithe	1.87	hastems	1.87
rousent	1.87	mesects	1.87
chertin	1.88	trinche	1.88
memsere	1.89	pritent	1.89
talisti	1.90	weastin	1.90
ancessa	1.90	sectori	1.90
arcents	1.91	screnta	1.91
segessa	1.92	clester	1.92
tueredo	1.93	anottac	1.93
mestead	1.94	sersice	1.94
amatera	1.95	slannet	1.95
opelese	1.96	whetesi	1.96
edisted	1.98	anotain	1.98
nathess	2.00	meentra	2.00
retsits	2.02	tessina	2.02
heashes	2.05	anceste	2.05
octeste	2.09	tatinse	2.08

References

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

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management, 39*, 45–65.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723.

Attardo, D., & Raskin, V. (1991). Script theory revis(it)ed: Joke similarity and joke representation model. *Humor: The International Journal of Humor Research, 4*, 293–347.

Baayen, R. H. (2008). *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. H. (2010). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon, 5*, 436–461.

Baayen, R. H. (2012). Learning from the Bible: Computational modeling of the costs of letter transpositions and letter exchanges in reading Classical Hebrew and Modern English. *Lingue & Linguaggio, 2*, 123–146.

Baayen, R. H., Hendrix, P., & Ramscar, M. (2013). Sidestepping the combinatorial explosion: Towards a processing model based on discriminative learning. *Language and Speech, 56*, 329–347.
Baayen, R. H., Milin, P., Filipovic Durdevic, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review, 118*, 438–482.
Baayen, R. H., & Ramscar, M. (2015). Abstraction, storage and naive discriminative learning. In E. Dabrowska & D. Divjak (Eds.), *Handbook of cognitive linguistics* (pp. 99–120). Berlin: De Gruyter Mouton.
Bain, A. (1865/2004). *The emotions and the will*. Whitefish, Montana: Kessinger Publishing.
Balsam, P., & Gallistel, C. (2009). Temporal maps and informativeness in associative learning. *Trends in Neurosciences, 32*, 73–78.
Binder, J., Medler, D., Westbury, C., Liebenthal, E., & Buchanan, L. (2006). Tuning of the human left fusiform gyrus to sublexical orthographic structure. *NeuroImage, 33*, 739–748.
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993–1022.
Chafe, W. (2007). *The importance of not being earnest: The feeling behind laughter and humor*. Philadelphia, PA: John Benjamins.
Clark, M. (1970). Humor and incongruity. *Philosophy, 45*, 20–32.
Davis, R. (1961). The fitness of names to drawings: A cross-cultural study in Tanganyika. *British Journal of Psychology, 52*, 259–268.
Freud, S. (1905/1960). *Jokes and their relation to the unconscious* (J. Strachey, Trans.). New York: W. W. Norton.
Gallistel, C. R. (2003). Conditioning from an information processing perspective. *Behavioural Processes, 62*, 89–101.
Gervais, M., & Wilson, D. (2005). The evolution and functions of laughter and humor: A synthetic approach. *The Quarterly Review of Biology, 80*, 395–430.
Goel, V., & Dolan, R. J. (2001). The functional anatomy of humor: Segregating cognitive and affective components. *Nature Neuroscience, 4*, 237–238.
Herrington, J., Taylor, J., Grupe, D., Curby, K., & Schultz, R. (2011). Bidirectional communication between amygdala and fusiform gyrus during facial recognition. *NeuroImage, 56*, 2348–2355.
Hirsh, J. B., Mar, R. A., & Peterson, J. B. (2012). Psychological entropy: A framework for understanding uncertainty-related anxiety. *Psychological Review, 119*, 304–320.
Holland, M. K., & Wertheimer, M. (1964). Some physiognomic aspects of naming, or, maluma and takete revisited. *Perceptual and Motor Skills, 19*, 111–117.
Hurley, M. H., Dennett, D. C., & Adams, R. B. (2011). *Inside jokes: Using humor to reverse-engineer the mind*. Cambridge, MA: MIT Press.
Hutcheson, F. (1725/1973). *An inquiry concerning beauty, order, harmony, design*. The Hague, Netherlands: Martinus Nijhoff.
James, W. (1890/1950). *The principles of psychology* (Vol. 1). New York, New York: Dover Publications Inc.
Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology, 66*, 115–124.
Jung, W. E. (2003). The inner eye theory of laughter: Mind reader signals cooperator value. *Evolutionary Psychology, 1*, 214–253.
Köhler, W. (1929). *Gestalt psychology*. New York: Liveright.
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.
Lipps, T. (1898). *Komik und Humor: Eine Psychologisch-Ästhetische Untersuchung* [*Comedy and humor: A psychological and aesthetic investigation*]. <https://www.gutenberg.org/cache/epub/8298/pg8298.txt>.
Martin, M. W. (1983). Humor and aesthetic enjoyment of incongruities. *British Journal of Aesthetics, 23*, 74–85.
McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech, 44*, 295–322.
Mobbs, D., Greicius, M. D., Abdel-Azim, E., Menon, V., & Reiss, A. L. (2003). Humor modulates the mesolimbic reward centers. *Neuron, 40*, 1041–1048.
Monaghan, P., Christiansen, M., & Chater, N. (2007). The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology, 55*, 259–305.
Monaghan, P., Christiansen, M., Farmer, T., & Fitneva, S. (2010). Measures of phonological typicality: Robust coherence and psychological validity. *The Mental Lexicon, 5*, 281–299.
Moran, J. M., Wig, G. S., Adams, R. B., Jr., Janata, P., & Kelley, W. M. (2004). Neural correlates of humor detection and appreciation. *NeuroImage, 21*, 1055–1060.
Mulder, K., Dijkstra, T., Schreuder, R., & Baayen, R. H. (2014). Effects of primary and secondary morphological family size in monolingual and bilingual word processing. *Journal of Memory and Language, 72*, 59–84.
Peterson, J. (1999). *Maps of meaning: The architecture of belief*. New York: Routledge.
Polimeni, J., & Reiss, J. P. (2006). The first joke: Exploring the evolutionary origins of humor. *Evolutionary Psychology, 4*, 347–366.
R Core Team (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. <http://www.R-project.org/>.
Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia—A window into perception, thought and language. *Journal of Consciousness Studies, 8*, 3–34.
Ramscar, M., Dye, M., & McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of mouses in adult speech. *Language, 89*, 760–793.
Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, R. H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in Cognitive Science, 6*, 5–42.
Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods, 41*, 647–656.
Reilly, J., Westbury, C., Kean, J., & Peele, J. (2012). Arbitrary symbolism in natural language revisited: When word forms carry meaning. *PLoS ONE, 7*, e42286. http://dx.doi.org/10.1371/journal.pone.0042286.
Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think. *American Psychologist, 43*, 151–160.
Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current theory and research* (pp. 64–99). New York: Appleton-Century-Crofts.
Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology, 12*, 239–255.
Schopenhauer, A. (1818). *The world as will and representation*. London, England: Ballantyne, Hanson, and Company.
Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 623–656.
Shaoul, C., & Westbury, C. (2006). *USENET orthographic frequencies for 111,627 English words (2005–2006)*. Edmonton, AB: University of Alberta. <http://www.psych.ualberta.ca/~westburylab/downloads/wlfreq.download.html>.
Singmann, H., & Bolker, B. (2014). *afex: Analysis of factorial experiments. R package version 0.12-135*. <http://CRAN.R-project.org/package=afex>.
Suls, J. (1972). A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. In J. H. Goldstein & P. E. McGhee (Eds.), *The psychology of humor: Theoretical perspectives and empirical issues* (pp. 81–100). New York: Academic Press.
Telfer, E. (1995). Hutcheson's reflections on laughter. *The Journal of Aesthetics and Art Criticism, 53*, 359–369.
Westbury, C. (2005). Implicit sound symbolism in lexical access: Evidence from an interference task. *Brain and Language, 93*, 10–19.
Westbury, C. (2007). *ACTUATE: Assessing cases*. The University of Alberta Testing Environment. <http://www.psych.ualberta.ca/~westburylab/downlaods/alfab.download.html>.
Westbury, C., Hollis, G., & Shaoul, C. (2007). LINGUA: The language-independent neighborhood generator of the University of Alberta. *The Mental Lexicon, 2*, 273–286.
Wittgenstein, L. (1953). *Philosophical investigations*. New York: Macmillan.