Are AI-generated Jokes Truly Original? Charting the "Joke Space"

Ori Amir
Psychology Department
Pomona College
& HaHator LLC
Claremont, CA 91771 USA
ori.amir@pomona.edu

Abstract

This research addresses the challenge of evaluating the originality of jokes generated by Large Language Models (LLMs), which operate as opaque "black boxes" with non-transparent algorithms, potentially relying heavily on training datasets to produce humor. Although LLMs excel at rephrasing, raising concerns about plagiarism, existing studies often assume their outputs are original without verification. A novel framework is proposed to assess joke originality by characterizing the "Joke Space"—the set of all possible verbal jokes in English. Drawing on the General Theory of Verbal Humor (GTVH), this framework defines a joke's "essence" as its script opposition (SO) and logical mechanism (LM), critical for determining similarity and potential plagiarism. The size of the Joke Space is estimated using combinatorial and pragmatic approaches, suggesting that the vast number of possible jokes (at minimum 1 to 32 billion) exceeds the capacity of LLM training datasets, indicating potential for novel outputs. To ensure originality, the framework prompts LLMs with randomly sampled noun pairs to generate jokes, enabling comparative evaluation with human outputs. This approach provides a systematic method to verify the originality of AI-generated humor, offering practical recommendations for researchers.

Introduction

Large Language Models (LLMs) can generate jokes on demand, producing humor that rivals the quality of professional comedy writers, as evidenced by audience reactions in comedy club settings (Toplyn and Amir 2025). However, the originality of LLM-generated jokes remains uncertain (Jentzsch and Kersting 2023; Veale 2024). This gap is a significant limitation in studies of AI-generated humor. While traditional plagiarism detection can identify verbatim matches in a corpus or, when accessible, an LLM's training data, LLMs excel at rephrasing, prompting critics to label them as plagiarism machines (Chomsky, Roberts, and Watumull 2023) or "Stochastic Parrots" (Bender et al. 2021). These challenges highlight the need for a systematic approach to assess whether AI-generated jokes are truly novel, beyond mere rephrasing of existing content. This paper proposes a framework to evaluate the true originality of AI-generated jokes.

The Problem

Unlike rule-based joke generation algorithms, which follow transparent processes (Binsted and Ritchie 1997), LLM-based systems function as opaque "black boxes," obscuring the mechanisms behind their joke creation. Numerous studies investigating the comedic capabilities of LLMs often implicitly assume, without substantiation, that the generated jokes are original. However, when prompted to produce a joke without specific contextual guidance, OpenAI's ChatGPT-3.5 tends to either repeatedly generate identical jokes—suggesting potential hard-coding—or reproduce well-known jokes verbatim or with minimal alterations (Jentzsch and Kersting 2023; Veale 2024).

This limitation may be mitigated by prompting LLMs to craft jokes tailored to highly specific contexts (Toplyn 2023). Nevertheless, verifying originality remains challenging due to restricted access to training datasets and the possibility that jokes may be sufficiently modified to evade detection within those datasets. To address this, the likelihood of a joke's originality can be evaluated by characterizing the "Joke Space," defined as the set of all possible jokes in any form or language, though this study focuses on the subspace of verbal jokes in English within a 5–250 word limit, which I explore to assess the potential for novel outputs.

Joke Plagiarism

A longstanding issue, predating the development of LLMs, is evident in the professional stand-up comedy community, where comedians frequently accuse one another of "stealing jokes" (Wilson 2014; Cain et al. 2024). While an allegedly stolen joke may be reproduced verbatim, it can also be modified in ways that preserve key comedic elements. Consider the following examples:

Joke 1: I used to do drugs. I still do, but I used to, too. (Mitch Hedberg)

Joke 2: When my husband was a little boy, he *used to* scream and cry every time someone switched the channel when he watched his favorite cartoons. He still does, but he used to, too. (Author)

Jokes may differ entirely in execution, yet comedians may still allege theft of a *premise*—the foundational idea or setup that establishes the context and perspective for a joke or routine (Dean and Allen 2000). For instance, Jokes 3 and 4, presented in the context of stand-up comedy, share the premise that "dating apps are treated like restaurant menus":

Joke 3: You're scrolling through profiles like it's a menu, and you're like, "Nah, not feeling this one, maybe I'll order a better human tomorrow." (Grok3)

Joke 4: You don't have to leave the house anymore to go on a dinner date. I use Uber Eats to get the food and Tinder to get the human. Today, I ordered Italian… yum yum! – Oh, and I also ordered some food. (Author)

While intentional or subconscious joke theft does occur, comedians often independently conceive similar jokes (Wilson 2014). Shahaf, Horvitz, and Mankoff (2015) examined submissions to The New Yorker's cartoon caption contest, where thousands of participants submit humorous captions for a published cartoon. Their analysis revealed that many competitors independently produced similar jokes, often relying on a specific association to humorously connect two primary elements in the cartoon. For example, in a cartoon depicting a cross between a car and a wolf, 1,096 participants—17% of the total—submitted jokes centered on the car being a "hybrid." Although the exact wording varied, enabling the selection of a contest winner, the conceptual similarity was striking. Two of the ~100 jokes referencing "horsepower" illustrate this:

Joke 5: With this baby, who needs horsepower?
Joke 6: Think of it like horsepower, redefined.

Joke Essence vs. Optimization

Determining whether a joke is original or substantially similar to an existing one, beyond verbatim reproduction, requires identifying its "essence"—the unique, critical elements that define its humor. Crafting a joke demands significant skill and iterative refinement. Stand-up comedians often spend months honing a joke, meticulously adjusting phrasing to maximize comedic impact (Seinfeld and Charles 2002; Wilson 2014). While evaluating artificial intelligence's ability to generate humor, I do not dismiss the importance of such fine-tuning or other elements of comedic craftsmanship. However, when assessing a joke's originality, I propose focusing on its "essence"—the core idea or metaphorical framework of the joke. For example, in jokes 5 and 6, the essence lies in the wolf-car context and the "horsepower" reference, while the exact wording represents optimization. This essence aligns closely with the critical

components that, if replicated, could lead to accusations of joke theft. To characterize a joke's essence, I explore an empirical method grounded in the General Theory of Verbal Humor (GTVH), later discussing how it aligns with a pragmatic approach based on formulae used by professional comedy writers.

GTVH The theory, in its various iterations (e.g, OSTH; Raskin, Hempelmann, and Taylor 2009), is the most comprehensive and widely cited framework for systematically analyzing humor's elements using a symbolic, rule-based approach. It posits that humor emerges from the interplay of six knowledge resources: script opposition (SO), logical mechanism (LM), situation (SI), target (TA), narrative strategy (NS), and language (LA). These components collaborate to create and *appear to* resolve incongruities that elicit humor (Attardo and Raskin 1991; Attardo, Hempelmann, and di Maio 2002).

Not all knowledge resources, however, contribute equally to judgments of joke similarity, with SO and LM typically playing more pivotal roles (Attardo and Raskin 1991; Ruch, Attardo, and Raskin 1993; Hempelmann 2004). Research by Raskin, Hempelmann, and Taylor (2009) suggests that SO, the clash of incompatible scripts, is the primary driver of humor and thus strongly influences perceptions of joke similarity. Similarly, LM, which resolves the incongruity introduced by SO, significantly shapes similarity judgments, as jokes sharing the same resolution mechanism often feel structurally similar (Attardo, Hempelmann, and di Maio 2002). In contrast, resources like SI, TA, NS, and LA are generally secondary, as they can vary widely without altering the core humorous effect (Hempelmann and Attardo 2011). For instance, Ruch, Attardo, and Raskin (1993) found that jokes with identical SO and LM but different SI or LA were rated as more similar by participants. Collectively, GTVH provides empirical support for defining a joke's essence as its SO and LM, with other elements serving to enhance or support this core (a result replicated with professional stand-up comedians; Cain et al. 2024).

Joke Space

To address the challenge of evaluating the originality of jokes, I propose characterizing the "joke space," defined as the collection of all possible instances of humor. To make this space finite and focused, I will restrict the analysis to English verbal humor within a word limit of 5–250 words, which accommodates most verbal humor (Dunbar, Launay, and Curry 2016). This "joke space" represents a subset of all word combinations of this length that are humorous. This study does not specify a method for determining humor but assumes a human or group of humans will classify each word combination as humorous or not and rate the humorous ones for funniness. I will estimate the size of the joke space (the number of possible jokes) using combinatorics and then propose methods to reduce this space by assessing joke similarity.

Joke Space Size Evaluation

Linguistic Upper Bound We can conceive of verbal "joke space" as a subset of all possible word combinations that are, with few exceptions, grammatically and semantically coherent (with the negligible exception of jokes that rely on syntactic violations and neologisms). The term "semantically coherent" is used broadly here to encompass any combination that elicits meaningful semantic representations in the reader's mind, including novel metaphors and nonsense humor. Such combinations have been shown to activate brain regions associated with semantic processing (Amir et al. 2015).

To characterize the space of possible jokes in the English language, it is necessary to first constrain word counts. Research indicates that popular English stand-up jokes typically range from 5 to 221 words (Dunbar, Launay, and Curry 2016), while one-liners generally consist of one to two concise sentences (Chauvin 2015), with an average of 17.5 words based on a large dataset of short jokes (Moudgil 2017). An average American adult has a vocabulary of approximately 42,000 lemmas (Brysbaert et al. 2016). Thus, the combinatorial upper bound for the shortest 5-word jokes is $42,000^5$, or approximately 10^{23} combinations.

Studies on context-free grammars suggest that grammatically correct sentences represent a significantly smaller subset, with estimates ranging from 10^{-3} to 10^{-5} of all possible word combinations (Charniak 1997; Stede 1998; Pereira 2000), resulting in an upper bound of approximately 10^{18} grammatically coherent 5-word sentences. This estimate surpasses the token count in the training sets of even the largest language models by at least a factor of 100,000 (Touvron et al. 2023), a margin substantially amplified when considering unique word sequences rather than individual tokens and further increased by accounting for inflected forms beyond lemmas. This indicates that large language models have considerable potential to generate novel, grammatically correct word sequences, even for short outputs. The 5-word example underscores the vast combinatorial space of possible jokes, though most jokes exceed this length and cannot be condensed to such brevity. For instance, the median short joke length of 17 words (Moudgil 2017) yields a roughly estimated 10^{70} coherent 17-word sequences. The subset of semantically coherent sentences is reduced by an unknown factor, though insights from plagiarism detection research may provide a lower bound for this estimate.

Research on plagiarism detection indicates that n-grams of 5–8 words are generally sufficient to suggest a high probability of verbatim copying (Meuschke and Gipp 2013; Gupta 2016; Foltýnek, Meuschke, and Gipp 2019). Content evaluated for plagiarism is typically human-generated, grammatically and semantically coherent, and meaningful. Consequently, the likelihood that a 5–8-word sentence identical to one found on the public internet is original is low. Indeed, searching for a 5–8-word sentence in quotation marks often retrieves a single, unique document containing that exact phrase. This evidence further supports the notion that, even for short, meaningful sentences, there remains significant potential to generate original outputs distinct from those in training datasets or available on the public internet.

Irony How can we estimate the proportion of humorous texts among all possible meaningful texts?

A trivial case is irony, where any sentence can be transformed into an ironic one by expressing its opposite, or may itself be the ironic version of its negation, making every sentence a potential instance of irony. However, not all statements lend themselves to irony equally; for example, subjective or evaluative sentences like "the weather is bad" can be effectively inverted to "the weather is so great" with clear ironic effect, while factual or neutral statements like "the table has four legs" yield much weaker irony, if they produce irony at all, when inverted to "the table does not have four legs" due to their objective nature (Giora 1995; Wilson and Sperber 2012). Furthermore, irony is often considered less sophisticated, may not qualify as humor under certain definitions, and typically relies on contextual cues beyond the verbal exchange to convey its intent (Ortega 2013).

Theoretical Estimates Multiple theories outline essential conditions for humor, such as script opposition (Raskin 1987) and benign violation (McGraw and Warren 2010). One theory-driven approach involves estimating the frequency of these necessary (and, ideally, sufficient) conditions. However, characterizing the space of all possible jokes predicted by *every* theory is beyond the scope of this paper, as theories often disagree on which conditions are necessary. Compounding this issue, some theories, such as Zillmann's (1983) disparagement theory, have been applied in a confirmatory manner, where advocates perceive a broadly interpreted condition (e.g., disparagement of a target) post hoc in all jokes, despite its vague definition (Amir 2016).

GTVH estimates The General Theory of Verbal Humor (GTVH; Attardo and Raskin 1991) provides a framework for estimating the total number of possible jokes through a combinatorial approach based on its six Knowledge Resources. The potential number of unique jokes can be expressed as the product of the possibilities within each resource:

$$|LA| \times |NS| \times |TA| \times |SI| \times |LM| \times |SO|$$

This calculation results in an exceptionally large number due to the vast variability within each resource. For example, the Target (TA) may include any noun, named individual, or institution, and even complex systems of beliefs, while Language (LA) encompasses all possible word combinations and associated semantic information, reflecting the combinatorial complexity of linguistic structures discussed previously. However, the interdependence of these resources limits the total number of viable combinations; thus, the product is not the appropriate approach to assess the size of the space. The Situation (SI), for instance,

constrains the applicable subset of Language (LA) to contextually relevant expressions (Raskin, Hempelmann and Taylor 2009). Furthermore, the essence of a joke's humor lies in its Script Opposition (SO) and Logical Mechanism (LM). Consequently, jokes can be conceptualized as clustering around specific SO and LM combinations, with variations in the other resources (LA, NS, TA, SI) modulating the humor's effectiveness to produce local maxima of funniness within clusters. The number of distinct joke clusters can thus be estimated as:

$$|LM| \times |SO|$$

At the most abstract level, Raskin (1985) identified three primary Script Oppositions: good/bad, normal/abnormal, and actual/non-actual. Attardo (2017) further elaborates on mid-abstract and concrete levels of SOs, adding granularity to the framework. Hempelmann and Attardo (2011) propose twelve Logical Mechanisms, each with potential subtypes and concrete instantiations. At this abstract level, the total number of joke clusters is approximately 12 LMs × 3 SOs = 36 distinct joke types. The precise number of LMs depends on the level of abstraction versus granularity, though elaborating a comprehensive ontology exceeds the scope of this study (see Attardo 2001; Hempelmann and Attardo 2011). While the development of a novel Logical Mechanism by artificial intelligence would be a remarkable advancement, it is more likely that humor innovation will occur at the concrete level through specific instantiations of existing LMs and SOs in particular contexts. To estimate, at the concrete level, the number of possible jokes that are sufficiently distinct to be perceived as different, assume the two opposing scripts in an SO can be represented by two nouns (e.g., for Jokes 3-4: "Dating" vs. "Food"), with the concrete instantiation of the LM encapsulated by a single-word association or connector (e.g., "Menu" or "Commodity"). This model aligns with comedy writing guides (e.g., Toplyn 2014), which describe connecting two topic-handles—approximated here as nouns—with an associative bridge, as discussed in the next section.

Comedy Writing-Based Estimates Several humor theorists, including comedy writers such as Dean and Allen (2000) and Toplyn (2014), have proposed variations of a general method for crafting jokes. This method, when simplified, involves selecting two topical elements—referred to as "topic handles"—and identifying an association or conceptual "bridge" that links them in a meaningful or pseudo-meaningful way. For instance, in a joke involving the topic handles "wolf" and "car," the association bridge might be "hybrid." These three components—two topic handles and one bridge—constitute the minimal set forming the core or essence of a joke.

This analysis assumes that any pair of nouns can serve as topic handles for a joke, an assumption supported by the generative model Witscript (Toplyn 2023). Furthermore, it is conservatively estimated that, on average, 10 association bridges exist between any pair of nouns, based on findings

reporting over 20 bridges in some cases (Shahaf, Horvitz and Mankoff 2015). Thus, the minimum number of possible unique jokes can be expressed as:

$$N^2 \times 10/2$$

Where N represents the number of nouns, the division by 2 is because the order of nouns does not matter. For an adult English speaker with a vocabulary of approximately 42,000 lemmas, roughly one-third of which are nouns (approximately 14,000), the calculation yields $14{,}000^2 \times 5 \approx 1$ billion possible jokes.

When considering the total number of nouns in the English language, estimates vary. Dictionary-based counts, excluding proper nouns, suggest approximately 80,000 nouns (Brysbaert et al., 2016). Alternatively, WordNet 3.0 reports around 117,000 nouns. However, to account for semantic overlap, it may be more appropriate to group synonymous nouns into semantic clusters (synsets), as this reduces redundancy while excluding puns that rely on distinct word forms. WordNet-based estimates indicate 82,115 synsets for English nouns (Kutuzov et al. 2019). Using the lower bound of 80,000 nouns, the calculation $80{,}000^2 \times 5$ results in approximately 32 billion possible jokes. The estimate of possible noun-based jokes is conservative, as it could be expanded by incorporating a subset of verbs and adjectives suitable as topic handles, along with proper nouns, such as names of celebrities and popular fictional characters. Additionally, it underestimates the total, as many jokes involve setups with relational dynamics between the two topic handles (e.g., a cross between a wolf and a car; a wolf driving a car; or a wolf chasing a car).

Given that the largest LLMs are trained on datasets ranging from 1 to 10 trillion tokens (Touvron et al. 2023), and assuming a concise joke averages 17 words or approximately 20 tokens, the 1 billion possible jokes based on adult vocabulary would require roughly 20 billion tokens (1 billion × 20). While this volume could theoretically fit within the training corpus of modern LLMs, it would constitute 0.2% to 2% of their training data. This fraction may appear small, but given that most LLM training data is non-humorous (Zhao et al. 2023) and humorous content often repeats topics, specific jokes, or essentially equivalent variations (Wilson 2014), the likelihood that such a dataset includes all unique jokes based on noun combinations remains low.

Extending the calculation to all possible nouns (32 billion jokes) would require approximately 640 billion tokens, equivalent to 6.4% to 64% of a typical LLM training set. Consequently, it is reasonable to conclude that no LLM training corpus could encompass all possible jokes. This suggests that, in principle, LLMs could be prompted to generate entirely novel jokes.

Solution: Randomly Sample Joke Space

We have estimated the space of all *possible* jokes, even when considering only the core elements (or "essence") of a joke, to be several orders of magnitude larger than the

number of jokes to which the largest LLMs or human comedians could have been exposed. Despite this vast potential, both AI systems and human comedians frequently produce non-original jokes. This phenomenon can be attributed to prior exposure to similar jokes—either within an AI's training data or a comedian's personal experience—or to a tendency to focus on highly salient topics (often referred to as "the elephant in the room"; Amir and Biederman 2016; Brawer and Amir 2021; Amir et al. 2022) or topics likely to evoke strong emotional responses, such as those related to sexuality (Hempelmann and Ruch 2005).

Although joke topics in practice are not randomly selected, originality in LLM-generated jokes can be encouraged by prompting the model to create jokes based on randomly sampled topics. This can be achieved by generating pairs of random nouns and instructing the LLM to "craft a joke involving these nouns…" Consider the following example using common English nouns ("apple" and "chair"):

Jokes 7-8:
Q: Why did the apple sit alone on the chair?
A1: Because it couldn't find its *ripe* match! (Grok3)
A2: Because Newton was late that day. (Author)

In the first joke, Grok3's association ("ripe") is general and applicable to any fruit, whereas the author's association ("Newton") is specific to "apple." Jokes that leverage specific associations with the provided nouns are more likely to be original, which is a crucial factor in assessing the likelihood of joke originality. Therefore, rigorous quality control is essential to ensure that jokes generated by LLMs, even when appropriately prompted, maintain specificity to the provided noun pairs. However, simply prompting an LLM to use unique associations may be insufficient; a multi-stage prompting approach, such as incorporating a stage where the LLM generates distinct associations (e.g., Toplyn 2023), may be necessary to enhance originality. Another example uses less common English nouns ("zephyr" and "quixotism"):

Jokes 9-10:
Q: Why did the zephyr whisper to the quixotism?
A1: It wanted to blow some sense into those dreamy ideals! (Grok3)
A2: It was due to pure quixotism (Author)

Jokes generated from randomly sampled noun pairs may not be as humorous as those addressing salient or emotionally charged topics, nor as clever as jokes about topics that elicit multiple associations (Halpern et al., 1994; Toplyn, 2014). The emotional salience and enhanced memory access resulting from numerous associative links likely explain why certain regions of joke space are extensively, and potentially exhaustively, explored. The method I propose allows both LLMs and human comedians to be prompted with identical noun pairs, enabling a comparative evaluation of their outputs (e.g., through independent ratings of funniness). This approach ensures the generation of truly original jokes, addressing a methodological challenge that has previously been overlooked.

Future Directions

Non-English Jokes The number of potential nouns used in the calculations approximates the number of semantic clusters (synsets) in WordNet, suggesting that the estimates should, in principle, be applicable to any language with a comparable semantic reference framework. Given that English texts predominate in the training data of major LLMs (Zhao et al. 2023), it is paradoxically more feasible for these models to generate original jokes in languages other than English.

Formalizing an Originality Score Future research should develop a formula to quantify the likelihood of a joke's originality. The algorithm should identify two or more topic handles, considering their mutual relationships as outlined in the setup. The calculation should incorporate the perplexity of the topic handles, their co-occurrence frequency in texts, and the LM's specificity to these handles and their interrelations.

LLM Decoding The proposed framework aims to evaluate the originality of AI-generated jokes without making claims about LLMs' capabilities or prescribing specific prompting strategies. In principle, LLMs could reproduce any joke within the Joke Space through verbatim prompting (e.g., "repeat the following sentences: …"). To enhance the generation of novel jokes and reduce reliance on memorized jokes from training datasets, decoding strategies such as top-k or top-p sampling, paired with higher temperature settings, can introduce randomness to explore less probable but meaningful outputs. Furthermore, structuring joke generation as a pipeline of subtasks—such as generating unique associations between topic handles (Toplyn 2023)—can minimize the reproduction of canned jokes by encouraging the model to construct humor from novel connections, with the added benefit of diversifying semantic associations. These approaches merit further exploration to support the evaluation of LLM-generated humor originality.

Acknowledgements

References

Amir, O., Biederman, I., Wang, Z. and Xu, X. 2015. Ha Ha! Versus Aha! A Direct Comparison of Humor to Nonhumorous Insight for Determining the Neural Correlates of Mirth. *Cerebral Cortex*, 25(5): 1405-1413.

Amir, O. 2016. The Frog Test: A Tool for Measuring Humor Theories' Validity and Humor Preferences. *Frontiers in Human Neuroscience*, 10: 40.

Amir, O. and Biederman, I. 2016. The Neural Correlates of Humor Creativity. *Frontiers in Human Neuroscience*, 10: 597.

Amir, O., Utterback, K.J., Lee, J., Lee, K.S., Kwon, S., Carroll, D.M. and Papoutsaki, A. 2022. The Elephant in the Room: Attention to Salient Scene Features Increases with Comedic Expertise. *Cognitive Processing*, 23(2): 203-215.

Attardo, S. and Raskin, V. 1991. Script Theory Revis (it) ed: Joke Similarity and Joke Representation Model. *Humor: International Journal of Humor Research,* 4(3-4): 293–347.

Attardo, S., 2001. *Humorous Texts: A Semantic and Pragmatic Analysis* (Vol. 6). Berlin: Walter de Gruyter.

Attardo, S. 2017. The GTVH and Humorous Discourse. In Chlopicki, W., and Brzozowska, D., eds., *Humorous Discourse*, 93–107. Berlin/Boston: De Gruyter Mouton.

Attardo, S., Hempelmann, C. F. and di Maio, S. 2002. Script Oppositions and Logical Mechanisms: Modeling Incongruities and Their Resolutions. *Humor: International Journal of Humor Research,* 15(1): 3–46.

Binsted, K. and Ritchie, G., 1997. Computational Rules for Generating Punning Riddles. *Humor: International Journal of Humor Research,* 10(1): 25–76.

Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S. 2021, March. On the dangers of Stochastic Parrots: Can Language Models be too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*: 610-623.

Brawer, J. and Amir, O. 2021. Mapping the 'Funny Bone': Neuroanatomical Correlates of Humor Creativity in Professional Comedians. *Social Cognitive and Affective Neuroscience*, 16(9): 915-925.

Brysbaert, M., Stevens, M., Mandera, P. and Keuleers, E. 2016. How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age. *Frontiers in Psychology*, 7: 1116.

Cain, K., Gimbel, S., Howard, L., Maronna, B. and Beirne, S. 2024. Joke Synonymy Sensitivity Among Working Comedians and the General Theory of Verbal Humor. *Humor: International Journal of Humor Research*, 37(4): 513-528.

Charniak, E. 1997. Statistical Parsing with a Context-Free Grammar and Word Statistics. *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97)*: 598-603. Providence, Rhode Island: AAAI Press.

Chauvin, C. 2015. One-Liners and Linguistics:(Re) Interpretation, Context and Meaning. *Angles: New Perspectives on the Anglophone World*, 1: 1-22.

Chomsky, N., Roberts, I. and Watumull, J. 2023. The False Promise of ChatGPT. *The New York Times*, 8 March. New York: The New York Times Company.

Dean, G. and Allen, S. 2000. *Step by Step to Stand-Up Comedy*. Portsmouth, NH: Heinemann.

Dunbar, R.I., Launay, J. and Curry, O. 2016. The Complexity of Jokes is limited by Cognitive Constraints on Mentalizing. *Human Nature*, 27(2): 130-140.

Foltýnek, T., Meuschke, N. and Gipp, B., 2019. Academic Plagiarism Detection: A Systematic Literature Review. *ACM Computing Surveys (CSUR)*, 52(6): 1-42.

Giora, R. 1995. On irony and negation. *Discourse processes*, 19(2): 239-264.

Gupta, D. 2016. Study on Extrinsic Text Plagiarism Detection Techniques and Tools. *Journal of Engineering Science & Technology Review*, 9(5): 8.

Halpern, C., Close, D. and Johnson, K. H. 1994. *Truth in Comedy*. Colorado Springs, CO: Meriwether.

Hempelmann, C.F. 2004. Script Opposition and Logical Mechanism in Punning. *Humor: International Journal of Humor Research*, 17(4): 381–392.

Hempelmann, C.F. and Ruch, W. 2005. 3WD meets GTVH: Breaking the Ground for Interdisciplinary Humor Research. *Humor: International Journal of Humor Research*, 18(4): 353-387.

Hempelmann, C.F. and Attardo, S. 2011. Resolutions and their Incongruities: Further Thoughts on Logical Mechanisms. *Humor: International Journal of Humor Research,* 24(2): 125–149.

Jentzsch, S. and Kersting, K. 2023, July. ChatGPT is Fun, but it is not Funny! Humor is still Challenging Large Language Models. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*: 325-340. Toronto, Canada: Association for Computational Linguistics.

Moudgil, A 2017. Short Jokes Dataset [Data set]. *Kaggle*. Retrieved May 19, 2025, from https://www.kaggle.com/datasets/abhinavmoudgil95/short-jokes

Kutuzov, A., Dorgham, M., Oliynyk, O., Biemann, C. and Panchenko, A., 2019, June. Learning Graph Embeddings from WordNet-based Similarity Measures. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (SEM-2019)*:125-135.

McGraw, A.P. and Warren, C. 2010. Benign Violations: Making Immoral Behavior Funny. *Psychological Science*, 21(8): 1141-1149.

Meuschke, N. and Gipp, B. 2013. State-of-the-Art in Detecting Academic Plagiarism. *International Journal for Educational Integrity*, 9(1): 50-71.

Ortega, M.B.A. 2013. An Approach to Verbal Humor in Interaction. *Procedia-Social and Behavioral Sciences*, 95: 594-603.

Pereira, F. 2000. Formal Grammar and Information Theory: Together Again?. *Phil. Trans. R. Soc. Lond. A*, 358: 1239-1253.

Raskin, V. 1985. *Semantic mechanisms of humor*. Dordrecht, Netherlands: D. Reidel.

Raskin, V. 1987. Linguistic Heuristics of Humor: A Script-Based Semantic Approach. *International Journal of the Sociology of Language*, 65: 11-26.

Ruch, W., Attardo, S. and Raskin, V. 1993. Toward an empirical verification of the general theory of verbal humor. *Humor: International Journal of Humor Research*, 6(2): 123–136.

Seinfeld, J. and Charles, C. 2002. *Comedian* [Documentary]. New York: Bridgnorth Films.

Shahaf, D., Horvitz, E. and Mankoff, R. 2015, August. Inside jokes: Identifying Humorous Cartoon Captions. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 1065-1074. Sydney, NSW, Australia: ACM.

Stede, M., 1998. A Generative Perspective on Verb Alternations. *Computational Linguistics*, 24(3): 401-430.

Toplyn, J. 2014. *Comedy Writing for Late-Night TV: How to Write Monologue Jokes, Desk Pieces, Sketches, Parodies, Audience Pieces, Remotes, and Other Short-Form Comedy*. Rye, New York: Twenty Lane Media, LLC.

Toplyn, J. 2023. Witscript 3: A Hybrid AI System for Improvising Jokes in a Conversation. *arXiv preprint arXiv:2301.02695*.

Toplyn, J. and Amir, O. 2025, January. Can AI Make Us Laugh? Comparing Jokes Generated by Witscript and a Human Expert. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*: 71-78. Online: Association for Computational Linguistics.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. and Bikel, D. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Veale, T. 2024. From Symbolic Caterpillars to Stochastic Butterflies: Case Studies in Re-Implementing Creative Systems with LLMs. In *Proceedings of the 15th International Conference on Computational Creativity (ICCC)*. Jönköping, Sweden: Association for Computational Creativity.

Wilson, G.D. 2014. *The Complete Guide to Stand-up: Everything You Need to Know, from Open-Mics to Going Pro!* Minneapolis: The Comedy Institute.

Wilson, D. and Sperber, D. 2012. Explaining Irony. In *Meaning and Relevance*: 123-145. Cambridge: Cambridge University Press.

Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z. and Du, Y. 2023. A Survey of Large Language Models. *arXiv e-prints*: arXiv-2303.

Zillmann, D., 1983. Disparagement humor. In *Handbook of Humor Research: Volume 1: Basic Issues*: 85-107. New York, NY: Springer New York.