The Funny Thing About Incongruity: A Noisy Channel Model of Puns

Justine Kao[1] (justinek@stanford.edu), **Roger Levy**[2] (rlevy@ucsd.edu), **Noah D. Goodman**[1] (ngoodman@stanford.edu)
[1]Department of Psychology, Stanford University. [2]Department of Linguistics, UCSD.

Abstract

"Researchers showed the robot ten puns, hoping that one of them would make it laugh. Unfortunately, no pun in ten did."

Keywords: Humor; language understanding; noisy channel; probabilistic models; sentence processing

Introduction

Humor plays an essential role in human interactions: it has important positive effects on children's development (Frank & McGhee, 1989), success in the work place (Duncan, Smeltzer, & Leap, 1990), coping with illness and traumatic events (Johnson, 2002; Gelkopf & Kreitler, 1996), and marital satisfaction (Ziv & Gadish, 1989). Indeed, in a study on gender differences in desired characteristics of relationship partners, both men and women rated sense of humor as more important than physical attractiveness and earning potential (Stewart, Stinnett, & Rosenfeld, 2000). In this paper, we are interested in understanding how this fundamental and ubiquitous phenomenon works from the perspective of cognitive science. What makes something funny? How might the defining characteristics of humor shed light on the ways in which the mind processes and evaluates information?

A leading theory of humor posits that incongruity—perceiving a situation from different viewpoints and finding the resulting interpretations to be incompatible—contributes to sensations of mirth (Veale, 2004; Forabosco, 1992; McGhee, 1979; Martin, 2007; Hurley, Dennett, & Adams, 2011); an idea that dates to Kant's theories about laughter and the sublime (Veatch, 1998; Forabosco, 1992). Although there is disagreement about whether incongruity alone is sufficient, most theorists accept that incongruity is necessary for producing humor: as Veale (2004) states, "Of the few sweeping generalizations one can make about humor that are neither controversial or trivially false, one is surely that humor is a phenomenon that relies on incongruity." However, definitions of incongruity are often ambiguous and difficult to operationalize in empirical research. In this paper, we use a computational model of language understanding to formalize a notion of incongruity and test its relationship to humor.

Language understanding in general, and particularly humor, relies on rich commonsense knowledge, reasoning, and discourse understanding. To somewhat limit the scope of our task, we focus on applying formalizations of incongruity to a subset of linguistic humor: puns. Writer and philosopher Henri Bergson defined a pun as "a sentence or utterance in which two ideas are expressed, and we are confronted with only one series of words." This definition highlights the fact that one sentence must evoke two different interpretations in order to be a pun, which aligns with the concept of incongruity as a requisite of humor.

We develop our model on homophone puns—puns that contain homophone words—because the space of possible interpretations of a homophone pun is relatively constrained and well-defined. An example helps to illustrate:

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

This sentence allows for two interpretations:

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

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

If the comprehender interprets the word "hare" as itself, he will arrive at interpretation (a); if he interprets the word as its homophone "hair," he will arrive at interpretation (b). In other words, the sentence-level differences between interpretations (a) and (b) can be approximated by the two interpretations of the observed word "hare." In general, distinct interpretations of a homophone pun hinges on one phonetically ambiguous word, allowing the two lexical forms of the homophone word to stand in for competing interpretations of the entire sentence.

Critically, even though the example we gave was a written pun and the reader sees the word "hare" explicitly on the page, the "hair" interpretation is still present and even salient in the context of the sentence. Noisy channel models of sentence processing posit that language comprehension is a rational process that incorporates uncertainty about the input at the word level to arrive at sentence-level interpretations that are globally coherent (Levy, 2008; Levy, Bicknell, Slattery, & Rayner, 2009). This process allows the comprehender to consider multiple word-level interpretations ("viewpoints") to arrive at more than one interpretation of a sentence, each coherent but potentially incongruous with each other. The notion of incongruity thus fits naturally into a noisy channel model of sentence comprehension.

Our purposes for developing a formal model of linguistic humor are two-fold. First, we wish to formalize the concept of incongruity and test assumptions adopted by leading theories in humor research. Secondly, we aim to show that a noisy channel of language processing allows for flexible context selection and sentence comprehension that gives rise to sophisticated linguistic and social meaning such as humor.

Model

We propose a computational model for general sentence comprehension that incorporates a noisy channel approach and semantic information to arrive at optimal global interpretations of a sentence. Based on the incongruity theory of humor,

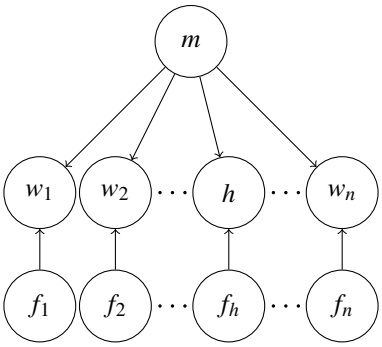


Figure 1: Generative model of a sentence

we derive measures of incongruity from the model to predict humans' judgements of funniness. Given a sentence that contains a phonetically ambiguous target word, a comprehender maintains uncertainty over the observed input and uses contextual cues to infer the latent interpretation of the target word. If the resulting interpretation is unambiguous, then no incongruity exists and the sentence is unlikely to be funny. If the resulting interpretation is ambiguous, and furthermore if each interpretation is supported by an aspect or "viewpoint" of the sentence, then the sentence is likely to evoke a sense of incongruity and humor.

Previous research suggests that semantic priming plays an important role in lexical disambiguation during sentence processing. In particular, efficient processing of sentences that contain lexical ambiguity involves leveraging semantic association with neighboring words (Seidenberg, Tanenhaus, Leiman, & Bienkowski, 1982; Simpson, 1981; Burke & Yee, 1984). We begin with a bag-of-words model of language and make the simplifying assumption that comprehenders use only semantic association information to compute the probabilities of a candidate interpretation.

Suppose a sentence is composed of a vector of content words $\vec{w} = \{w_1, \ldots, w_n\}$ and a phonetically ambiguous word h. We make the simplifying assumption that candidate interpretations of h are constrained by phonetic similarity. As a result, only the two homophone interpretations of h are salient and probable: m_1 and m_2. For example, in the magician pun described above, h is the phonetically ambiguous target word "hare," and m_1 and m_2 are the candidate interpretations *hare* and *hair*, respectively.

We propose a generative model of a sentence in which the latent meaning variable m and a latent indicator variable $\vec{f}$ are responsible for generating the observed words (see Figure 1). Although the target word h is special in the sense that it constrains the space of interpretations to its homophone meanings m_1 and m_2, in the model we treat h as a contextual cue similar to other words in its relationship to the latent meaning variable m and indicator variable f_h. For simplicity of notation, in the following derivations, $\vec{w}$ includes the target word h, and $\vec{f}$ includes the indicator variable for h, f_h. $\vec{f}$ is a vector of indicator values that determines whether each word in a sentence is in semantic focus. When $f_i = 1$, w_i is in semantic focus and is generated due to semantic relevance to m. Otherwise if $f_i = 0$, w_i is simply drawn from a default unigram distribution. Similar approaches have been used in generative models of language to account for words that serve purposes in a sentence beyond providing semantic information, such as topic models that incorporate syntax (Griffiths, Steyvers, Blei, & Tenenbaum, 2005). For our purposes, the indicator variable allows the model to consider subsets of the sentence as viewpoints that may differentially support the two candidate interpretations m_1 and m_2.

Using the generative model of a sentence, we can infer the joint probability distribution of the latent variables m and $\vec{f}$ given the observed words $\vec{w}$ and h. This distribution can be factorized into:

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

We derive two measures separately from the factors $P(m|\vec{w})$ and $P(\vec{f}|m, \vec{w})$, which we will call *ambiguity* and *distinctiveness*. Ambiguity means the presence of two similarly likely interpretations, which is required for incongruity because a single unambiguous interpretation cannot be incongruous with itself. Distinctiveness means the two interpretations are supported by different viewpoints of the sentence, which we approximate as words that are in semantic focus with the two values of m. Together, these two measures represent our formalization of incongruity.

Ambiguity Let M denote the distribution $P(m|\vec{w})$, a binomial distribution over the two meaning values m_1 and m_2 given the observed words. If the entropy of this distribution $H(M)$ is low, this means that the probability mass is concentrated on only one meaning, and the alternative meaning is unlikely given the observed words. If $H(M)$ is high, this means that the probability mass is more evenly distributed among m_1 and m_1, and the two interpretations are similarly likely given the contexts. $H(M)$ is thus a natural measure of the degree of ambiguity present in a sentence.

We compute $P(m|\vec{w})$ as follows:

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

From Bayes' theorem, this is proportional to the following:

$$\sum_{\vec{f}} P(\vec{w}|m, \vec{f})P(m)P(\vec{f}) = \sum_{\vec{f}} \left(P(m)P(\vec{f}) \prod_i P(w_i|m, f_i) \right)$$

From the generative model,

$$P(w_i|m, f_i) = \begin{cases} P(w_i), & \text{if } f = 0 \\ P(w_i|m), & \text{if } f = 1 \end{cases}$$

We then compute the entropy of $P(m|w)$ as a measure of ambiguity.

Distinctiveness Let F_1 denote the distribution $P(f|m_1,\vec{w})$ and F_2 denote the distribution $P(f|m_2,\vec{w})$. F_1 and F_2 represent the distributions over semantic focus sets given the observed words and m_1 and m_2, respectively. We use a symmetrised KullbackLeibler divergence score $D_{KL}(F_1||F_2)+D_{KL}(F_2||F_1)$ to measure the distance between the two distributions. This score measures how "distinct" the semantic focus sets are given m_1 and m_2. A low KL score would indicate that m_1 and m_2 are likely to be in semantic focus with similar subsets of the sentence; a high KL score would indicate that m_1 and m_2 are likely to be in semantic focus with different subsets of the sentence.

We compute the measure as follows:

$$P(\vec{f}|m,\vec{w}) \propto P(\vec{f},\vec{w}|m) = P(\vec{w}|m,\vec{f})P(\vec{f}|m)$$

From the generative model, since $\vec{f}$ and m are independent, $P(\vec{f}|m) = P(\vec{f})$. We assume a uniform probability over focus sets in our computations.

We have described a generative model of sentences containing phonetically ambiguous words that incorporates the idea of semantic focus sets. We then used the model to derive two measures that formalize incongruity. By combining a noisy channel model of language comprehension and standard information theoretic measures, our model presents a formalization of incongruity to be tested on humans' judgments of potentially humorous sentences.

Evaluation

We evaluate our model and measures on a set of 235 sentences, consisting of 65 puns, 40 "de-punned" control sentences that are matched with a subset of the puns, and 130 control non-pun sentences that match the puns in containing the same phonetically ambiguous words.

Materials

We selected 40 pun sentences from a large collection of puns on a website called Pun of the Day, which contains over one thousand puns. Puns were selected such that the ambiguous item is a single phonetically ambiguous word, and no two puns in the collection have the same ambiguous item. To obtain more homophone pun items, a research assistant generated an addition 25 pun sentences based on a separate list of homophone words.

We constructed 40 sentences to be minimally different from the pun sentences that we collected from "Pun of the Day," which we will call de-punned sentences. A second research assistant who was blind to the hypothesis was asked to replace one word in each of the pun sentences (without changing the homophone word itself) so that the sentence is still grammatical but is no longer a pun. This resulted in sentences that were only one word different from the pun sentences.

The 130 non-pun sentences were chosen to match each pun sentence on its ambiguous word as well as the alternative homophone. The sentences were taken from an online version of Heinle's Newbury House Dictionary of American English (http://nhd.heinle.com/). We selected sample sentences included in the definition of the homophone word. This design ensured that puns and non-pun sentences contain the same phonetically ambiguous words.

Figure 2 shows example sentences from each category.

Type	Example
Pun	The magician got so mad he pulled his hare out.
De-pun	The professor got so mad he pulled his hare out.
Non-pun	The hare ran rapidly across the field.
Non-pun	Some people have lots of hair on their heads.

Figure 2: Example sentences from each category

Human ratings of semantic relatedness

As described in the model section, in order to compute our measures, we need the prior probabilities of the interpretations $P(m)$, the prior probabilities of the words $P(w)$, and the conditional probabilities of each word in the sentence given an interpretation $P(w|m)$. We computed $P(w)$ and $P(m)$ directly from the Google unigram corpus. However, $P(w|m)$ is difficult to obtain through corpora due to data sparsity. Given these constraints, we approximate $P(w|m)$ using empirical measures of the semantic relatedness between w and m, denoted as $R(c,m)$. We use $R(c,m)$ as a proxy for point wise mutual information between c and m, defined as follows:

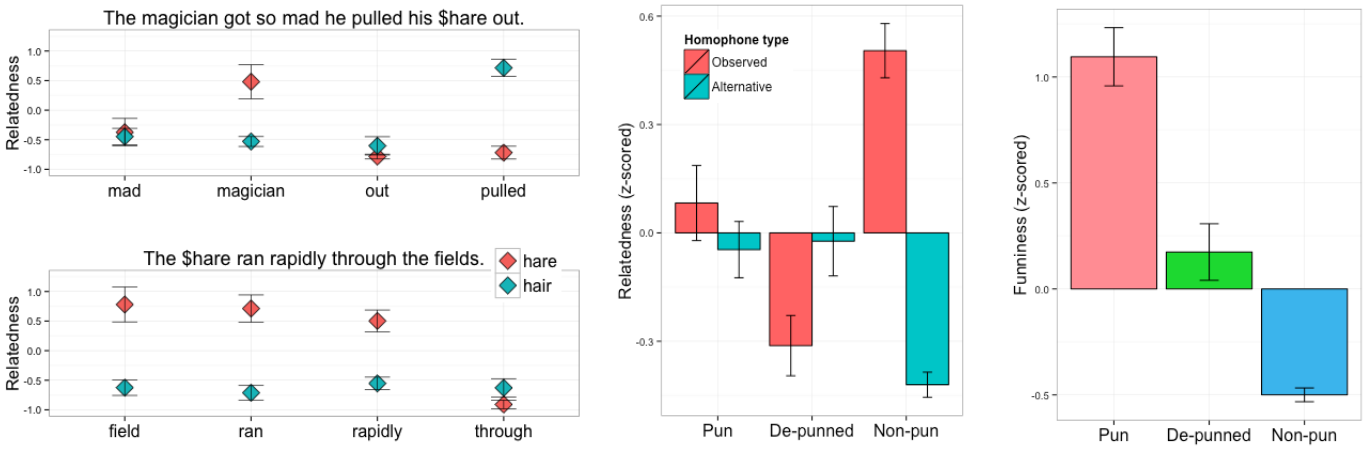
$$R(w,m) = \log\frac{P(w,m)}{P(w)P(m)} = \log\frac{P(w|m)}{P(w)} = \log P(w|m) - \log P(w)$$

With the proper substitutions and transformations,

$$P(w|m) = e^{R(w,m)}P(w)$$

To obtain $R(w,m)$ for each of the words w in the stimuli sentences, we recruited 200 subjects on Amazon's Mechanical Turk to rate distinct word pairs on their semantic relatedness. Function words were removed from the sentences, and the remaining words were paired with each of the interpretations of the homophone sequence (e.g., "magician" and "hare" is a legitimate word pair, as well as "magician" and "hair"). This resulted in 1460 distinct word pairs. Each subject saw 146 pairs of words in random order and were asked to rate how related each word pair is from 1 to 10. The average split-half correlation of the relatedness ratings was 0.916.

Figure 3 (a) shows the relatedness of each content word with the two homophone interpretations for two example sentences. In the top sentence, which is a pun, the word "magician" is rated as significantly more related to "hare" than it is to "hair", while the word "pulled" is rated as significantly more related to "hair" than it is to "hare." In the bottom sentence, which is a non-pun, all words except the neutral word "through" are more related to the word "hare" than to "hair."



(a) Relatedness of word pairs in example pun and non-pun

(b) Average relatedness

(c) Average funniness

Figure 3: Analyses of human ratings

Figure 3 (b) shows the average relatedness ratings of words and the two homophone interpretations across the three types of sentences. In pun sentences, the average relatedness of words to the two homophone interpretations are roughly equivalent. In the de-punned sentences, the average relatedness of words to the observed homophone is significantly lower than to the alternative homophone. In the non-pun sentences,the average relatedness of words to the observed homophone is significantly higher than to the alternative homophone. These analyses suggest that relatedness ratings for the two candidate meanings capture the presence of alternative interpretations in puns, de-punned sentences, and non-puns. It further supports the intuition that ambiguity of meaning and the distinctiveness of supporting context words can help distinguish among the three types of sentences.

Human Ratings of Funniness

We obtained funniness ratings of the 235 sentences from 100 subjects on Amazon's Mechanical Turk. Each subject read roughly 60 sentences in random order, counterbalanced for the sentence types, and rated each sentence on funniness and correctness. The average split-half correlation of the funniness ratings was 0.83. Figure 3 (c) shows the average funniness ratings of puns, de-punned, and non-pun sentences. Pun sentences are rated as significantly funnier than de-punned sentences, and de-punned sentences are rated as significantly funnier than non-pun sentences ($F(2,232) = 415.3, p < 0.0001$). Figure 3 (b) and Figure 3 (c) together suggest that the balance of relatedness between the two interpretations is a predictor of funniness.

Results

We computed an ambiguity and distinctiveness value for each of the 235 sentences following the derivations described in the model section and using the relatedness measures described above.

As predicted, ambiguity differs significantly across sentence types ($F(2,232) = 11.94, p < 0.0001$) and correlates significantly with human ratings of funniness across the 235 sentences ($r = 0.27, p < 0.0001$). Furthermore, distinctiveness of semantic focus sets differs significantly across sentence types as well ($F(2,232) = 5.76, p < 0.005$) and correlates significantly with human ratings of funniness, although less strongly ($r = 0.21, p < 0.005$).

Using both ambiguity and distinctiveness as the two dimensions that formalize incongruity, we can distinguish among puns, non-puns, and de-punned sentences. We built a linear regression model using ambiguity and distinctiveness as predictors of human ratings of funniness. Results showed that both ambiguity and distinctiveness are significant predictors of funniness and together capture a significant amount of the variance in funniness ratings ($F(2,232) = 19.6, R^2 = 0.137, p < 0.001$) (see Table 1).

	Estimate	Std. Error	p value
Intercept	−0.687	0.140	< 0.0001
Ambiguity	1.011	0.193	< 0.0001
Distinctiveness	0.237	0.054	< 0.0001

Table 1: Regression model for funniness ratings

Semantic focus sets Beyond predicting the funniness of a sentence, our model can also tell us which particular features of a pun make it amusing. By finding the semantic focus sets $\vec{f}$ that are most likely given the observed words and latent meaning variable m, we can identify words in a sentence that are likely to be in semantic focus given each of the two values of m. Table 2 shows the most likely semantic focus sets given each meaning, the ambiguity and distinctiveness scores, and funniness ratings for two sets of sentences. Although each sentence

$m1$	$m2$	Type	Sentence and Semantic Focus Sets	Amb.	Disj.	Funniness
hare	hair	Pun	The **magician** got so mad he **pulled** his **hare** out.	0.378	2.291	1.714
		De-pun	The professor got so mad he **pulled** his **hare** out.	0.048	1.832	−0.103
		Non-pun	The **hare ran rapidly** through the **fields**.	0.447	1.677	−0.400
		Non-pun	Most **people** have **lots** of **hair** on their **heads**.	0.0004	2.807	−0.343
tiers	tears	Pun	It was an **emotional wedding**. Even the **cake** was in **tiers.**	0.612	2.311	1.541
		De-pun	It was an **emotional wedding**. Even the **mother-in-law** was in **tiers**.	0.189	1.802	0.057
		Non-pun	**Boxes** are **stacked** in **tiers** in the warehouse.	0.194	2.089	−0.560
		Non-pun	**Tears** ran down her **cheeks** as she watched a **sad movie**.	0.0003	3.283	−0.569

Table 2: Semantic focus sets of example sentences. Words in red are in semantic focus with m_1; green with m_2; navy blue with both.

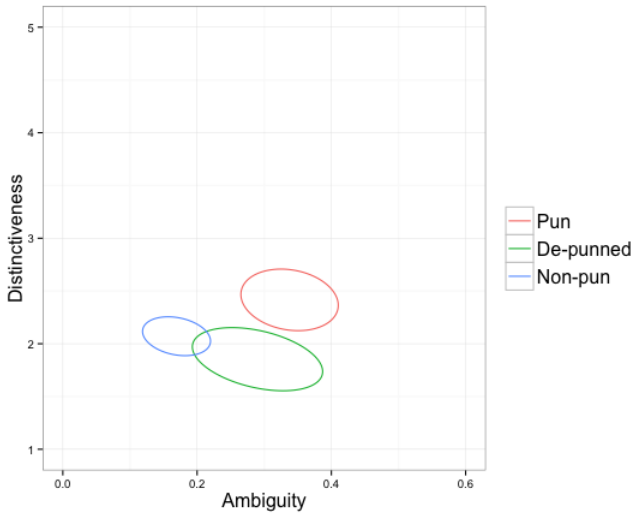


Figure 4: Standard error ellipses of ambiguity and disjointedness across sentence types

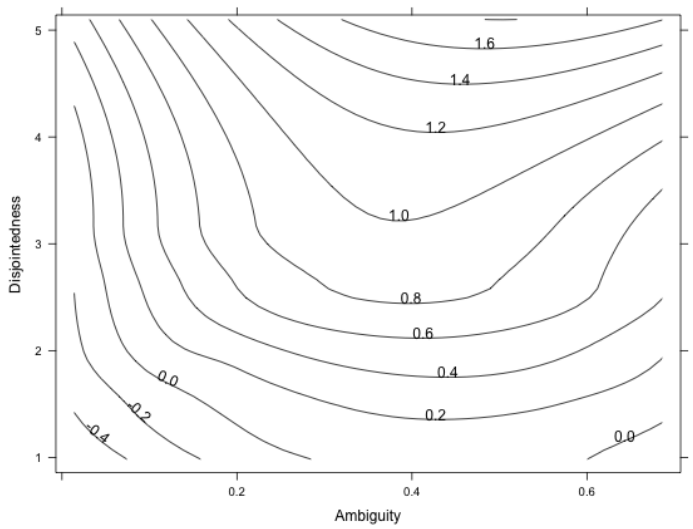


Figure 5: Ambiguity and disjointedness with funniness contours

Discussion

Mostly incomplete

Given the prevalence of humor in human communication, researchers in artificial intelligence have argued that computers should be able to generate and detect humor in order to interact with humans more naturally and effectively (Mihalcea & Strapparava, 2006). As a result, computational humor has made important progress and also received attention from popular press in the last decade (insert NYT citation here). However, most of the work in computational humor has focused either on utilizing joke-specific templates and schemata (Binsted, 1996; Kiddon & Brun, 2011), or identifying linguistic features such as slang and alliteration that strongly predict humorous intent (Mihalcea & Strapparava, 2006; Reyes, Buscaldi, & Rosso, 2010). The former type of studies is restricted to identifying jokes with a very specific format and structure, while the latter type falls short of testing or building upon deeper and more general theories of humor involving the management of incongruity.

Our work moves beyond these two types of approaches and directly utilizes incongruence to identify humorous texts. Given that humor theorists view incongruity as an essential component of jokes, we examined whether human judgments of funniness can be predicted by the presence of incongruous interpretations of the same input. In particular, we developed a formal model of linguistic humor that fits naturally into the framework of normal language processing. We propose that a noisy channel approach to language processing allows the comprehender to consider alternative viewpoints and interpretations of the same linguistic input and can account for the possibility of incongruity.

Incomplete Describe its advantages over previous work on computational humor. Describe potential applications and generalizations to other forms of humor.

Acknowledgments

References

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

Burke, D., & Yee, P. (1984). Semantic priming during sentence processing by young and older adults. *Developmental Psychology*, *20*(5), 903.

Duncan, W., Smeltzer, L., & Leap, T. (1990). Humor and work: Applications of joking behavior to management. *Journal of Management*, *16*(2), 255–278.

Forabosco, G. (1992). Cognitive aspects of the humor process: The concept of incongruity. *Humor: International Journal of Humor Research; Humor: International Journal of Humor Research*.

Frank, M., & McGhee, P. (1989). *Humor and children's development: A guide to practical applications*. Routledge.

Gelkopf, M., & Kreitler, S. (1996). Is humor only fun, an alternative cure or magic? the cognitive therapeutic potential of humor. *Journal of Cognitive Psychotherapy*, *10*(4), 235–254.

Griffiths, T., Steyvers, M., Blei, D., & Tenenbaum, J. (2005). Integrating topics and syntax. *Advances in neural information processing systems*, *17*, 537–544.

Hurley, M., Dennett, D., & Adams, R. (2011). *Inside jokes: Using humor to reverse-engineer the mind*. Mit Pr.

Johnson, P. (2002). The use of humor and its influences on spirituality and coping in breast cancer survivors. In *Oncology nursing forum* (Vol. 29, pp. 691–695).

Kiddon, C., & Brun, Y. (2011). Thats what she said: double entendre identification. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 89–94).

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).

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.

Martin, R. (2007). *The psychology of humor: An integrative approach*. Academic Press.

McGhee, P. (1979). *Humor, its origin and development*. WH Freeman San Francisco.

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

Reyes, A., Buscaldi, D., & Rosso, P. (2010). The impact of semantic and morphosyntactic ambiguity on automatic humour recognition. In H. Horacek, E. Mtais, R. Muoz, & M. Wolska (Eds.), *Natural language processing and information systems* (Vol. 5723, p. 130-141). Springer Berlin / Heidelberg.

Seidenberg, M., Tanenhaus, M., Leiman, J., & Bienkowski, M. (1982). Automatic access of the meanings of ambiguous words in context: Some limitations of knowledge-based processing. *Cognitive psychology*, *14*(4), 489–537.

Simpson, G. (1981). Meaning dominance and semantic context in the processing of lexical ambiguity. *Journal of Verbal Learning and Verbal Behavior*, *20*(1), 120–136.

Stewart, S., Stinnett, H., & Rosenfeld, L. (2000). Sex differences in desired characteristics of short-term and long-term relationship partners. *Journal of Social and Personal Relationships*, *17*(6), 843–853.

Veale, T. (2004). Incongruity in humor: Root cause or epiphenomenon? *Humor-International Journal of Humor Research*, *17*(4), 419–428.

Veatch, T. (1998). A theory of humor. *Humor*, *11*, 161–215.

Ziv, A., & Gadish, O. (1989). Humor and marital satisfaction. *The journal of social psychology*, *129*(6), 759–768.