Title:

A computational model of linguistic humor in puns

Authors:

Justine T. Kao, Department of Psychology, Stanford University

Roger Levy, Department of Linguistics, UCSD

Noah D. Goodman, Department of Psychology, Stanford University

Corresponding author:

Justine T. Kao (justinek@stanford.edu)

Jordan Hall, Building 420, 450 Serra Mall, Stanford, CA 94305

Keywords: humor, computational modeling, language processing, ambiguity, noisy-channel

Abstract

Humor plays an essential role in human interactions. Precisely what makes something funny, however, remains elusive. While research on natural language understanding has made significant advancements in recent years, there has been little direct integration of humor research with computational models of language understanding. In this paper, we propose two information-theoretic measures—ambiguity and distinctiveness—derived from a simple model of sentence processing. We test these measures on a set of puns and regular sentences and show that they correlate significantly with human judgments of funniness. Moreover, within a set of puns, the distinctiveness measure distinguishes exceptionally funny puns from mediocre ones. Our work is the first, to our knowledge, to integrate a computational model of general language understanding and humor theory to quantitatively predict humor at a fine-grained level. We present it as an example of a framework for applying models of language processing to understand higher-level linguistic and cognitive phenomena.

1. Introduction

Love may make the world go round, but humor is the glue that keeps it together. Our everyday experiences serve as evidence that humor plays a critical role in human interactions and composes a significant part of our linguistic, cognitive, and social lives. Previous research has shown that humor is ubiquitous across cultures (Martin, 2010; Kruger, 1996), increases interpersonal attraction (Lundy, Tan & Cunningham, 1998), helps resolve intergroup conflicts (Smith, Harrington & Neck, 2000), and improves psychological wellbeing (Martin, Kuiper, Olinger & Dance, 1993). However, little is known about the cognitive basis of such a pervasive and enjoyable experience. By providing a formal model of linguistic humor, we aim to solve part of the mystery of what makes us laugh.

Theories of humor have existed since the time of Plato and Aristotle (see Attardo, 1994 for review). A leading theory in modern research posits that incongruity, loosely characterized as the presence of multiple incompatible meanings in the same input, may be critical for humor (Koestler, 1964; Veale, 2004; Forabosco, 1992; McGhee, 1979; Martin, 2007; Hurley, Dennett, & Adams, 2011; Vaid & Ramachandran, 2001). However, despite relative consensus on the importance of incongruity, definitions of incongruity vary across informal analyses of jokes. As Ritchie (2009) wrote, “There is still not a rigorously precise definition that would allow an experimenter to objectively determine whether or not incongruity was present in a given situation or stimulus” (p. 331). This lack of precision makes it difficult to empirically test the role of incongruity in humor or extend these ideas to a concrete computational understanding. On the other hand, most work on computational humor focuses either on joke-specific templates and schemata (Binsted, 1996, Taylor & Mazlack, 2004) or surface features and properties of individual words (Mihalcea & Strapparava, 2006; Kiddon & Brun, 2011; Reyes, Rosso & Buscaldi, 2012). One exception is Mihalcea et al. (2010), which used features inspired by incongruity theory to detect humorous punch lines; however, the incongruity features proposed did not significantly outperform a random baseline, leading the authors to conclude that joke-specific features may be preferable. While these dominant approaches in computational humor are able to identify humorous stimuli within certain constraints, they fall short of testing a more general cognitive theory of humor.

In this work, we suggest that true measures of incongruity in linguistic humor may require a model that infers meaning from words in a principled manner. We build upon theories of humor and language processing to formally measure the multiplicity of meaning in puns -- sentences "in which two different sets of ideas are expressed, and we are confronted with only one series of words," as described by Philosopher Henri Bergson (Bergson, 1914). Puns provide an ideal test bed for our purposes because they are simple, humorous sentences with multiple meanings. Here we focus on phonetic puns, defined as puns containing words that sound identical or similar to other words in English[1]. The following is an example:

(1) "The magician got so mad he pulled his hare out."

Although the sentence's written form unambiguously contains the word "hare," previous work has suggested that phonetic representations play a central role in language comprehension even during reading (Niznikiewicz & Squires, 1996; Pexman et al., 2001; Pollatsek et al., 1992).

---

[1] An early version of this work appeared in the proceedings of the 35th Annual Meeting of the Cognitive Science Society. In this extended paper, we examine a wider range of sentences, including puns that contain identical homophones as well as puns with words that sound similar (but not identical) to other words in English.

Taking the lexical ambiguity of its phonetic form into account, this sentence thus implicitly expresses two "ideas," or meanings[2]:

(1a) The magician got so mad he performed the trick of pulling a rabbit out of his hat.

(1b) The magician got so mad he pulled out the hair on his head.

At the most basic level, the humor in this pun relies on the fact that it contains the word "hare," which is phonetically confusable with "hair." However, the following sentence also contains a phonetically ambiguous word, but is clearly not a pun:

(2) "The hare ran rapidly across the field."

A critical difference between (1) and (2) is that *hare* and *hair* are both probable meanings in the context of sentence (1), whereas *hare* is much more likely than *hair* in sentence (2). From this informal analysis, it seems that both meanings are compatible with context in a phonetic pun, suggesting that a sentence must contain ambiguity to be funny. However, another example shows that ambiguity alone is insufficient. Consider the sentence:

(3) "Look at that hare."

This sentence is also ambiguous between *hare* and *hair*, but is unlikely to elicit chuckles. A critical difference between (1) and (3) is that while each meaning is strongly supported by distinct groups of words in (1) (*hare* is supported by "magician" and "hare"; *hair* is supported by "mad" and "pulled"), both meanings are weakly supported by all words in (3). This comparison suggests that in addition to ambiguity, distinctiveness of support may also be an important criterion for humor. Observations on the putative roles of ambiguity of sentence meaning and

[2] In this work we focus on written sentences that contain phonetic ambiguity. In the future, it would be interesting to examine humorous effects in spoken sentences, where ambiguity cannot be partially resolved by the orthographic form.

distinctiveness of support will motivate our formal measures of humor. [3]

How should we represent the meaning of a sentence in order to measure its ambiguity and distinctiveness? While formally representing sentence meanings is a complex and largely unsolved problem (Grefenstette et al., 2014; Socher et al., 2012; Liang et al., 2013), we can utilize certain properties of phonetically ambiguous sentences to simplify the problem. We notice that in sentence (1), meaning (1a) arises if the word "hare" is interpreted as *hare*, while meaning (1b) arises if "hare" is interpreted as its homophone *hair*. Each sentence-level meaning directly corresponds to the meaning of a phonetically ambiguous word. As a result, we can represent sentence meaning (1a) with *hare* and (1b) with *hair.* This approximation is coarse and captures only the "gist" of a sentence rather than its full meaning. However, we will show that it is sufficiently powerful for modeling the interpretation of sentences with only a phonetic ambiguity.

Given the space of candidate sentence meanings, a comprehender's task is to infer a distribution over these meanings from the words she observes. Formally, a phonetically ambiguous sentence such as (1) is composed of a vector of words $\vec{w} = \{w_1, \ldots, w_i, h, \ w_{i+1}, \ldots, \ \ w_n\}$, where h is phonetically confusable with its homophone h'. The sentence meaning is a latent variable m, which we assume has two possible values m_a and m_b. These sentence meanings can be identified with h and h', respectively. Consistent with a noisy channel approach (Levy, 2008; Levy et al., 2009; Gibson et al., 2013), we construe the task of understanding a sentence as inferring m using probabilistic integration of noisy evidence

[3] Note that it is not necessary for both meanings to be completely compatible with the full context, as illustrated by puns such as *I used to be addicted to soap, but I'm clean now*, in which the most common meaning of *clean* is actually ruled out, rather than supported, by full compositional interpretation of the context. What instead seems necessary is that the support derived from the subset of context for each meaning is balanced.

given by $\vec{w}$. We construct a simple probabilistic generative model that captures the relationship between the meaning of a sentence and the words that compose it (Fig. 1). If a word is semantically relevant ($f_i = 1$), we assume that it is sampled based on semantic relatedness to the sentence meaning; if the word is irrelevant, or "noise," it only reflects general language statistics and is sampled from an n-gram model. Because the comprehender maintains uncertainty about which words are relevant, it is possible for her to arrive at multiple interpretations of a sentence that are each coherent but incongruous with one another, a situation that we hypothesize gives rise to humor. To capture this intuition, we introduce two measures of humor derived from the distribution over sentence meanings (details in Methods section).

-----------------------------Insert Figure 1 about here ------------------------------

Given words in a sentence, we infer the joint probability distribution over sentence meanings and semantically relevant words, which can be factorized into the following:

$$P(m, \vec{f} \mid \vec{w}) = P(m \mid \vec{w})\, P(\vec{f} \mid m, \vec{w}) \qquad (Eq. 1)$$

We compute a measure of humor from each of the two terms on the right-hand side. Ambiguity is quantified by the entropy of the distribution $P(m \mid \vec{w})$. If entropy is high, then the sentence is ambiguous because both meanings are near-equally likely. Distinctiveness captures the degree to which the relevant words differ given different sentence meanings. Given one meaning m_a, we compute $F_a = P(\vec{f} \mid m_a, \vec{w})$. Given another meaning m_b, we compute $F_b = P(\vec{f} \mid m_b, \vec{w})$. Distinctiveness is quantified by the symmetrized Kullback-Leibler divergence between these two distributions, $D_{KL}(F_a \,||\, F_b) + D_{KL}(F_b \,||\, F_a)$. If the symmetrized KL distance is high, it suggests that the two sentence meanings are supported by distinct subsets of words in the sentence.

Derivation details of these two measures are in the Methods section below. We empirically evaluate ambiguity and distinctiveness as predictors of humor in a set of phonetically ambiguous sentences.

2. Methods

2.1. Computing model predictions

We define the ambiguity of a sentence as the entropy of $P(m \mid \vec{w})$, where $\vec{w}$ is a vector of observed content words in a sentence (which contains a phonetically ambiguous word h) and m is the latent sentence meaning. Given the simplifying assumption that the distribution over sentence meanings is not affected by function words, each w_i in $\vec{w}$ is a content word. The distribution over sentence meanings given words can be derived using Bayes' rule:

$$P(m \mid \vec{w}) = \sum_{\vec{f}} P(m, \vec{f} \mid \vec{w})$$

$$\propto \sum_{\vec{f}} P(\vec{w} \mid m, \vec{f}) P(m) P(\vec{f})$$

$$= \sum_{\vec{f}} \left(P(m) P(\vec{f}) \prod_i P(w_i \mid m, f_i) \right) \qquad (Eq. 2)$$

Each value of m is approximated by either the meaning of the observed phonetically ambiguous word h (e.g. "hare" in sentence (1)) or its unobserved homophone h' (e.g. "hair"). We can thus represent $P(m)$ as the unigram frequency of h or h'. For example, $P(m = hare)$ is approximated as proportional to $P(\text{"hare"})$. We assume equal prior probability that each subset of the words is semantically relevant, hence $P(\vec{f})$ is a constant. $P(w_i|m, f_{\square})$ depends on the value of the semantic relevance indicator variable f_i. If $f_i = 1$, w_i is semantically relevant and is sampled in proportion to its relatedness with the sentence meaning m. If $f_i = 0$, then w_i is generated from a noise process and sampled in proportion to its probability given the previous two words in the sentence. Formally,

$$P(w_i|m, f_i) = \begin{cases} P(w_i|m) & \text{if } f_i = 1 \\ P(w_i|bigram_i) & \text{if } f_i = 0 \end{cases} \qquad (Eq. 3)$$

We estimate $P(w_i|m)$ using empirical association measures described in the Experiment section and compute $P(w_i|bigram_i)$ using the Google N-grams corpus (Brants & Franz, 2006). Once we derive $M = P(m\,|\vec{w})$, we compute its information-theoretic entropy as a measure of ambiguity:

$$Amb(M) = -\sum_{k\in\{a,b\}} P(m_k|\vec{w}) \log P(m_k|\vec{w}) \qquad (Eq. 4)$$

We next compute the distinctiveness of words supporting each sentence meaning. Using Bayes' Rule:

$$P(\vec{f}|m, \vec{w}) \propto P(\vec{w}|m, \vec{f})P(\vec{f}|m) \qquad (Eq. 5)$$

Since $\vec{f}$ and m are independent, $P(\vec{f}\mid m) = P(\vec{f})$, which is a constant. Let $F_a = P(\vec{f}\,|m_a, \vec{w})$ and $F_b = P(\vec{f}\,|m_b, \vec{w})$. We compute the symmetrized Kullback-Leibler divergence score $D_{KL}(F_a\,||\,F_b) + D_{KL}(F_b\,||\,F_a)$, which measures the difference between the distribution of supporting words given one sentence meaning and the distribution of supporting words given another sentence meaning. This results in the distinctiveness measure[4]:

$$Dist(F_a, F_b) = \sum_i \left(\ln\left(\frac{F_a(i)}{F_b(i)}\right) F_a(i) + \ln\left(\frac{F_b(i)}{F_a(i)}\right) F_b(i) \right) \quad (Eq. 6)$$

Given these derivations, we conducted the following experiment to implement and test the ambiguity and distinctiveness measures.

2.2. Experiment

We collected 435 sentences consisting of phonetic puns and regular sentences that contain phonetically ambiguous words. We obtained the puns from a website called "Pun of the Day" (http://www.punoftheday.com), which at the time of collection contained over a thousand puns submitted by users. We collected 40 puns where the phonetically ambiguous word has an identical homophone, for example "hare." Since only a limited number of puns satisfied this criterion, a research assistant generated an additional 25 pun sentences based on a separate list of homophone words, resulting in a total of 65 identical-homophone puns. We selected 130 corresponding non-pun sentences from an online version of Heinle's Newbury House Dictionary of American English (http://nhd.heinle.com). 65 of the non-pun sentences contain the ambiguous words observed in the pun sentences (e.g. "hare"); the other 65 contain the unobserved

[4] In addition to the symmetrized KL divergence of Eq. 6, we also experimented with non-symmetrized KL divergence in both directions and found qualitatively identical results.

homophone words (e.g. “hair”)[5]. To test whether our measures generalize to sentences containing phonetically ambiguous words that do not have identical homophones, we collected 80 puns where the phonetically ambiguous word sounds similar (but not identical) to other words in English (e.g. “tooth” sounds similar to “truth”). We also collected 160 corresponding non-pun sentences. Table 1 shows an example sentence from each category. The full set of sentences can be found here: http://web.stanford.edu/~justinek/pun-paper/results.html

------------------------------Insert Table 1 about here -------------------------------

We obtained funniness ratings for each of the 435 sentences. 100 participants on Amazon’s Mechanical Turk[6] rated the 195 sentences that contain identical homophones. Each participant read roughly 60 sentences in random order, counterbalanced for the sentence types, and rated each sentence on funniness (“How funny is this sentence?”) on a scale from 1 (not at all) to 7 (extremely). We removed 7 participants who reported a native language other than English and z-scored the ratings within each participant. A separate group of 160 participants on Mechanical Turk rated the 240 near homophone sentences. Each participant read 40 sentences in random order, counterbalanced for the sentence types, and rated each sentence on funniness on a scale from 1 to 7. We removed 4 participants who reported a native language other than English and z-scored the ratings within each participant. We used the average z-scored ratings across participants as human judgments of funniness for all 435 sentences.

As described in the measure derivations, computing ambiguity and distinctiveness requires the conditional probabilities of each word given a sentence meaning, i.e. $P(w_i \mid m)$. In

[5] Results for the 195 identical homophone sentences were reported in Kao et al. (2012), which was published in the proceedings of the 35th Annual Meeting of the Cognitive Science Society (a non-archival publication).

[6] The sample sizes were chosen such that each sentence would receive roughly 20-30 funniness ratings, in order for the uncertainty in funniness measurement to be reasonably low, while keeping the number of sentences rated by each participant manageably small.

practice, this value is difficult to obtain reliably and accurately in an automated way, such as through WordNet distances or semantic vector space models (Gabrilovich & Markovitch, 2007; Zhang et al., 2011; Mihalcea et al., 2010)[7]. Instead of tackling the challenging problem of automatically learning $P(w_i \mid m)$ from large corpora, we observe that $P(w_i \mid m)$ is related to point wise mutual information (PMI) between w_i and m, an information-theoretic measure defined mathematically as the following:

$$\log \frac{P(w_i, m)}{P(w_i)P(m)} = \log P(w_i|m) - \log P(w_i) \qquad (Eq. 7)$$

Intuitively, PMI captures the relatedness between w_i and m, which can be measured empirically by asking people to judge the semantic relatedness between two words. This allows us to harness people's rich knowledge of the relationships between word meanings without relying solely on co-occurrence statistics in corpora. We assume that the z-scored human ratings of relatedness between two words, denoted $R(w_i, m)$, approximates true PMI. With the proper substitutions and transformations[8] from Eq. 7, we derive the following:

$$P(w_i \mid m) = e^{R(w_i, m)} P(w_i) \qquad (Eq. 8)$$

[7] We experimented with computing these values from corpora in early stages of this work. However, we found that it is difficult to obtain reliable co-occurrence statistics for many word pairs of interest (such as "hare" and "magician"), due to the sparsity of these topics in most corpora. Future work could further explore methods for extracting these types of commonsense-based semantic relationships from corpus statistics.

[8] By assuming $R(w_i, m) = \log \frac{P(w_i, m)}{P(w_i)P(m)}$, we get $R(w_i, m) = \log P(w_i|m) - \log P(w_i)$ from Eq. 7; exponentiating both sides gives us Eq. 8.

To obtain $R(w_i, m)$ for each of the words in the stimuli sentences, function words were removed from each of the sentences in our dataset, and the remaining words were paired with the phonetically ambiguous word h and its homophone h' (e.g., for the pun in Table, [“magician”, “hare”] is a legitimate word pair, as well as [“magician”, “hair”]). This resulted in 1460 distinct word pairs for identical homophone sentences and 2056 word pairs for near homophone sentences. 200 participants on Amazon’s Mechanical Turk rated the semantic relatedness of word pairs for identical homophone sentences. Each participant saw 146 pairs of words in random order and were asked to rate how related each word pair is using a scale from 1 to 10. We removed 5 participants who reported a native language other than English. A separate group of 120 participants rated word pairs for near homophone sentences. We removed 2 participants who reported a native language other than English. Since it is difficult to measure the relatedness of a word with itself, we assume that it is constant for all words and treat it as a free parameter, r. After computing our measures, we fit this parameter to people’s funniness judgments (resulting in r = 13). We used the average z-scored relatedness measure for each word pair to obtain $R(w_i, m)$ and Google Web unigrams to obtain $P(w_i)$. This allowed us to compute $P(w_i \mid m)$ for all word and meaning pairs.

3. Results

We computed an ambiguity and distinctiveness score for each of the 435 sentences (see Methods). We found no significant differences between identical and near homophone puns in terms of funniness ratings ($t(130.91) = 0.13$, $p = 0.896$), ambiguity scores ($t(137.80) = 1.13$, $p = 0.261$), and distinctiveness scores ($t(134.91) = -0.61$, $p = 0.543$), suggesting that ambiguity and distinctiveness are fairly robust to the differences between puns that involve identical or near homophone words. As a result, we collapsed across identical and near homophone sentences for the remaining analyses. We found that ambiguity was significantly higher for pun sentences than non-pun sentences ($t(159.48) = 7.89$, $p < 0.0001$), which suggests that the ambiguity measure successfully captures characteristics distinguishing puns from other phonetically ambiguous sentences. Distinctiveness was also significantly higher for pun sentences than non-pun sentences ($t(248.99) = 6.11$, $p < 0.0001$). Fig. 2 shows the standard error ellipses for the two sentence types in a two-dimensional space of ambiguity and distinctiveness. Although there is a fair amount of noise in the predictors (likely due to simplifying assumptions, the need to use empirical measures of relatedness, and the inherent complexity of humor), pun sentences (both identical and near homophone) tend to cluster at a space with higher ambiguity and distinctiveness, while non-pun sentences score lower on both measures.

------------------------------Insert Figure 2 about here ------------------------------

We constructed a linear mixed-effects model of funniness judgments with ambiguity and distinctiveness as fixed effects, a by-item random intercept, and by-subject random slopes for entropy and distinctiveness. We found that ambiguity and distinctiveness were both highly significant predictors, with funniness increasing as each of ambiguity and distinctiveness increases (Table 2). Furthermore, the two measures capture a substantial amount of the reliable variance in funniness ratings averaged across subjects ($F(2,432) = 74.07$, $R^2 = 0.25$, $p < 0.0001$). A linear mixed effects model including a term for the interaction between ambiguity and distinctiveness (both as fixed effect and by-subjects random slope) showed no significant interaction between the two ($t = 1.39$, $p > 0.15$).

------------------------------Insert Table 2 about here ------------------------------

We then examined whether the measures are able to go beyond distinguishing puns from non-puns to predicting fine-grained levels of funniness within puns. We found that ambiguity does not correlate with human ratings of funniness within the 145 pun sentences ($r = 0.03$, $p = 0.697$). However, distinctiveness ratings correlate significantly with human ratings of funniness within pun sentences ($r = 0.28$, $p < 0.001$). By separating the puns into four equal bins based on their distinctiveness, we found that puns with distinctiveness measures in the top-most quartile were significantly funnier than puns with distinctiveness measures in the lower quartiles ($t(90.15) = 3.41$, $p < 0.001$) (Fig. 3). This suggests that while ambiguity helps distinguish puns from non-puns, high distinctiveness characterizes exceptionally humorous puns. To our knowledge, our model provides the first quantitative measure that predicts fine-grained levels of funniness within humorous stimuli.

------------------------------Insert Figure 3 about here ------------------------------

Besides predicting the funniness of a sentence, the model can also be used to reveal critical features of each pun that make it amusing. For each sentence, we identified the set of words that is most likely to be semantically relevant given $\vec{w}$ and each sentence meaning m. Formally, we computed $\arg\max_{\vec{f}} P(\vec{f} \mid m_a, \vec{w})$ and $\arg\max_{\vec{f}} P(\vec{f} \mid m_b, \vec{w})$. Table 3 shows a group of identical-homophone sentences and a group of near-homophone sentences. Sentences in each group contain the same pair of candidate meanings for the homophone; however, they differ on ambiguity, distinctiveness, and funniness. Words that are most likely to be relevant given sentence meaning m_a are in boldface; words that are most likely to be relevant given m_b are in italics. Qualitatively, we observe that the two pun sentences (which are significantly funnier) have more distinct and balanced sets of meaningful words for each sentence meaning than other sentences in their groups. Non-pun sentences tend to have no words in support of the meaning that was not observed. Furthermore, the boldfaced and italicized words in each pun sentence are what one might intuitively use to explain why the sentence is funny—for example, the fact that magicians tend to perform magic tricks with hares, and people tend to be described as pulling out their hair when angry.

------------------------------Insert Table 3 about here ------------------------------

4. Discussion

In this paper, we presented a simple model of gist-level sentence processing and used it to derive formal measures that predict human judgments of humor in puns. We showed that a noisy-channel model of sentence processing facilitates flexible context selection, which enables a single series of words to express multiple meanings. Our work is one of the first to integrate a computational model of sentence processing to analyze humor in a manner that is both intuitive and quantitative. In addition, it is the first computational work to our knowledge to go beyond classifying humorous versus regular sentences to predict fine-grained funniness judgments within humorous stimuli.

The idea of deriving measures of humor from a model of general language understanding is closely related to previous approaches, where humor is analyzed within a framework of semantic theory and language comprehension. Raskin’s (1985) Semantic Script Theory of Humor (SSTH) builds upon a theory of language comprehension in which language is understood in terms of scripts. Under this analysis, a text is funny when it activates two scripts that are incompatible with each other. This theory explains a number of classic jokes where the punch line introduces a script that is incongruous with the script activated by the joke’s setup.

Attardo and Raskin (1991) proposed a revision to SBST in the General Theory of Verbal Humor (GTVH), which details six hierarchically organized knowledge resources that inform the understanding of texts as well as the detection of humor. Nirenburg and Raskin (2004) further formalized the ideas proposed in SBST and GTVH by developing a system for computational semantics termed Ontological Semantics, which includes a large concept ontology, a repository of facts, and an analyzer that translates texts into an ontology-based knowledge representation. This system provides rich ontological knowledge to support in-depth language comprehension and has been applied productively to a variety of domains (Nirenburg and Raskin, 2004; Beale et al., 2004; Taylor et al., 2011). Hempelmann et al. (2006) used a classic joke to show that an extension to the Ontological Semantics system can in principle detect as well as generate humorous texts. However, to our knowledge the system has not yet been tested on a larger body of texts to demonstrate its performance in a quantitative manner (Raskin, 2008; Taylor, 2010). While providing detailed analyses that reveal many important characteristics of humor, much of the work on formalizing humor theories falls short of predicting people's fine-grained judgments of funniness for a large number of texts (Raskin & Attardo, 1994; Ritchie 2001; Attardo et al. 2002; Hempelmann, 2004; Veale, 2006; Brône et al., 2006). In this regard, we believe that our work advances the current state of formal approaches to humor theory. By implementing a simple but psychologically motivated computational model of sentence processing, we derived measures that distinguish puns from regular sentences and correlate significantly with fine-grained humor ratings within puns. Our approach also provides an intuitive but automatic way to identify features that make a pun funny. This suggests that a probabilistic model of general sentence processing (even without the support of rich ontological semantics) may enable powerful explanatory measures of humor.

In addition to advancing computational approaches, our work contributes to cognitive theories of humor by providing evidence that different factors may account for separate aspects of humor appreciation. Some humor theorists argue that while incongruity is necessary for humor, resolving incongruity—discovering a cognitive rule that explains the incongruity in a logical manner—is also key (Ritchie, 1999; Ritchie, 2009; Suls, 1972). We can construe our measures as corresponding roughly to incongruity and resolution in this sense, where ambiguity represents the presence of incongruous sentence meanings, and distinctiveness represents the degree to which each meaning is strongly supported by different parts of the stimulus. Our results would then suggest that incongruity distinguishes humorous input from regular sentences, while the intensity of humor may depend on the degree to which incongruity is resolved by focusing on two different supporting sets. Future work could more specifically examine the relationship between incongruity resolution and the measures presented in our framework.

Although our task in this paper was limited in scope, it is a step towards developing computational models that explain higher-order linguistic phenomena such as humor. To address more complex jokes, future work may incorporate more sophisticated models of language understanding to consider the time course of sentence processing (Kamide et al., 2003; McRae et al., 1998), effects of pragmatic reasoning and background knowledge (Kao et al., 2014a; Kao et al., 2014b), and multi-sentence discourse (Polanyi, 1988; Chambers & Jurafsky, 2008). Our approach could also benefit greatly from the rich commonsense knowledge encoded in the Ontological Semantics system and may be combined with it to measure ambiguity and distinctiveness at the script level rather than at the level of the sentence.

Previous research on creative language use such as metaphor, idioms, and irony has contributed a great deal to our understanding of the cognitive mechanisms that enable people to infer rich meanings from sparse and often ambiguous linguistic input (Lakoff & Turner, 2009; Nunberg et al. 1994; Gibbs & O'Brien, 1991). We hope that our work on humor contributes to theories of language understanding to account for a wider range of linguistic behaviors and the social and affective functions they serve. By deriving the precise properties of sentences that make us laugh, our work brings us one step closer to understanding that funny thing called humor (pun intended).

References

Attardo, S. (1994). *Linguistic theories of humor.* Walter de Gruyter.

Attardo, S. (2001). *Humorous texts: A semantic and pragmatic analysis* (Vol. 6). Walter de Gruyter.

Attardo, S., & Raskin, V. (1991). Script theory revis (it) ed: Joke similarity and joke representation model. *Humor: International Journal of Humor Research*.

Attardo, S., Hempelmann, C. F., & Di Maio, S. (2002). Script oppositions and logical mechanisms: Modeling incongruities and their resolutions. *Humor: International Journal of Humor Research.*

Beale, S., Lavoie, B., McShane, M., Nirenburg, S., & Korelsky, T. (2004, July). Question answering using ontological semantics. In *Proceedings of the 2nd Workshop on Text Meaning and Interpretation* (pp. 41-48). Association for Computational Linguistics.

Bergson, H. (1914). *Laughter: An essay on the meaning of the comic.* Macmillan.

Binsted, K. (1996). *Machine humour: An implemented model of puns*.

Brants, T. & Franz, A. (2006). Web 1T 5-gram Version 1 LDC2006T13. Web Download. *Philadelphia: Linguistic Data Consortium*.

Brône, G., Feyaerts, K., & Veale, T. (2006). Introduction: Cognitive linguistic approaches to humor. *Humor–International Journal of Humor Research*, 19(3), 203-228.

Chambers, N., & Jurafsky, D. (2008). Unsupervised Learning of Narrative Event Chains. In *ACL* (pp. 789-797).

Gabrilovich, E., & Markovitch, S. (2007). Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *IJCAI* (Vol. 7, pp. 1606-1611).

Gibbs, R. W., & O'Brien, J. (1991). Psychological aspects of irony understanding. *Journal of pragmatics*, 16(6), 523-530.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20), 8051-8056.

Grefenstette, E., Sadrzadeh, M., Clark, S., Coecke, B., & Pulman, S. (2014). Concrete sentence spaces for compositional distributional models of meaning. *In Computing Meaning* (pp. 71-86).

Hempelmann, C. F. (2004). Script opposition and logical mechanism in punning. *Humor-International Journal of Humor Research*, 17(4), 381-392.

Hempelmann, C., Raskin, V., & Triezenberg, K. E. (2006). Computer, Tell Me a Joke... but Please Make it Funny: Computational Humor with Ontological Semantics. In *FLAIRS* Conference (pp. 746-751).

Kamide, Y., Altmann, G., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, 49(1), 133-156.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014a). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences,* 111(33), 12002-12007.

Kao, J. T., Bergen, L., & Goodman, N. D. (2014). Formalizing the pragmatics of metaphor understanding. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*.

Kiddon, C., & Brun, Y. (2011). That's what she said: double entendre identification. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* , 89-94.

Kruger, A. (1996). The nature of humor in human nature: Cross-cultural commonalities. *Counselling Psychology Quarterly , 9* (3), 235-241.

Lakoff, G., & Turner, M. (2009). *More than cool reason: A field guide to poetic metaphor*. University of Chicago Press.

Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. *In Proceedings of the conference on empirical methods in natural language processing* (pp. 234-243). Association for Computational Linguistics.

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, 106(50), 21086-21090.

Liang, P., Jordan, M. I., & Klein, D. (2013). Learning dependency-based compositional semantics. *Computational Linguistics*, *39*(2), 389-446.

Lundy, D. E., Tan, J., & Cunningham, M. R. (1998). Heterosexual romantic preferences: The importance of humor and physical fitness for different types of relationships. *Personal Relationships , 5* (3), 311-325.

Martin, R. A., Kuiper, N. A., Olinger, L., & Dance, K. A. (1993). Humor, coping with stress, self-concept, and psychological well-being. *Humor: International Journal of Humor Research*.

Martin, R. (2010). *The psychology of humor: An integrative approach.*

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38(3), 283-312.

Mihalcea, R., & Strapparava, C. (2006). Learning to laugh (automatically): Computational models for humor recognition. *Computational Intelligence , 22* (2), 126-142.

Mihalcea, R., Strapparava, C., & Pulman, S. (2010). Computational models for incongruity detection in humour. In *Computational Linguistics and Intelligent Text Processing* (pp. 364-374). Springer Berlin Heidelberg.

Nirenburg, S., & Raskin, V. (2004). *Ontological semantics*. MIT Press.

Niznikiewicz, M., & Squires, N. K. (1996). Phonological processing and the role of strategy in silent reading: behavioral and electrophysiological evidence. *Brain and language*, 52(2), 342-364.

Nunberg, G., Sag, I. A., & Wasow, T. (1994). Idioms. *Language*, 491-538.

Pexman, P. M., Lupker, S. J., & Jared, D. (2001). Homophone effects in lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(1), 139.

Pollatsek, A., Lesch, M., Morris, R. K., & Rayner, K. (1992). *Phonological codes are used in integrating information across saccades in word identification and reading*. Journal of Experimental Psychology: Human Perception and Performance, 18(1), 148.

Polanyi, L. (1988). A formal model of the structure of discourse. *Journal of pragmatics*, 12(5), 601-638.

Raskin, V. (1985). *Semantic mechanisms of humor* (Vol. 24). Springer Science & Business Media.

Raskin, V. (2008). Theory of humor and practice of humor research: Editor's notes and thoughts. *The Primer of Humor Research*, Berlin, 1-15.

Raskin, V., & Attardo, S. (1994). Non-literalness and non-bona-fide in language: An approach to formal and computational treatments of humor. *Pragmatics & Cognition*, 2(1), 31-69.

Ricoeur, P. (2003). *The rule of metaphor: The creation of meaning in language*. Psychology Press.

Ritchie, G. (1999). *Developing the incongruity-resolution theory.*

Ritchie, G. (2001). Current directions in computational humour. *Artificial Intelligence Review*, 16(2), 119-135.

Ritchie, G. (2009). Variants of Incongruity Resolution. *Journal of Literary Theory*, 3(2), 313-332.

Smith, W. J., Harrington, K. V., & Neck, C. P. (2000). Resolving conflict with humor in a diversity context. *Journal of Managerial Psychology , 15* (6), 606-625.

Socher, R., Huval, B., Manning, C. D., & Ng, A. Y. (2012, July). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 1201-1211). Association for Computational Linguistics.

Suls, J. (1983). Cognitive processes in humor appreciation. *Handbook of humor research* , 39-57.

Suls, J. M. (1972). A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. *The psychology of humor: Theoretical perspectives and empirical issues , 1*, 81-100.

Taylor, J. M. (2010). Ontology-based view of natural language meaning: the case of humor detection. *Journal of Ambient Intelligence and Humanized Computing*, 1(3), 221-234.

Taylor, J., & Mazlack, L. (2004). Computationally recognizing wordplay in jokes. *Proceedings of Annual Meeting of the Cognitive Science Society*.

Taylor, J. M., Raskin, V., & Hempelmann, C. F. (2011). From disambiguation failures to common-sense knowledge acquisition: A day in the life of an Ontological Semantic System. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*-Volume 01 (pp. 186-190). IEEE Computer Society.

Veale, T. (2006). Computability as a test on linguistics theories. *Applications of Cognitive Linguistics,* 1, 461.

Venour, C., Ritchie, G., & Mellish, C. (2010). Quantifying humorous lexical incongruity. In *Proceedings of the International Conference on Computational Creativity* (pp. 268-277).

Zhang, Z., Gentile, A. L., & Ciravegna, F. (2011). Harnessing different knowledge sources to measure semantic relatedness under a uniform model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 991-1002). Association for Computational Linguistics

Homophone	**Type**	**Example**
Identical	Pun	The magician was so mad he pulled his hare out.
Identical	Non-pun	The hare ran rapidly across the field.
Identical	Non-pun	Some people have lots of hair on their heads.
Near	Pun	A dentist has to tell a patient the whole tooth.
Near	Non-pun	A dentist examines one tooth at a time.
Near	Non-pun	She always speaks the truth.

Table 1. Example sentence from each category. Identical homophone sentences contain phonetically ambiguous words that have identical homophones; near homophone sentences contain phonetically ambiguous words that have near homophones. Pun sentences were selected from a pun website; non-pun sentences were selected from an online dictionary (see main text for details).

	Estimate	Std. Error	p-value
Intercept	-2.139	0.306	< 0.0001
Ambiguity	1.915	0.221	< 0.0001
Distinctiveness	0.264	0.040	< 0.0001

Table 2. Regression coefficients using ambiguity and distinctiveness to predict funniness ratings for all 435 sentences; *p*-values are computed assuming that the *t* statistic is approximately normally distributed.

m_a	m_b	Type	Sentence	Amb.	Dist.	Funni.
hare	*hair*	Pun	The **magician** got so mad he *pulled* his **hare** out.	0.15	7.87	1.71
		Non	The **hare ran rapidly** through the **fields**.	$1.43E^{-5}$	7.25	-0.40
tooth	*truth*	Pun	A **dentist** has to *tell* a **patient** the *whole* **tooth**.	0.1	8.48	1.41
		Non	A **dentist** ***examines*** one **tooth** at a time.	$8.92E^{-5}$	7.65	-0.45

Table 3. Semantically relevant words, ambiguity/distinctiveness scores, and funniness ratings for sentences from each category. Words in boldface are semantically relevant to m_a; words in italics are semantically relevant to m_b.

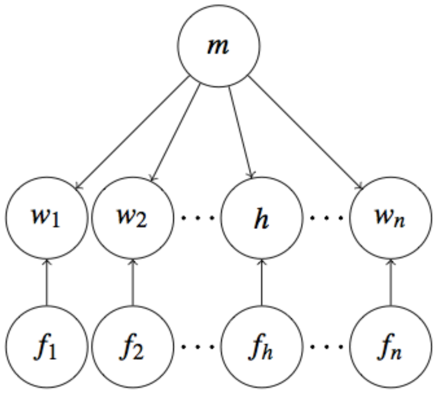

Figure 1. Graphical representation of a generative model of a sentence. If the indicator variable f_i has value 1, w_i is generated based on semantic relatedness to the sentence meaning m; otherwise, w_i is sampled from a trigram language model based on the immediately preceding two words.

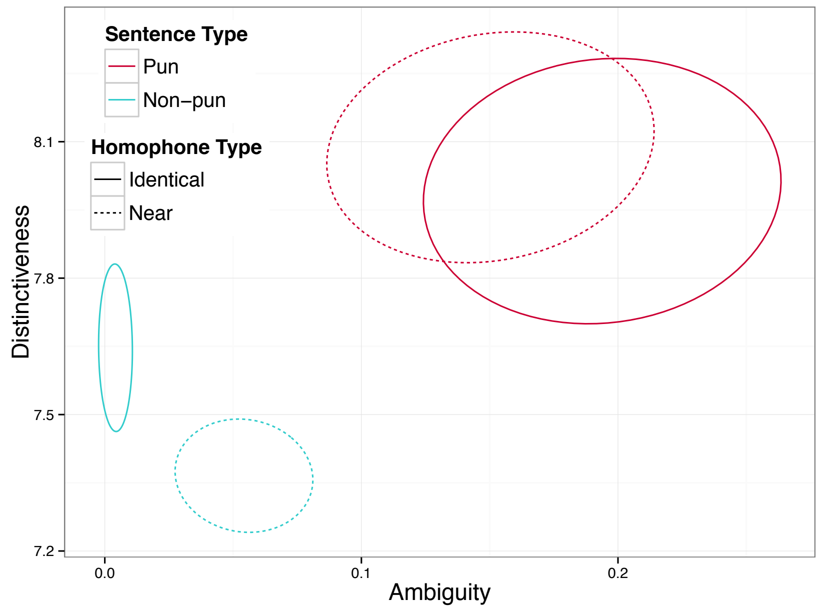

Figure 2. Standard error ellipses of ambiguity and distinctiveness for each sentence type. Puns (both identical and near homophone) score higher on ambiguity and distinctiveness; non-pun sentences score lower.

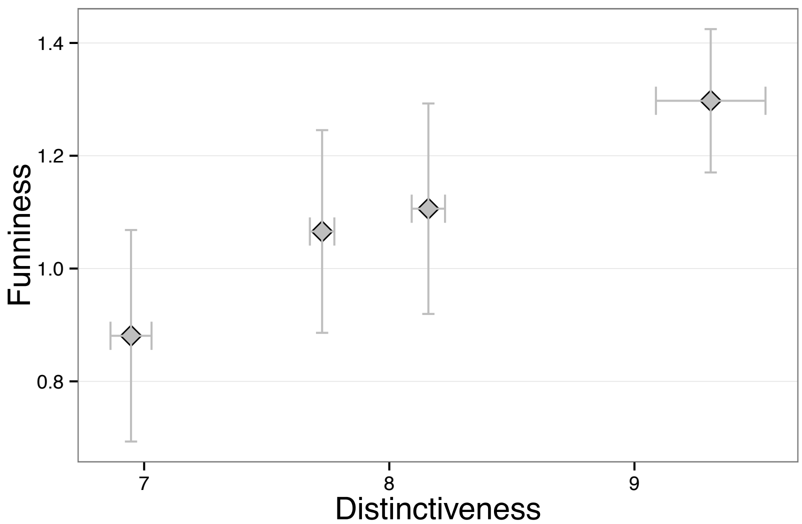

Figure 3. Average funniness ratings and distinctiveness of 145 pun sentences binned according to distinctiveness quartiles. Error bars are confidence intervals.