A Computational Model of Linguistic Humor in Puns

**Abstract**

Humor plays an essential role in human interactions. However, its precise nature 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 then test these measures on a set of puns and regular sentences and show that they correlate significantly with human judgments of funniness. Our model is one of the first to integrate general linguistic knowledge and humor theory to model humor computationally. We present it as an example of a framework for applying models of language processing to understand higher-level linguistic and cognitive phenomena.

**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, the cognitive basis of humor remains largely a mystery. By providing a formal model of linguistic humor, we aim to shed light on the precise properties of sentences that make us laugh.

Many theories of humor have been proposed since Plato and Aristotle (see Attardo, 1994 for review). A leading theory 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 consensus on the importance of incongruity, its precise definitions differ across informal analyses of individual jokes, making it difficult to empirically test the role of multiplicity of meaning in humor on a larger scale. On the other hand, most work on computational humor focuses either on joke-specific templates and schemata (Binsted, 1996) or surface features and properties of individual words (Mihalcea & Strapparava, 2006; Kiddon & Brun, 2011; Reyes, Rosso & Buscaldi, 2012). While these approaches are able to produce or identify humorous stimuli within certain constraints, they fall short of testing a more general cognitive theory of humor.

In this paper, we build upon theories of humor and language processing to formally measure the multiplicity of meaning in puns. Philosopher Henri Bergson described puns as sentences “in which two ideas are expressed, and we are confronted with only one series of words.” Puns provide an ideal test bed for our purposes because they are by definition 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. The following is an example:

(1) “The **magician** got so *mad* he *pulled* his **hare** out.”

The phonetic form of this sentence generates two “ideas,” or meanings:

(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 the word “hare” is phonetically confusable with its homophone “hair.” However, the following sentence contains the same 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 intuitive that ambiguity of meaning is an important criterion for puns. 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 words in bold; *hair* is supported by words in italics), both meanings are weakly supported by all words in (3). This suggests that in addition to ambiguity, distinctiveness of support is also an important criterion for humor. These insights on the roles of ambiguity of sentence meaning and distinctiveness of support motivate our formal measures of humor.

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 make simplifying assumptions. 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.* Although this approximation is coarse and captures the “gist” of a sentence rather than its full meaning, it is sufficiently powerful as a first step for modeling the interpretation of phonetically ambiguous sentences.

Given a space of candidate sentence meanings, a language 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 , where *h* is phonetically confusable with its homophone *h’*. The sentence meaning is a latent variable , which we assume has two possible values and . As described earlier, these sentence meanings can be approximated by the meanings of *h* and *h’*, respectively. Consistent with a noisy channel approach, we formally construe the task of understanding a sentence as inferring using probabilistic integration of noisy evidence given by . We construct a simple generative model that captures the relationship between the meaning of a sentence and the words that compose it (Figure). If a word is a semantically relevant observation, it is sampled based on semantic similarity to the sentence meaning; if the word is irrelevant, or “noise,” it reflects the sentence’s sequential structure 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 each other, which we hypothesize gives rise to humor.

We derive measures of humor from the distribution over sentence meanings. 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:

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 binomial distribution . If entropy is high, then the sentence is ambiguous because both meanings are similarly likely. Distinctiveness captures the degree to which distributions of relevant words differ given different sentence meanings. Given one meaning , we compute . Given another meaning , we compute . Distinctiveness is quantified by the Kullback-Leibler divergence score between these two distributions, ). If the KL score is high, it suggests that the two sentence meanings are supported by distinct subsets of the sentence. Together, ambiguity and distinctiveness constitute a two-dimensional formalization of humor in puns. We empirically evaluate these measures on a set of phonetically ambiguous sentences.

**Methods**

*Measure derivations*

We define the ambiguity of a sentence as the entropy of , where is a vector of observed words in a sentence (which contains a phonetically ambiguous word ) and is the latent sentence meaning. This distribution can be derived using Bayes’ rule[[1]](#footnote-1):

Each value of is approximated by either the meaning of the phonetically ambiguous word or its homophone . We can thus represent as the unigram frequency of or . For example, is approximated as . Since we assume a uniform prior probability over all subsets of the words being semantically relevant, is a constant. depends on the value of the semantic relevance indicator variable . If , is semantically relevant and is sampled in proportion to its relatedness with the sentence meaning . If , then is generated from a noise process and sampled in proportion to its probability given the previous two words (including function words). Formally,

We estimate using measures described in Experiment 2 and compute using the Google Ngrams corpus. Once we derive , we compute its information-theoretic entropy as a measure of ambiguity:

We next compute the distinctiveness of words supporting each sentence meaning Using Bayes’ Rule, we derive the following:

Since and are independent, , which is a constant. Let

Let and . We compute the Kullback-Leibler divergence score ) as a measure of distinctiveness:

We implement and test these measures using the following two experiments.

*Experiment 1*

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 homophone words (e.g. “hair”). 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 shows an example sentence from each category.

|  |  |  |
| --- | --- | --- |
| 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. |

We obtained funniness ratings for each of the 435 sentences. 93 participants on Amazon's Mechanical Turk 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 and correctness. 158 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. We z-scored the ratings and used the average z-scored ratings across participants as human judgments of funniness.

*Experiment 2*

As described in the measure derivations, computing ambiguity and distinctiveness requires the conditional probabilities of each word given a sentence meaning. However, this value is difficult to obtain reliably through traditional topic models trained on corpora due to data sparsity. As a result, we decided to measure it empirically with the following experiment.

We approximated using an empirical measure of the semantic relatedness between and , which we denote as . We construe of as a proxy for point wise mutual information between and , defined as follows:

We assume that human ratings of relatedness between two words approximate true relatedness up to an additive constant and assume . With the proper substitutions and transformations,

To obtain 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. 195 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. 120 participants rated word pairs for near homophone sentences. Since it is difficult to measure the relatedness rating of a word with itself, we used a free parameter and fit it to data (r=13). We used the average z-scored relatedness measure for each word pair to obtain and Google Web unigrams to obtain . This allowed us to compute for all word and meaning pairs.

**Results**

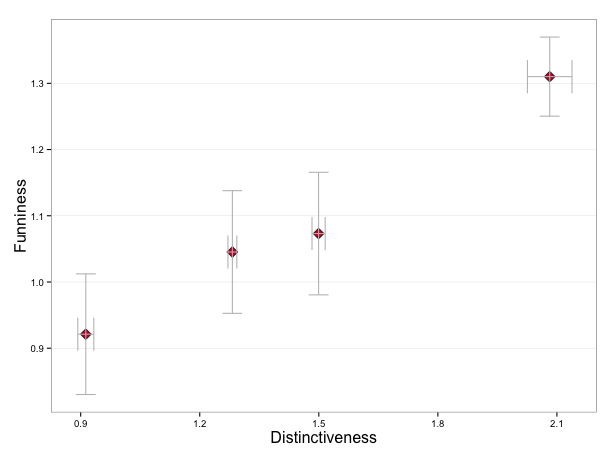
We computed an ambiguity and distinctiveness score for each of the 435 sentences (see Methods). We found that ambiguity was significantly higher for pun sentences than non-pun sentences (F(1, 433) = 108.4, p < 0.0001), which suggests that our ambiguity measure successfully captures characteristics distinguishing puns from other phonetically ambiguous sentences. Distinctiveness was also higher for pun sentences than non-pun sentences, although the difference is only marginally significant (F(1, 433) = 47.1, p < 0.1).

Figure 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. A linear regression model showed that both ambiguity and distinctiveness are significant predictors of funniness ratings across all 435 sentences. Together, the two predictors capture a modest but significant amount of the reliable variance in funniness ratings (F(2,432) = 76.79, r = 0.51, p < 0.0001; see Table).



|  |  |  |  |
| --- | --- | --- | --- |
|  | Estimate | Std. Error | p-value |
| Intercept | -0.830 | 0.107 | < 0.0001 |
| Ambiguity | 1.899 | 0.212 | < 0.0001 |
| Distinctiveness | 0.568 | 0.082 | < 0.0001 |

We now examine whether the measures are able to predict fine-grained levels of funniness within puns. Ambiguity does not correlate with human ratings of funniness within the 145 pun sentences (r = 0.03, p > 0.05). However, distinctiveness ratings correlate significantly with human ratings of funniness within pun sentences (r = 0.28, p < 0.001). Furthermore, puns with distinctiveness measures in the upper quartile were significantly funnier than the other puns. This suggests that while ambiguity distinguishes puns from non-puns, distinctiveness separates very funny puns from mediocre ones. We thus provide the first quantitative measure to our knowledge that predicts fine-grained levels of funniness within humorous stimuli. (Which figure should we show?)



Besides predicting the funniness of a sentence, the model also reveals 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 and a sentence meaning . Formally, we computed and . Table 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 meaningful given sentence meaning are in red; words that are most likely given are in green; and words that are most likely given both meanings are in blue. 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 colorful words in each pun sentence—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—are what one might intuitively use to explain why the sentence is funny. Besides producing quantitative predictions of funniness, the model also provides an intuitive way of explaining which aspects of a pun make it funny.

**Discussion**

In this paper, we presented a simple model of sentence processing and derived formal measures to 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 at once. Our work is one of the first to integrate a general model of sentence processing to analyze humor in an manner that is both intuitive and quantitative. To our knowledge, it is also the first computational work to go beyond classifying humorous versus regular sentences to predict fine-grained funniness judgments within humorous stimuli.

Previous research in artificial intelligence has applied computational tools to identify and generate humorous input, with the hopes of building computers that can interact with humans in a more natural and engaging manner (Mihalcea & Strapparava, 2006). Our work contributes to this literature by incorporating a simple and psychologically driven model of sentence processing to formally capture multiplicity of meaning in sentences. Our measures distinguish puns from regular sentences, correlate significantly with fine-grained humor ratings within puns, and provide an intuitive way of identifying critical features that make a pun funny. This suggests that models of general sentence processing may help derive richer and more explanatory measures of humor.

Besides potential applications in computational humor, 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 manger—is also key (Ritchie, 1999; Suls, 1972). Under this theory, we can construe our measures as each corresponding roughly to incongruity and resolution, 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 logical support. Future work could more specifically examine the roles of incongruity and resolution using a similar framework.

Although our task in this paper is limited in scope, it is a first step towards developing computational models that explain higher-order linguistic phenomena such as humor. Future work may incorporate more sophisticated models of language understanding to consider the time course of sentence processing, deeper semantic representations, and multi-sentence discourse. Besides seeking to understand linguistic humor for its own sake, creative language use can serve as probes for developing models of language processing that account for a wider range of lsinguistic behavior. We believe that our work contributes to research in humor theory, computational humor, and language understanding, such that some day we can build robots that make us laugh and understand the appreciation for humor that makes us uniquely human.

# References

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

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.*

Bartolo, A., Benuzzi, F., Nocetti, L., Baraldi, P., & Nichelli, P. (2006). Humor comprehension and appreciation: an FMRI study. *Journal of Cognitive Neuroscience* *, 18* (11), 1789-1798.

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

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).

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.

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.*

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

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

Samson, A. C., Hempelmann, C. F., Huber, O., & Zysset, S. (2009). Neural substrates of incongruity-resolution and nonsense humor. *Neuropsychologia* *, 47* (4), 1023-1033.

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.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Type | Sentence | Amb. | Dist. | Funni. |
| hare | hair | Pun | The magician got so mad he pulled his hare out. | 0.15 | 1.36 | 1.71 |
| Non | The hare ran rapidly through the fields. | 1.43E-5 | 1.07 | -0.40 |
| Non | Most people have lots of hair on their heads. | 9.47E-11 | 1.55 | -0.34 |
| tooth | truth | Pun | A dentist has to tell a patient the whole tooth. | 0.1 | 1.64 | 1.41 |
| Non | A dentist examines one tooth at a time. | 8.92E-5 | 1.23 | -0.45 |
| Non | She always speaks the truth. | 3.85E-10 | 0.72 | -0.46 |

1. For simplification, our model disregards function words. Each in is a content word. [↑](#footnote-ref-1)