Linguistic differences in humour: a feature-based comparison between human and AI-generated jokes

Freya Pan
Crofton House School, Canada

https://doi.org/10.36505/TheLinguisticProceedings/2025/17/02/015/000701

Abstract

Humour is one of the most creative and socially embedded forms of human communication. With the rise of large language models (LLMs), questions about whether artificial intelligence can identify and generate jokes have become central to computational humour studies. This paper conducts a systematic comparison of human- and AI-generated jokes across multiple linguistic dimensions: lexical, syntactic, affective, semantic, and prosodic. To this end, we first constructed a parallel joke corpus by instructing LLMs to extract meta-information from human-authored texts and generate comparable jokes. By applying Elastic Net logistic regression on our parallel corpus of human and AI humour, we found that AI-generated jokes relied on connectives and retrospective narration. In contrast, human humour was characterised by greater lexical density, conditional subordination, emotive language, and semantic incongruity, as well as an increased use of nouns, interrogative words, social references, and intentional rhythmic devices. These findings highlight the gap between human and AI humour in terms of creativity, emotional depth, and prosodic design. The study contributes both to the theoretical validation of humour mechanisms and to the practical advancement of explainable and human-like humour generation in LLMs.

Keywords: computational humour, quantitative linguistics, stylometric, machine learning

Introduction

Humour was once thought of as a uniquely human quality. It is deeply embedded in the human condition, shaped by our shared vulnerabilities and capacity to find mirth amidst hardship (Morreall, 2011; Martin & Ford, 2018). However, with the rapid development of large language models (LLMs) in recent years, whether machines can truly understand or produce humour has become an increasingly controversial topic. Although existing studies have shown that LLMs demonstrate certain humour-generation capabilities (Gorenz & Schwarz, 2024), systematic linguistic evidence explaining the fundamental differences between human and machine humour remains lacking. Hence, the primary goal of this study is to identify which linguistic features most effectively distinguish and explain the differences between human and machine humour. To achieve this, this paper first constructs a comparable corpus. Based on representative human humour corpora, prompt strategies, grounded in humour theory, were designed to extract joke meta-information. Then, three mainstream large language models

were used to generate machine humour texts with controlled length and comparable semantics. Second, this paper proposes a systematic linguistic indicator framework covering five dimensions: lexical, syntactic, semantic, emotional, and prosodic. Finally, based on these linguistic features, this paper employs interpretable machine learning for human-machine humour classification and feature importance analysis.

Methodology

Comparable human and AI-generated humour data were collected for this study to enable systematic analysis of human-AI humour differences. For human humour, three datasets were selected: Reddit Jokes, Short Jokes, and Pun of the Day. For AI-generated humour, three large language models (GPT-4.1-mini, Llama-3.3-70B, DeepSeek-V3-Chat) were employed. Human jokes were first analysed using a prompt template to extract meta-information, including humour keywords, semantic incongruity, and current events, based on semantic incongruity theory and the General Theory of Verbal Humour (GTVH) (Ruch et al., 1993; Attardo, 1997). A second prompt guided the models to generate jokes within word limits and context comparable to the original human jokes.

A multidimensional metric system covering lexical, syntactic, emotional, semantic, and prosodic features was applied to analyse stylistic differences. Texts were vectorised across these features for computational modelling. Elastic Net logistic regression was used to distinguish human and machine humour, with separate datasets split 8:2 for training and testing. Three classification experiments (Human vs GPT-4.1-mini, Human vs Llama-3.3-70B, Human vs DeepSeek-V3-Chat) achieved average F1-scores above 0.8. Top-20 indicators per dimension were extracted, and features consistently appearing across all three experiments were marked as strong indicators of human-machine humour differences. The multidimensional metric system and the complete statistical results of this study are provided in:
https://github.com/adventurer-chy/human_AI_humor_comparison.

Results and discussion

Lexical distinctions

At the lexical and syntactic levels, there are several linguistic metrics that distinguish human and machine humour. At the lexical level, Lexical Density (LD3) human humour (7.459) surpasses machine humour (5.292). Human humour tends to rely on high-information-load content words (nouns, verbs, adjectives), while machine humour uses more function words, which results in reduced content density and a lack of lexical innovation. Similarly, the Noun Ratio (0.190 vs. 0.168) suggests that human humour achieves narrative vividness by introducing specific settings and characters. This aligns with Barnden's theory

(Barnden, 2017), which emphasises that concrete nouns are crucial for linguistic expressiveness. By contrast, machine humour lacks this scenic texture. In addition, the Wh-Pronoun Ratio (0.013 vs. 0.004) shows that human humour uses interrogative structures more frequently. Meanwhile, machine humour relies frequently on declarative statements and thus tends to produce flatter and more expected punchlines.

Syntactic distinctions

At the syntactic level, Average Sentence Length (17.211 vs. 13.365) indicates that human humour favours longer and more elaborate structures. Noun Subject Ratio (0.128 vs. 0.131) suggests machine humour depends on explicit SVO templates. In contrast, human humour more frequently features omitted or implicit subjects and invites the audience to co-construct meaning.

Affective and semantic distinctions

At the affective and semantic level, human and machine humour differ in emotional and contextual expressiveness. The proportion of social words is higher in human humour (0.130) than in machine humour (0.115), which reflects that human-created jokes are more grounded in shared human experiences. In contrast, machine humour tends to rely on object- or event-focused descriptions, which often lack relational depth. The proportion of emotive words (LIWC) is also higher in human humour (0.839) compared to machine humour (0.804), which suggests how human jokes often integrate emotional cues, whereas machine humour leans toward neutrality and emotional flatness. In addition, auxiliary and modal verbs are used more often by humans. Human humour's higher frequency of modals reflects a greater capability for handling ambiguity and subtlety, while this lack in machines suggests rigid propositional framing. Lastly, semantic incongruity (0.620 vs. 0.605) and WSD-based incongruity (0.694 vs. 0.681) are higher in human humour. This indicates that human-generated jokes are more adept at creating unexpected contrasts in conceptual combinations. Together, these findings indicate that human humour balances emotional resonance, social context, as well as surprise, while machine humour remains relatively literal and semantically predictable.

Rhythmic and phonetic distinctions

Human humour (0.030) slightly exceeds machine humour (0.027) in rhythmic features. Machines generate rhymes or repetitions incidentally, and thus lack a deliberate rhythm. Human humour (0.894) also far exceeds machine humour (0.193) in the use of repetitive dyads. Repetition is one of the key techniques for comedians, as it helps establish rhythmic patterns and dramatizes tension (Schwarz, 2009). In contrast, AI-generated humour rarely employs repetition deliberately or strategically, which results in noticeably weaker rhythmic patterns

and reduced dramatic structure in its narrative. Human humour (0.049) again outperforms machine humour (0.021) in rhyme similarity. Machines rarely design rhyme explicitly, which reduces the musicality that well-crafted humour often relies on. Overall, human humour exhibits more intentional musicality, while machine humour remains comparatively limited in its use of sound.

Conclusion

Having systematically compared human- and AI-generated humour across lexical, syntactic, emotional, semantic, and prosodic dimensions, findings show that, despite structural similarities, human jokes demonstrate greater creativity, emotional depth, as well as more deliberate rhythmic expressiveness. These differences highlight how humour reflects uniquely human cognitive and social capacities that AI has yet to replicate. By identifying where AI humour differs from human-like expression, this research provides empirical support for humour theory and a framework for enhancing computational humour. Future work could integrate cross-cultural data and interpretability metrics to develop more context-sensitive humour systems, ultimately improving AI's ability to engage in authentic, human-like communication.

References

Attardo, S. 1997. The semantic foundations of cognitive theories of humor. Humor 10, 395–420.

Barnden, J. 2017. Exploding the creativity myth: The computational foundations of linguistic creativity by Tony Veale. Metaphor and Symbol 32, 52–54.

Gorenz, D., Schwarz, N. 2024. How funny is ChatGPT? A comparison of human- and A.I.-produced jokes. PLOS ONE 19, e0305364.

Martin, R. A., Ford, T. E. 2018. The psychology of humor: An integrative approach. London, Academic Press.

Morreall, J. 2010. Comic relief: a comprehensive philosophy of humor. New York, Choice Reviews Online.

Ruch, W., Attardo, S., Raskin, V. 1993. Toward an empirical verification of the General Theory of Verbal Humor. Humor 6, 123–136.

Schwarz, J. 2009. Linguistic aspects of verbal humor in stand-up comedy. Saarbrücken, Universität des Saarlandes.