Humor in Cartoon Captions

Pedram Razavi (prazavi@stanford.edu)

Abstract

Why do we laugh at some seemingly random occurrences and not the others? One of the reigning theories of humor is incongruity theory. Although, different variations of incongruity theory might dictate the necessary conditions that lead to the sensation of mirth, but they all stop short of offering the sufficient conditions. In this paper, I study some of the entries to *The New Yorker* cartoon caption contest to establish what might make a caption funnier than its alternatives. Using topic modeling algorithms such as Latent Dirichlet Allocation, I compare captions of a cartoon to the literal descriptions of the same drawing to establish whether I can reaffirm incongruity factors such as ambiguity of meaning and distinctiveness of viewpoint (Kao, Levy, & Goodman, 2013, To appear) in the domain of cartoon captions and humorous drawings. At this point, I could not find a significant correlation between humor ratings of a cartoon and the degree of abstractness of a caption. However, the present work stands as an initial step in the rigorous study of humor in the natural yet constrained world of cartoon captions.

Keywords: Humor; language understanding; probabilistic models; topic models; Latent Dirichlet Allocation

Introduction

What makes something funny? Studying humor is a difficult task yet it is crucial because not only it has numerous applications in natural language processing, but humor studies can also give us glimpses into creative aspects of human cognition. For starters, to see why studying humor is difficult, we should note that the concept of humor is elusive and difficult to define. While we have an intuitive sense of what humor might be; researchers have struggled with a definition that can capture different facets in which people use or consume humor in their daily lives (Hurley, Dennett, & Adams, 2011). We use the word "humor" to describe so many seemingly different activities: in embarrassment, laughing at a joke, when getting tickled by someone, in relief of when something unwanted does not occur, and so on.

In this paper, I take a data-driven approach to look at this fundamental and ubiquitous phenomenon. Specifically, using a large dataset of *The New Yorker* cartoon caption contests (Figure 1), I am interested in studying the defining characteristics of humor and the underlying patterns in these captions. To achieve this goal, I use Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003), a widely used topic modeling probabilistic algorithm, to uncover the hidden thematic structures of these captions. This in turn would enable us to naturally classify captions in smaller subsets of equivalent topics and study them in more constrained settings. These categorizations could shed light into different ways that people capture and generate humorous contents. Moreover, by comparing the topics distributions of captions with the literal descriptions of the same drawing, a measure of abstractness of a caption can be obtained. This bottom up study of humor compliments the evolutionary top down accounts of why humor exists and psychological models of what makes something funny.

"We have to move out—I just sold a painting."

Figure 1: *The New Yorker* cartoon caption contest number 462 and the winning caption. A weekly contest established in 2005, where readers of the magazine submit captions for a printed cartoon in the back of each issue. The magazine receives around 6,000 caption submissions each week.

Background

Throughout the study of humor, there has been the prevalent assumption that there must be some essential underlying essence shared by activities that lead to laughter. Once a sufficient underlying model is found, we could explain the inner workings of humor and how it is generated, the reasoning goes. Hurley et al. even go further and claim that: " [Our book is] about the epistemic predicament of agents in the world and a class of models of cognition that can successfully deal with that predicament. It argues that emotions govern all our cognitive abilities, large and small, and that humor is thus a rich source of insight into the delicate machinery of our minds (Hurley et al., 2011)." This shows the great importance of studying the intricacies of humor.

One of the many humor theories is put forth by Freud (Freud, 1989). He theorizes that humans have numerous internal censors (shaped by social constructs) for the most repressed emotions such as sexual desire and hostility. When one generates or consumes a joke, he or she bypasses the internal censors. Consequently, the psychic energy that is usually used for repression becomes superfluous, and is released in the form of laughter. Later, Minsky extends this theory and defines the role of these censors more clearly and illustrates the dynamics of learning that goes on in such censors (Minsky, 1984). Another more high-level and prevalent explanation of why humor leads to the sensation of mirth is the

incongruity theory. Simply put, proponents of this theory assert that a humorous anecdote resolves the incongruity of a context. This agrees with our intuitive "punchline" nature of humor. However, the main problem associated with most of the variations of such theory is that there is no agreed upon definition for the umbrella term "incongruity", and most definitions are hard to operationalize. Works by Kao et al. (Kao et al., 2013, To appear) have elucidated some aspects of the notion of incongruity through rigorous formulation. More specifically, Kao et al. study how incongruity can be concretely and computationally defined by combining a noisy channel model of language comprehension and standard information theoretic measures. Using these models they establish two dimensions of incongruity: ambiguity of meaning and distinctiveness of viewpoints. However, the domain that they tested and applied this model to is a constrained domain of textual puns.

To explain why humor came into being in the first place, there have been multiple theories for the source of humor, among them the evolutionary understanding of humor (Hurley et al., 2011; Minsky, 1984). Most of such theories equate jokes with problem solving. Consequently, when someone "gets" a joke, the feeling it evokes is similar to the feeling of solving a riddle. This theory is compatible with variations of incongruity theory because in a sense in both cases there is some notion of puzzle or incongruity that needs to be resolved, however the evoked emotions differ: we laugh in reaction to a joke but we feel satisfied by solving a riddle. Nonetheless, in both situations a buildup tension is resolved or relieved. In their book Hurley et al. claim that humor serves as an "error correction" mechanism (Hurley et al., 2011). Using which, people debug certain mistakes that enter the conscious mind when they should not. In other words, our brain with its limited resources needs a self-correcting mechanism that filters-out information that hinders its function and/or leads the individual to commit potentially costly errors from erroneous conclusions. Humor filters resolve the tension when such errors or incongruences arise and the reward system that motivates us to conduct debugging is manifested by laughter and the sensation of mirth. Along the same lines, Minsky formulates this issue using the Freudian censors analogy. He asserts that each person is constantly extending a private collection of "cognitive censors" to suppress past mistakes, because he or she does not have a systematic way of avoiding all inconsistencies of commonsense logic a priori (Minsky, 1984). A joke is never as funny the second time it is heard, this theory explains.

In this paper, I am more concerned with an operational way of assessing humor, specifically in the case of cartoon captions. The aim is to concretely examine whether similar factors as devised by Kao et al. can be extended to the domain of humor in cartoon captions and humorous drawings. More concretely, let's call the KL-Divergence between topics distribution of the literal description of a cartoon drawing and topics distribution of a caption the "caption abstractness" measure. This measure is related both to the ambiguity of meaning and distinctiveness of perspective. For instance, a more abstract caption does not directly use the words that describe the drawing, therefore it is considered as ambiguous and distinctive and hence funny, yet at the same time a caption that is perceived to be unrelated to a cartoon is unlikely to be considered as funny. The question that I set forth is that, do abstract captions tend to be funnier on average or vice versa? And how much abstract should a cartoonist go?

Model

The main model used in this project is the generative Latent Dirichlet Allocation (Blei et al., 2003) which is widely used to extract topics from corpora of texts. To reduce the dimension of captions other methods such as k-means clustering of term frequency-inverse document frequency matrix was used, however, LDA proved more robust in this domain and the final results reported in this paper do not use this clustering method.

Latent Dirichlet Allocation

Latent Dirichlet Allocation (Blei et al., 2003; Blei, 2012) is a generative model of text. This model casts the problem of discovering themes in large document collections as a posterior inference problem. For instance, in the domain of cartoon captions, a salient topic for a subset of captions might be "copyright attorney" or "casual friday" for the cartoon illustrated in Figure 1. This classification lets us visualize the hidden thematic structure in large collections, and generalize new data to fit into that structure. To qualitatively evaluate the result of LDA, topic model visualization tools can be used such as the one developed by Chaney et al. (Chaney & Blei, 2012).

LDA relies on the assumption that a small number of latent topics suffice to effectively represent a large corpus. The main intuitions behind LDA is as following: we assume that some number of "topics," exist which concisely describe our collection of documents. These topics are probability distributions over a fixed vocabulary. Each document is generated as following: first, chose a distribution over the topics, then, for each word, choose a topic assignment and choose the word from the corresponding topic. More specifically:

1: **for** document d_d in corpus D **do**
2: Choose $\theta_d \sim \text{Dirichlet}(\alpha)$
3: **for** position w in d_d **do**
4: Choose a topic $z_w \sim \text{Multinomial}(\theta_d)$
5: Choose a word w_w from $p(w_w|z_w, \beta)$, a multinomial distribution over words conditioned on the topic and the prior β.
6: **end for**
7: **end for**

Tokenization

After experimenting with different caption tokenization techniques, I used the following method:

- Initial tokenization using the NLTK tokenizer (Bird, Klein, & Loper, 2009).
- Remove punctuations and common English stop words.
- Keep nouns, verbs, adjectives, and adverbs after using NLTK part-of-speech tagger (Bird et al., 2009).
- Use WordNet (Miller, 1995) to lemmatize and find synonyms set of each token.

We should note that there are some overlaps between these steps but since runtime was not a bottleneck, no attempt was made to optimize the tokenization pipeline.

Number of Topics

LDA assumes that the number of topics is known a priori. Although, estimating the number of topics might be easy in some cases, there are no general procedures for finding the most expressive number of latent topics for a given corpus. To find the number of topics that best describe our corpus of captions, I used the observation and presented heuristic by Arun et al. (Arun, Suresh, Madhavan, & Murthy, 2010). They approach LDA as a matrix factorization method and present a measure based on the symmetric KL-Divergence of salient distributions that are derived from LDA matrix factors. Although this method is a heuristic and not a theoretical result, they have observed that in numerous experimental studies their presented measure is higher for non-optimal number of topics. This observation translates to this finding: the appropriate number of topics can be loosely evaluated by finding a dip in the graph of their presented symmetric KL-Divergence measure (Figure 5).

Evaluation

The New Yorker magazine features a weekly cartoon caption contest, in which readers can submit a short caption (less than 250 words) for a cartoon printed in the last page of each issue. This contest has been running since 2005 (more than 470 contests so far) and attracts between 5,000 to 6,000 submissions per week for each cartoon. Although these captions constitute a limited subset of what seems to be the infinite ocean of humor, the dataset is rich, yet constrained enough to be studied computationally.

Human Ratings of Funniness

I selected 5 cartoons from a collection of 470 past *The New Yorker* cartoon caption contests. For each contest, I randomly selected 17 caption submissions and added the 3 finalist captions, curated by the cartoon editor of the magazine, to this pool. Several entries that were incomplete were discarded and replaced. Using Amazon's Mechanical Turk, I obtained funniness ratings of the 20 captions from 20 participants. After seeing the related cartoon, the subjects read 20 captions in a random order and were asked to rate each one from 0 to 100 using a slider with increments of 10. In the end, they were asked to write their own suggestion for a caption, however, I did not use these suggestions for this project.

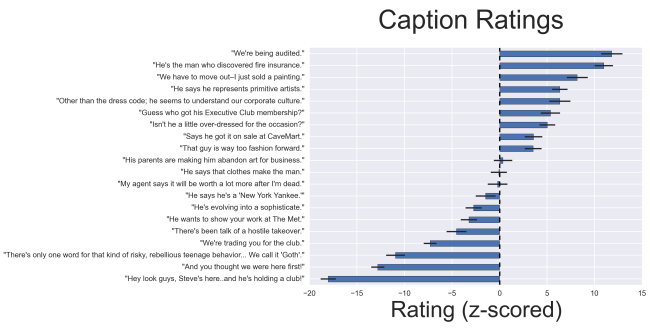

Figure 2: Normalized humor ratings for contest number 462. Twenty subjects were asked to rate 20 captions. Average split-half correlation of 0.51 with two-sided p-value of 0.04.

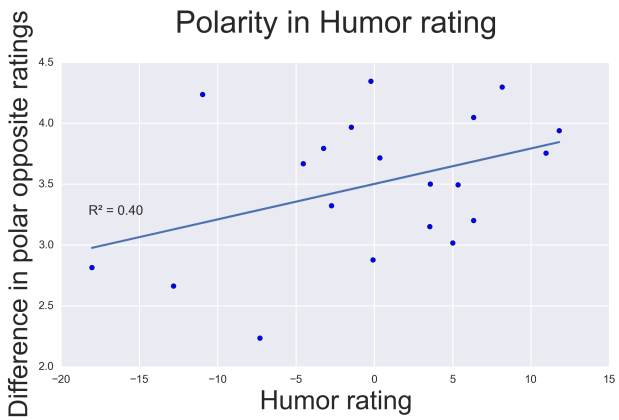

Figure 3:

The average split-half correlation of the humor ratings for the five different contests were around 0.6. This means that although humor ratings are considered to be subjective per individual, in a population, the aggregate caption rating converges quickly even for a rather small sample size. Figure 2 illustrates the normalized humor ratings plot of one of the five contests that was chosen in this experiment. A similar pattern is apparent in the ratings plot of the other four contests.

Contest Number	Split-half Correlation	P-value
453	0.48	0.063
454	0.61	0.013
455	0.59	0.018
462	0.51	0.047
463	0.67	0.004

By examining the difference between most and least favored rankings for each caption after normalization of ratings, one can see that there exist a correlation between how polar the ratings for a caption are and its aggregate humor ratings. In sum, funnier caption seem to be more polar on average (Figure 3).

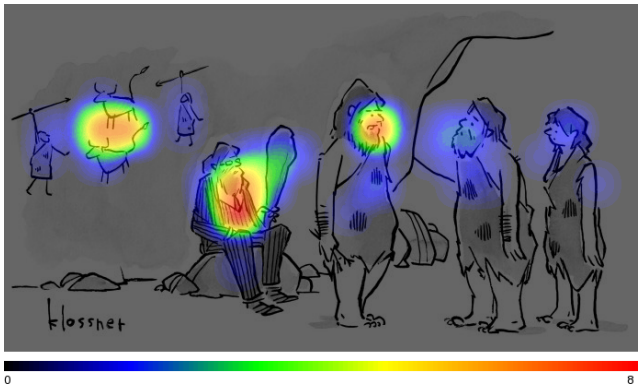

Figure 4: Self-reported clicks heat map of the most salient features of a drawing. Twenty participants were asked to click on different parts of the image that attracted their attention the most.

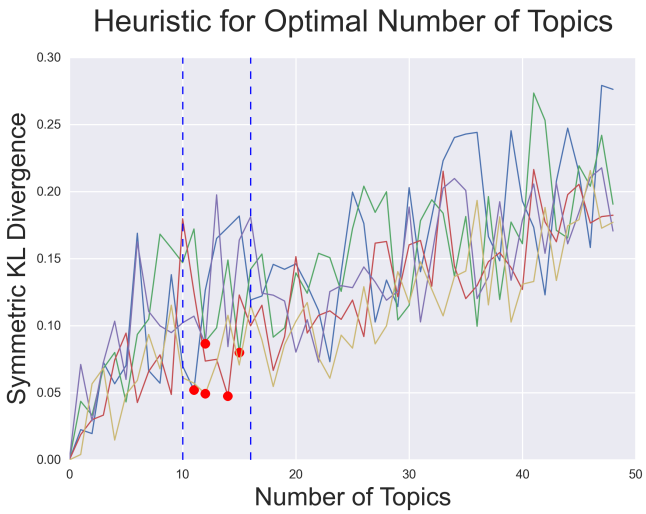

Figure 5: Number of LDA Topics vs. Symmetric KL Divergence for five contests. For each contest, the dip in divergence happens when the number of topics is in the range of 10-15. Based on divergence measure presented by Arun et al. (Arun et al., 2010), these dips correlate well with the optimal number of topics in the corpus.

Literal Description of a Cartoon

In another independent experiment, again for the same five selected cartoons, 20 subjects on Amazon Mechanical Turk were asked to write what they saw in each drawing. In this experiment the participants were not shown any of the captions. The aim of this experiment was to gather a literal description of the most salient features of a cartoon drawing. Two examples of participants' responses to cartoon number 462 are as following:

- "I see cavemen lined up or grouped talking. A Freud looking figure in a suit sitting on a rock and cave painting on the wall of the cave."
- "I see several cavemen standing around in a cave, dressed in loose furs, apparently talking to each other. There is one cavewoman among them. Behind them is another caveman sitting on a rock. He is dressed in a business suit and has glasses, and he is also holding a club. Behind him is the cave wall, marked by drawings of cavemen hunting horned beasts. In the lower left corner is the name of the artist, Klossnet."

Participants were also asked to self-report the areas of a drawing that attracted their attention the most by directly clicking on the drawing. The resulting clicks heat map for one of the five contests is depicted in Figure 4. The data from this part of the experiment is not directly used in this paper, besides qualitative comparison of the clicks heat maps with the aggregate textual participants' responses.

Results

Number of Topics

Using Arun et al. (Arun et al., 2010) heuristic for finding optimal number of topics, I found that 10 to 15 topics are sufficient to express the captions corpora for five different contests. Figure 5 shows the result of running the divergence measure for the five contests. The evaluated small number of topics between 10 to 15 was also confirmed by manually looking at the output LDA topics. The limited number of topics that contestants use to generate their captions from is surprising. Firstly, it reestablishes the notion that subtle differences in captions with the same topic can make a considerable difference in the funniness of a caption. Secondly, if we assume that there are two to four salient features in each cartoon (as depicted in the clicks heat map in Figure 4), there should be on average three to eight ideas or topics for each salient feature in a cartoon that contestants generate their captions from. Therefore, qualitatively, contestants seem to aim for captions that are directly relevant to some aspects of the cartoon drawing. This is not surprising, but at the same time it brings out the fine boundary between an incongruent yet nonsensical caption and an incongruent and funny one.

Topics Distribution

I used the Gensim LDA library (Řehůřek, Sojka, et al., 2010) on each of the five corpora of captions using the suggested number of topics discussed previously. Although the dataset for each contest is very sparse, after inspection of the resulting topics, it seemed that LDA could extract different topics such as "art critic", "casual Friday", "intellectual property", "clubbing", "security", "tax audit", "fire insurance", and so on for the cartoon depicted in Figure 1. Using the trained LDA model, topics distribution for the 20 rated captions was evaluated as illustrated in Figure 6. Captions are usually assigned to one or two topics. Although, this might be the result of extraction of relevant topics by LDA, it is also partially caused by the short length of some of the captions. If the number of tokens for a caption are few, a word that might

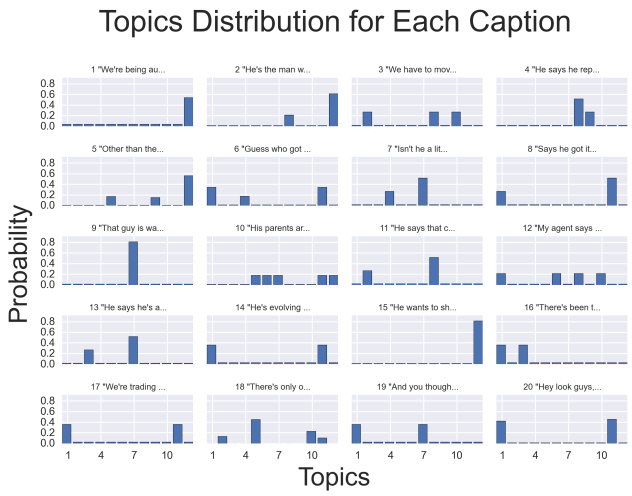

Figure 6: Topics distributions of 20 rated captions sorted based on their humor ratings. The probability of each caption belonging to a specific topic usually peaks around one or two topics.

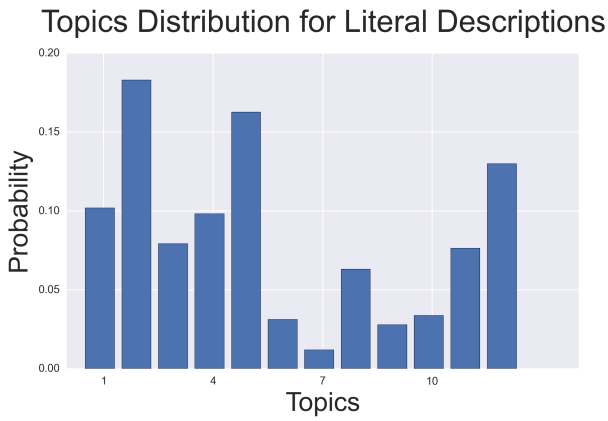

Figure 7: Topics distributions for aggregate literal descriptions of a drawing. Peaks on the topics that mention 'painting', 'security guard', 'cavemen', etc.

not be completely relevant to the meaning of a caption can significantly skew its assignment probability to a topic.

To compare the topics distributions of the captions with their literal descriptions, I again used the trained LDA model to find the topics distribution of the aggregate literal descriptions of a drawing (Figure 7). The literal descriptions are longer than the captions and they also tend to be more comprehensive and include details about all the features of a drawing. Therefore as it can bee seen in Figure 7, the assignment probability is non-zero for most of the topics.

By taking the KL-Divergence between the topics distribution of literal descriptions of a drawing and the captions, we can heuristically find a measure of how abstract a caption is. The higher the divergence measure, the more abstract a caption is. I compared this abstractness measure with the humor ratings of captions to obtain Figure 8. It does not seem that there is a significant correlation between the abstractness of a caption and its funniness ratings. There are several reasons for why this might be the case, for instance: the data set is too sparse, captions are too short, tokenization pipeline is losing too much information or it is not filtering the noisy words that contribute adversely to the caption topic models, and so on. LDA also has its own limitation as we treat each sentence as bag-of-words, consequently this model fails to capture word plays and idiomatic expressions in a caption.

Discussion

In this paper I compared the literal topic models of cartoon to the topic models of a caption to devise an abstractness metric for each caption. I could not find any correlation between this abstractness metric and how funny a caption is perceived to be. This is mostly due to the sparsity of the dataset, short length of the captions, and limitations of LDA which treats a caption as a bag-of-words. However the former factors play a bigger role as the aggregate topics resulting from LDA were qualitatively relevant. I found that 10 to 15 topics tend to capture the content of captions well. In future, it would be worthwhile to annotate a larger subset of captions with more metadata such as salient topics, number of features in the drawing that it addresses directly, etc. Moreover, the ratings experiment can be replicated by recording the demographics of each participant to see how perceived funniness is dependent on demographical contexts. Using this potential extended survey metadata, we can directly reformulate the abstractness measure and try to more rigorously formalize the relationship of a cartoon with its caption, and evaluate the factors that contribute to making a caption funny.

Forabosco's conclusion to his paper on assessing current state of humor incongruity theory and his outlook for the field is as follows: "We can soundly state that the concept of incongruity is still a useful and fruitful construct for humor investigation... Much of the research on humor is projected towards the future, and relies upon rigorous methods of investigation, and on advancements of technology (Forabosco, 2008)." Like Forabosco, I am optimistic about the future of incongruity theory, especially with development of more rigorous models such as the one proposed by Kao et al. (Kao et al., 2013, To appear). I also believe that datasets such as the cartoon caption contest are rich, yet constrained enough to be approachable by extensions to our current computational models.

References

Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Advances in knowledge discovery and data mining* (pp. 391–402). Springer.

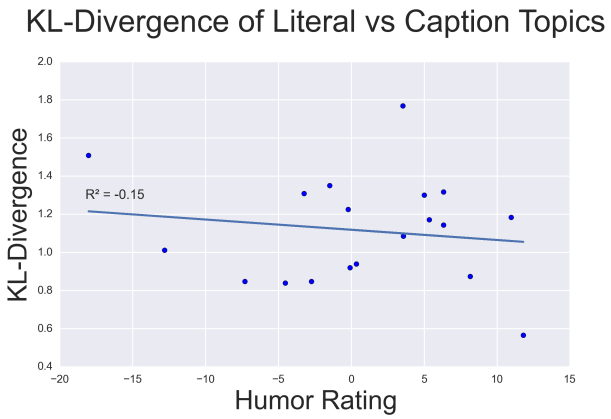

Figure 8: KL-Divergence between topics distribution of literal descriptions of a drawing and topics distribution of captions, plotted against humor ratings of the captions. An upward slope in this graph would show that captions that are perceived as funny tend to be less literal (or more abstract) on average. In contrast, a downward slope would show that more literal captions tend to be funnier.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python*. " O'Reilly Media, Inc.".

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993–1022.

Chaney, A. J.-B., & Blei, D. M. (2012). Visualizing topic models. In *Icwsm*.

Forabosco, G. (2008). Is the concept of incongruity still a useful construct for the advancement of humor research? *Lodz Papers in Pragmatics*, *4*(1), 45–62.

Freud, S. (1989). *Jokes and their relation to the unconscious* (No. 145). WW Norton & Company.

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

Kao, J. T., Levy, R., & Goodman, N. D. (2013). The funny thing about incongruity: A computational model of humor in puns. In *Proceedings of the 35th annual meeting of the cognitive science society*.

Kao, J. T., Levy, R., & Goodman, N. D. (To appear). A computational model of linguistic humor in puns. *Cognitive Science*.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

Minsky, M. (1984). *Jokes and the logic of the cognitive unconscious*. Springer.

Řehůřek, R., Sojka, P., et al. (2010). Software framework for topic modelling with large corpora.