Reverse-Engineering Satire, or "Paper on Computational Humor Accepted despite Making Serious Advances"

Robert West*
EPFL
robert.west@epfl.ch

Eric Horvitz
Microsoft Research
horvitz@microsoft.com

Abstract

Humor is an essential human trait. Efforts to understand humor have called out links between humor and the foundations of cognition, as well as the importance of humor in social engagement. As such, it is a promising and important subject of study, with relevance for artificial intelligence and human–computer interaction. Previous computational work on humor has mostly operated at a coarse level of granularity, e.g., predicting whether an entire sentence, paragraph, document, etc., is humorous. As a step toward deep understanding of humor, we seek fine-grained models of attributes that make a given text humorous. Starting from the observation that satirical news headlines tend to resemble serious news headlines, we build and analyze a corpus of satirical headlines paired with nearly identical but serious headlines. The corpus is constructed via *Unfun.me*, an online game that incentivizes players to make minimal edits to satirical headlines with the goal of making other players believe the results are serious headlines. The edit operations used to successfully remove humor pinpoint the words and concepts that play a key role in making the original, satirical headline funny. Our analysis reveals that the humor tends to reside toward the end of headlines, and primarily in noun phrases, and that most satirical headlines follow a certain logical pattern, which we term *false analogy*. Overall, this paper deepens our understanding of the syntactic and semantic structure of satirical news headlines and provides insights for building humor-producing systems.

1 Introduction

Humor is a uniquely human trait that plays an essential role in our everyday lives and interactions. Psychologists have pointed out the role of humor in human cognition, including its link to the identification of surprising connections in learning and problem solving, as well as the importance of humor in social engagement (Martin 2010). Humor is a promising area for studies of intelligence and its automation: it is hard to imagine a computer passing a rich Turing test without being able to understand and produce humor. As computers increasingly take on conversational tasks (e.g., in chat bots and personal assistants), the ability to interact with users naturally is gaining importance, but human–computer interactions will never be truly natural without giving users the option to say something funny and have it understood that way; e.g., recent work has shown that misunderstanding of playful quips can be the source of failures in conversational dialog in open-world interaction (Andrist et al. 2016).

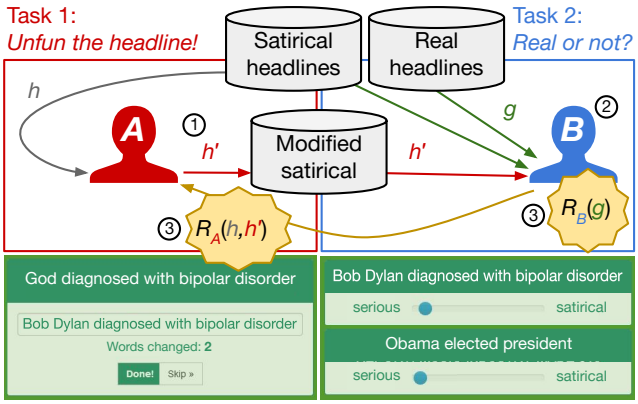


Figure 1: *Unfun.me*, a game for building a corpus of pairs (h, h') of satirical and similar-but-serious-looking headlines. Numbers: order of steps. Screenshots: running example ($h =$ *God diagnosed with bipolar disorder;* $h' =$ *Bob Dylan diagnosed with bipolar disorder;* $g =$ *Obama elected president*).

Given how tied humor is to the human condition, the phenomenon has challenged some of the greatest thinkers throughout history and has been the subject of much academic research across over 20 disciplines (Raskin 2008), including computer science (Binsted et al. 2006), where researchers have developed algorithms for detecting, analyzing, and generating humorous utterances (cf. Sec. 6).

The automated analysis of humor is complicated by the fact that most humorous texts have a complex narrative structure that is difficult to disentangle; e.g., typical jokes—the type of humorous text studied most in the literature—carefully set the stage to build certain expectations in the audience, which are then turned upside down in the punchline. To circumvent the difficulties imposed by narrative structure, we focus on a specific humorous genre: satirical news. Satirical news articles, on the surface, mimic the format typical of mainstream journalism, but unlike serious news articles, they do not aim to relate facts, but rather to ridicule individuals, groups, or society. Crucially, though, satirical news stories are typically written headline-first: only if the

*Research done partly at Microsoft Research.

headline is funny in and of itself is the rest of the story written (Glass 2008). This is markedly different from real news stories and means that **satirical news headlines** can be studied in isolation from the full stories, whose essence they convey in a concise form with minimal narrative structure.

An additional advantage of satirical headlines is that they mimic the formulaic style of serious news headlines, which limits their syntactic variability and allows us to better control for syntax and focus on semantics. Moreover, satirical headlines are similar to serious news headlines not only in style but also in content: changing a single word often suffices to make a satirical headline sound like serious news.

Running example. For instance, changing *God* to *Bob Dylan* turns the satirical headline ***God diagnosed with bipolar disorder,*** which was published in the satirical newspaper *The Onion*, into ***Bob Dylan diagnosed with bipolar disorder,*** which could appear *verbatim* in a serious newspaper.

A large corpus of such pairs of satirical and similar-but-serious-looking headlines would open up exciting opportunities for humor research. For instance, it would allow us to understand why a satirical text is funny at a finer granularity than previously possible, by identifying the exact words that make the difference between serious and funny. This is a striking difference from most previous research, where usually the *average* satirical headline is compared to the *average* serious one (Mihalcea and Pulman 2007). Moreover, while the principal goal of this research has been to achieve new insights about humor, we also imagine new applications. For example, if we attained a grasp on the precise differences between satirical and serious headlines, we might be able to create procedures for transforming real news headlines into satirical headlines with minimal changes.

To create an aligned corpus, a first idea would be to automatically pair satirical with serious news headlines: start with a satirical headline and find the most similar serious headline written around the same time. It is hard to imagine, though, that this process would yield many pairs of high lexical and syntactic similarity. An alternative idea would be to use crowdsourcing: show serious headlines to humans and ask them to turn them into satirical headlines via minimal edits. Unfortunately, this task requires a level of creative talent that few people have. Even at *The Onion,* America's most prominent satirical newspaper, only 16 of 600 headlines generated each week (less than 3%) are accepted (Glass 2008).

The crucial observation is that the task is much easier in the reverse direction: it is typically straightforward to **remove the humor** from a satirical headline by applying small edits that turn the headline into one that looks serious and could conceivably be published in a real news outlet. In other words, reversing the creative effort that others have already invested in crafting a humorous headline requires much less creativity than crafting the headline in the first place. We thus adopt this reverse-crowdsourcing approach, by designing a **game with a purpose** (von Ahn and Dabbish 2008).

The game is called ***Unfun.me*** and is described graphically in Fig. 1. A player A of the game is given a satirical news headline h and asked to modify it in order to fool other players into believing that the result h' is a real headline from a serious news outlet. The reward $R_A(h,h')$ received by the player A who modified the satirical headline increases with the fraction of other players rating the modified headline h' as serious and decreases with the number of words changed in the original headline h.

Contributions. Our main contributions are twofold. First, we present *Unfun.me*, an online game for collecting a corpus of pairs of satirical news headlines aligned to similar-but-serious-looking headlines (Sec. 2). Second, our analysis of these pairs (Sec. 3–5) reveals key properties of satirical headlines at a much finer level of granularity than prior work (Sec. 6). Syntactically (Sec. 4), we conclude that the humor tends to reside in noun phrases, and with increased likelihood toward the end of headlines, giving rise to what we term "micro-punchlines". Semantically (Sec. 5), we observe that original and modified headlines are usually opposed to each other along certain dimensions crucial to the human condition (e.g., *high vs. low stature, life vs. death*), and that satirical headlines are overwhelmingly constructed according to a *false-analogy* pattern. We conclude the paper by discussing our findings in the context of established theories of humor (Sec. 7).

2 Game description: *Unfun.me*

Here we introduce *Unfun.me*, our game for collecting pairs of satirical and similar-but-serious-looking headlines. The game, available online at `http://unfun.me` and visually depicted in Fig. 1, challenges players in two tasks.

Task 1: ***Unfun the headline!*** This is the core task where the reverse-engineering of satire happens (left panel in Fig. 1). A player, A, is given a satirical headline h and is asked to turn it into a headline h' that could conceivably have been published by a serious news outlet, by changing as few words as possible.

Task 2: ***Real or not?*** Whether on purpose or not, player A may have done a bad job in task 1, and h' may still be humorous. Detecting and filtering such cases is the purpose of task 2 (right panel in Fig. 1), where h' is shown to another player, B, who is asked to indicate her belief $p_B(h')$ that h' comes from a serious news outlet using a slider bar ranging from 0% to 100%. We shall refer to $p_B(h')$ as B's **seriousness rating** of h'. For reasons that will become clear below, player B also indicates her belief $p_B(g)$ for a second, unmodified headline g (unrelated to h) that originates from either a serious or a satirical news outlet. The two headlines h' and g are presented in random order, in order to avoid biases.

For the purpose of incentivizing players to make high-quality contributions, we reward them as follows.

Reward for task 1. As player A is supposed to *remove the humor* from h via a *minimal modification,* his reward $R_A(h,h')$ increases (1) with the average rating $r(h')$ that the modified headline h' receives from all n players $B_1,\ldots,B_n$ who rate it and (2) with the similarity $s(h,h')$ of h and h':

$$R_A(h,h') = \sqrt{r(h')\, s(h,h')}\,, \tag{1}$$

$$\text{where} \quad r(h') = \frac{1}{n}\sum_{i=1}^{n} p_{B_i}(h'), \quad s(h,h') = 1 - \frac{d(h,h')}{\max\{|h|,|h'|\}},$$

where, in turn, $|x|$ is the number of tokens (i.e., words) in a string x, and $d(h,h')$, the **token-based edit distance** (Navarro 2001) between h and h', i.e., the minimum number of insertions, deletions, and substitutions by which h can be transformed into h', considering as the basic units of a string its tokens, rather than its characters. The geometric mean was chosen in Eq. 1 because it is zero whenever one of the two factors is zero (which is not true for the more standard arithmetic mean): a modified headline that seems very serious, but has nothing to do with the original, should not receive any points, nor should a headline that is nearly identical to the original, but retains all its humor.

Reward for task 2. Since player B's very purpose is to determine whether h' is without humor, we do not have a ground-truth rating for h'. In order to still be able to reward player B for participating in task 2, and to incentivize her to indicate her true opinion about h', we also ask her for her belief $p_B(g)$ regarding a headline g for which we do have the ground truth of "serious" vs. "satirical". The reward $R_B(g)$ that player B receives for rating headline g is then

$$R_B(g) = \begin{cases} \log(p_B(g)) & \text{if } g \text{ is serious,} \\ \log(1-p_B(g)) & \text{if } g \text{ is satirical.} \end{cases} \quad (2)$$

Note that this is a *proper scoring rule* (Gneiting and Raftery 2007), i.e., player B maximizes her expected reward by indicating her true belief. This would not be true for the more straightforward scoring formula without logarithms, which would drive players to report beliefs of 0 or 1 instead of their true beliefs. Also, as h' and g are shown in random order, B does not know which is which, and her optimal strategy is to indicate her true belief on both.

Overall game flow. Whenever a user wants to play, we generate a type-1 task with probability $\alpha = 1/3$ and a type-2 task with probability $1-\alpha = 2/3$, such that we can collect two ratings per modified headline. As mentioned, ratings from task 2 can serve as a filter, and we can increase its precision at will by decreasing α. To make rewards more intuitive and give more weight to the core task 1, we translate and scale rewards such that $R_A(\cdot,\cdot) \in [0,1000]$ and $R_B(\cdot) \in [0,200]$. We also implemented additional incentive mechanisms such as badges, high-score tables, and immediate rewards for participating, but we omit the details for space reasons.

Satirical and serious headlines. The game requires corpora of satirical as well as serious news headlines as input. Our satirical corpus consists of 9,159 headlines published by the well-known satirical newspaper *The Onion;* our serious corpus, of 9,000 headlines drawn from 9 major news websites.

Data and code. We make the data collected via *Unfun.me*, as well as our code for analyzing it, publicly available online (West and Horvitz 2019).

3 Analysis of game dynamics

Via *Unfun.me*, we have collected 2,801 modified versions h' for 1,191 distinct satirical headlines h (2.4 pairs per satirical headline). All but 7 modified headlines have received at least one rating, and 1,806 (64%), at least two (mean/median: 2 ratings per modified headline). The modified headlines (ratings) came from 582 (546) unique user ids (mean/median: 4.8/2 modified headlines per user; 10/4 ratings per user).

We start by analyzing the edit operations players perform in task 1 and the seriousness ratings they provide in task 2. The main objects of study are pairs (h,h') consisting of an original satirical headline h and a modified version h', which we shall simply call **pairs** in what follows.

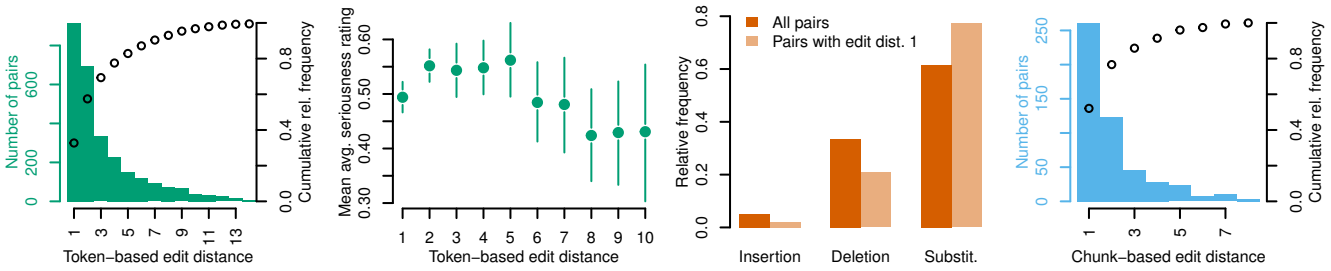
Edit distance. The first interesting question is how much players tend to modify original satirical headlines h in order to expunge the humor from them. We quantify this notion via the token-based edit distance $d(h,h')$ between the satirical headline h and the modified version h' (cf. Sec. 2). Fig. 2(a), which plots the distribution of edit distance, shows that very small edits are most common, as incentivized by the reward structure of the game (Eq. 1). In particular, 33% of all pairs have the smallest possible edit distance of 1, and 57% (69%) have a distance up to 2 (3).

Tradeoff of edit distance vs. seriousness rating. The reward structure of the game (Eq. 1) does not, however, exclusively encourage small edits. Rather, there is a tradeoff: larger edits (bad) make it easier to remove the humor (good), while smaller edits (good) run the risk of not fully removing the humor (bad). Fig. 2(b), which plots the mean average seriousness rating $r(h')$ of modified headlines h' as a function of the edit distance $d(h,h')$, shows how this tradeoff plays out in practice. For edit distances between 1 and 5 (83% of all pairs, cf. Fig. 2(a)), seriousness ratings correlate positively with edit distance. In particular, it seems harder to remove the humor by changing one word than by changing two words, whereas the marginal effect is negligible when allowing for even larger edits. The positive correlation does not hold for the much smaller number (17%) of pairs with an edit distance above 5. Inspecting the data, we find that this is caused by headlines so inherently absurd that even large edits cannot manage to remove the humor from them.

Seriousness ratings. Recall that, in task 2, players attribute seriousness ratings to modified headlines h', as well as to unmodified serious or satirical headlines g. We find that, in all three cases, the distribution of seriousness ratings is bimodal, with extreme values close to 0 or 1 being most common. Hence, we binarize ratings into two levels, "satirical" (rating below 0.5) and "serious" (rating above 0.5).

In order to see how people rate serious, satirical, and modified headlines, respectively, Table 1 aggregates ratings by headline (considering only the 1,806 headlines with at least two ratings) and splits the headlines into three groups: "consensus serious" (over 50% "serious" ratings), "no consensus" (exactly 50%), and "consensus satirical" (under 50%).

We make two observations. First, modified headlines h' (column 3 of Table 1) are distributed roughly evenly over the three groups; i.e., there are about as many headlines from which the humor has been successfully removed ("consensus serious") as not ("consensus satirical"). The most useful modified headlines for our purposes are those from the "consensus serious" group, as they likely do not carry the humor of the original h anymore. Hence, we shall restrict our subse-



(a) (b) (c) (d)

Figure 2: **(a)** Distribution of token-based edit distance in headline pairs collected via *Unfun.me*. **(b)** Tradeoff of edit distance vs. seriousness rating (only pairs with at least 2 ratings; with bootstrapped 95% confidence intervals). **(c)** Distribution of token-based edit operations (successful pairs only, cf. Sec. 3). **(d)** Distribution of chunk-based edit distance (successful pairs only).

Table 1: Rating distributions for pairs with at least 2 ratings. "No consensus" large as most pairs have exactly 2 ratings.

Aggregate rating	Serious	Satirical	Modified
Consensus serious	777 (57%)	105 (8%)	654 (36%)
No consensus	447 (33%)	368 (27%)	570 (32%)
Consensus satirical	133 (10%)	871 (65%)	582 (32%)

quent analyses to the corresponding 654 **successful pairs.**[1] Second, the ratings are heavily skewed toward the ground truth for unmodified serious (column 1) and satirical (column 2) headlines; i.e., players can typically well distinguish serious from satirical headlines (but cf. discussion in Sec. 7).

Insertions, deletions, substitutions. When computing the edit distance $d(h,h')$ using dynamic programming, we can also keep track of an optimal sequence of edit operations (insertions, deletions, substitutions) for transforming h into h' (Navarro 2001). In Fig. 2(c), we plot the distribution of edit operations, macro-averaged over all pairs. We see that substitutions clearly dominate (61%), followed by deletions (34%), with insertions being very rare (5%).

Pairs with edit distance 1 are particularly interesting, as they are the most similar, as well as the most frequent (Fig. 2(a), footnote 1). Also, the optimal edit sequence may not be unique in general, but for edit distance 1 it is. Hence, Fig. 2(c) also displays the distribution over edit operations for pairs with edit distance 1 only. Here, substitutions dominate even more (77%), and insertions are even rarer (2%).

Reversing the direction of the editing process, we hence conclude that writers of satirical headlines tend to work overwhelmingly by substituting words in (hypothetical) similar-but-serious headlines, and to a certain degree by adding words, but very rarely by deleting words.

4 Syntactic analysis of aligned corpus

Next, we go one level deeper and ask: what parts of a satirical headline should be modified in order to remove the humor from it, or conversely, what parts of a serious headline should be modified in order to add humor? We first tackle this question from a syntactic perspective, before moving to a deeper, semantic perspective in Sec. 5.

[1] As a sanity check, we computed the edit-distance distribution for successful pairs only, finding no big differences from Fig. 2(a).

From tokens to chunks. We analyze syntax at an intermediate level of abstraction between simple sequences of part-of-speech (POS) tags and complex parse trees, by relying on a *chunker* (also called *shallow parser*). We use OpenNLP's maximum entropy chunker (Berger, Pietra, and Pietra 1996), after retraining it to better handle pithy, headline-style text. The chunker takes POS-tagged text as input and groups subsequent tokens into meaningful phrases **(chunks)** without inferring the recursive structure of parse trees; e.g., our running example (Sec. 1) is chunked as [NP *Bob Dylan*] [VP *diagnosed*] [PP *with*] [NP *bipolar disorder*] (chunk labels expanded in Table 2). Chunks are handy because they abstract away low-level details; e.g., changing *God* to *Bob Dylan* requires a token-based edit distance of 2, but a chunk-based distance of only 1, where the latter is more desirable because it more closely captures the conceptual modification of one entity being replaced by another entity.

Chunking all 9,159 original headlines from our *The Onion* corpus, we find the most frequent chunk pattern to be NP VP NP PP NP (4.8%; e.g., H2 in Fig. 4(a)), followed by NP VP NP (4.3%; e.g., H4) and NP VP PP NP (3.3%; e.g., H9).

To control for syntactic effects, it is useful to study a large number of pairs (h,h') where all original headlines h follow a fixed syntactic pattern. We therefore gave priority to headlines of the most frequent pattern (NP VP NP PP NP) for a certain time period when sampling satirical headlines as input to task 1, such that, out of all 2,801 (h,h') pairs collected in task 1, h follows that pattern in 21% of all cases.

Chunk-based edit distance. Recomputing edit distances at the chunk level, rather than the token level, we obtain the chunk-based edit distance distribution of Fig. 2(d). It resembles the token-based edit distance distribution of Fig. 2(a), with the difference that the smallest possible distance of 1 is even more prevalent (52% vs. 33% of pairs), due to the fact that modifying a single chunk frequently corresponds to modifying multiple tokens. Since, moreover, the vast majority (97%) of all single-chunk edits are substitutions, we now focus on 254 (h,h') pairs where exactly one chunk of h has been modified (henceforth **single-substitution pairs**). This accounts for about half of all successful pairs (after discarding pairs that were problematic for the chunker).

Dominance of noun phrases. We now ask which syntactic chunk types (noun phrases, verb phrases, etc.) are modified to remove humor. In doing so, we need to be careful,

Table 2: Distribution of syntactic chunk types in single-substitution pairs (only showing types modified at least once).

Label	Chunk type	Modified	Prior	Lift
NP	Noun phrase	89.37%	58.63%	1.52
VP	Verb phrase	9.45%	20.15%	0.47
ADJP	Adjective phrase	0.79%	1.49%	0.53
PP	Preposition	0.39%	17.40%	0.02

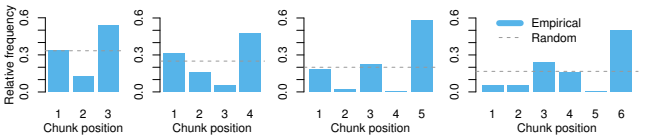


Figure 3: Distributions of modified chunk positions in single-substitution pairs, for original headlines containing 3 to 6 chunks (number of pairs for each length: 24, 38, 123, 38).

as some chunk types are more common *a priori* than others; e.g., 59% of all chunks in original satirical headlines are noun phrases, 20%, verb phrases, etc. We therefore compare the empirical distribution of modified chunks with this prior distribution, via the ratio of the two (termed *lift*). Table 2 shows that noun phrases constitute 89% of the modified chunks (lift 1.52), whereas all other chunk types are less frequent than under the prior. We conclude that the humor of satirical news headlines tends to reside in noun phrases.

Micro-punchlines. We now ask where in terms of location within a headline the humor tends to reside. To answer this question, we compute the position of the modified chunk in each headline's chunk sequence and plot the distribution of modified positions in Fig. 3. We see that, regardless of headline length, modifications to the last chunk are particularly overrepresented.[2] This is an important finding: we have previously (Sec. 1) argued that satirical headlines consist of a punchline only, with minimal narrative structure, and indeed it was this very intuition that led us to investigate headlines in isolation. Given Fig. 3, we need to revise this statement slightly: although satirical headlines consist of a single sentence, they are often structured—at a micro-level—akin to more narrative jokes, where the humorous effect also comes with the very last words. Put differently, the final words of satirical headlines often serve as a "micro-punchline".[3]

5 Semantic analysis of aligned corpus

After characterizing aligned pairs syntactically, we now move to the semantic level. We first analyze the aligned pairs obtained from *Unfun.me* and later discuss our findings in the broader context of established theories of humor (Sec. 7).

Example. Before a more general analysis, let us first consider again our running example (Sec. 1), *God diagnosed with bipolar disorder*. This satirical headline works by blending two realms that are fundamentally opposed—the human and the divine—by talking about God as a human. Although the literally described situation is impossible (God is perfect and cannot possibly have a disease), the line still makes sense by expressing a crucial commonality between bipolar humans and God, namely that both may act unpredictably. But for humans, being unpredictable (due to bipolarity) is a sign of imperfection, whereas for God it is a sign of perfection ("The Lord moves in mysterious ways"), and it is this opposition that makes the line humorous.

The main advantage of our aligned corpus is that it lets us generalize this *ad-hoc* analysis of a particular example to a large and representative set of satirical headlines by pinpointing the essential, humor-carrying words in every headline: if the humor has been successfully removed from a headline h by altering certain words, then we know that these very words are key to making h funny.

This is especially true for single-substitution pairs; e.g., in the running example, *God* was replaced by *Bob Dylan* (a particular human), giving rise to the serious-sounding *Bob Dylan diagnosed with bipolar disorder*. The automatically extracted chunk pair {*God*, *Bob Dylan*} surfaces both the crucial commonality in the context of the headline (unpredictability) and the crucial opposition (God vs. human; unpredictability as a good vs. bad trait).

While the semantic analysis of original vs. substituted chunks may be difficult to automate, having access to explicit chunk pairs tremendously facilitates a large-scale human analysis. Conducting such an analysis revealed that the above pattern of a crucial commonality combined with a crucial opposition occurs in a large fraction of satirical headlines, and particularly in nearly all single-substitution pairs.

Script opposition. The crucial opposition has been called *script opposition* by humor theorists (cf. Sec. 7), and we henceforth adopt the same term. Inspecting all 254 single-substitution pairs, we found each pair to be in at least one of 6 oppositions, all representing "good"-vs.-"bad" dichotomies that are essential to the human condition, such as *high/low stature*, *life/death,* or *non-obscene/obscene.* All 6 oppositions, alongside examples, are listed in Fig. 4(a).

We manually labeled all pairs with their (sometimes multiple) oppositions and observe that most pairs (68%) feature an opposition of *high/low stature* (as in the running example), and surprisingly few pairs (7%), one of *non-obscene/obscene*. Due to its dominance, Fig. 4(a) further splits the *high/low stature* opposition into 10 subtypes.

Main mechanism: false analogy. Moving to a more formal analysis, we represent the running example schematically in Fig. 4(b), while Fig. 4(c) abstracts away from the example and depicts the generic template it implements, which may be verbalized as follows. The pair involves two entities, x (God) and x' (Bob Dylan), who share a crucial common property P (unpredictability), but whereas statement $P(x')$ ("Bob Dylan is unpredictable") could potentially entail the serious headline $H(x') = h'$ (*Bob Dylan diagnosed with bipolar disorder*), the analogous statement $P(x)$ ("God is unpredictable") cannot entail the analogous headline $H(x) = h$ (*God diagnosed with bipolar disorder*), for x and x' are cru-

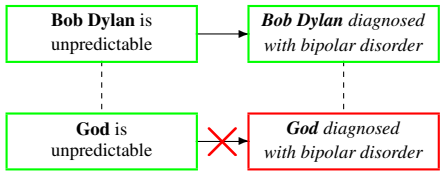
[2] We ascertained that the effect is not due to trailing chunks potentially being (1) longer and (2) more likely to be noun phrases.

[3] Strictly speaking, the findings of Sec. 4 only pertain to satirical headlines that are already similar to hypothetical serious headlines, due to selection bias (we only study headlines that players of *Unfun.me* chose to modify) and due to our focus on single-substitution pairs (about 50% of successful pairs).

(a) Script oppositions in single-substitution pairs (H3 has two substitutions), with percentage of pairs following each opposition. Examples show satirical headline $h = H(x)$ and a modified version $h' = H(x')$ (cf. diagram (c)); format: $\{x, x'\}$.

Script opposition	Perc.	Example pair
high/low stature	68%	
sublime/*mundane*	*15%*	H1. *Bush picks {laser,* ***rural****} background for presidential portrait*
success/*failure*	*14%*	H2. *Iraqis {arming,* ***preparing****} selves for independence*
authority/*no authority*	*13%*	H3. *Fort Knox receives {$85,* ***$85 million****} from {Cash4Gold,* ***Fed****}*
sophisticated/*simple*	*10%*	H4. *City opens new art {jail,* ***museum****}*
human/*object*	*6%*	H5. *{Local bar,* ***NFL star****} comes out as gay*
human/*animal*	*5%*	H6. *Hollywood mourns passing of {16th or 17th Lassie,* ***Robin Williams****}*
modern/*outdated*	*5%*	H7. *General Motors reports record sales of new {disposable,* ***eco****} car*
rich/*poor*	*4%*	H8. *Asian economic woes force layoffs of 700,000 {****pop stars****, workers}*
no religion/*religion*	*3%*	H9. *{God,* ***Bob Dylan****} diagnosed with bipolar disorder* (cf. diagram (b))
animal/*object*	*1%*	H10. *New delicious {****species****, fruit} discovered*
good/bad intentions	25%	H11. *BP ready to resume oil {spilling,* ***drilling****}*
reasonable/absurd response	17%	H12. *Conservation group condemns {waterboarding,* ***baths****} as wasteful*
no violence/violence	10%	H13. *Russian officials promise low {death,* ***highway****} toll for Olympics*
life/death	9%	H14. *Cancer victim given second chance at {death,* ***life****}*
non-obscene/obscene	7%	H15. *Tiger Woods announces return to {sex,* ***golf****}*

(b) Example of false-analogy headline.



(c) Abstract false-analogy template.

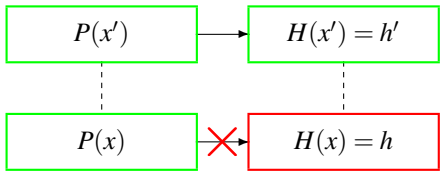


Figure 4: **(a)** Script oppositions and examples. **(b)** Example of false-analogy headline. **(c)** Abstract false-analogy template.

cially opposed via one of the script oppositions of Fig. 4(a) (*religion/no religion;* or, *God,* for whom unpredictability is a sign of perfection, vs. *humans,* for whom it is a sign of imperfection). Hence, we call this mechanism *false analogy.*

As the examples of Fig. 4(a) show, the analogy is never marked lexically via words such as *like;* rather, it is evoked implicitly, e.g., by blending the two realms of human psychiatry and biblical lore into a single headline. Only the satirical headline $H(x)$ itself (red box in Fig. 4(c)) is explicit to the reader, whereas x' and P (and thus all the other 3 boxes) need to be inferred. A main advantage of our method is that it also makes x' explicit and thereby facilitates inferring P and thus the semantic structure that induces humor (as in Fig. 4(b)).

We emphasize that the script opposition that invalidates the logical step from $P(x)$ to $H(x)$ is not arbitrary, but must be along certain dimensions essential to human existence and contrasting "good" vs. "bad" (Fig. 4(a)). Interestingly, in typical jokes, the "good" side is explicit and the "bad" side must be inferred, whereas in satirical headlines, either the "good" or the "bad" side may be explicit. And indeed, as shown by the examples of Fig. 4(a) (where the "good" side is marked in bold), satirical headlines differ from typical jokes in that they tend to make the "bad" side explicit.

Single vs. multiple edit operations. A large fraction of all headlines from *The Onion*—and an overwhelming fraction of those in single-substitution pairs—can be analyzed with the false-analogy template (and we indeed encourage the reader to apply it to the examples of Fig. 4(a)). Additionally, many of the pairs with two substitutions also follow this template. H3 in Fig. 4(a), which plays on the opposition of the Federal Reserve being a serious institution vs. Cash4Gold being a dubious enterprise exploiting its customers, exemplifies how, whenever multiple substitutions are applied, they all need to follow the same opposition (e.g., Fed : Cash4Gold = $85 million : $85 = serious : dubious).

6 Related work

The most widely accepted theory of verbal humor is the so-called *General Theory of Verbal Humor* by Attardo and Raskin (1991), an extension of Raskin's (1985) *Semantic-Script Theory of Humor,* which we summarize when discussing our findings in its context in Sec. 7.

Much follow-up work has built on these theories; see the excellent primer edited by Raskin (2008). Here, we focus on contributions from computer science, where most work has been on the **detection of humor** in various forms, e.g., irony (Reyes, Rosso, and Veale 2013; Wallace, Choe, and Charniak 2015), sarcasm (Davidov, Tsur, and Rappoport 2010; González-Ibánez, Muresan, and Wacholder 2011), and satire (Burfoot and Baldwin 2009; Goldwasser and Zhang 2016), sometimes with the goal of deciding which of two texts is funnier (Shahaf, Horvitz, and Mankoff 2015). These works use documents or sentences as the smallest unit of analysis, whereas we operate at a finer granularity, analyzing the very words causing the switch from serious to funny.

Another cluster of work has considered the **generation of humor,** mostly via fixed templates such as acronyms (Stock and Strapparava 2006), puns (Binsted and Ritchie 1997; Ritchie et al. 2007), two-liners (Labutov and Lipson 2012), or cross-reference ambiguity (Tinholt and Nijholt 2007).

Finally, our work also relates to efforts of **constructing humor corpora** (Filatova 2012; Khodak, Saunshi, and Vodrahalli 2018). Here, too, we increase the granularity by actively generating new data, rather than compiling humorous texts that have already been produced. Crucially, ours is a corpus of aligned pairs, rather than individual texts, which enables entirely novel analyses that were infeasible before.

7 Discussion and future work

Summary of findings. Comparing satirical to similar-but-serious-looking headlines within the pairs collected via *Unfun.me* reveals that the humor tends to reside in the final words of satirical headlines, and particularly in noun phrases. In order to remove the humor, players overwhelmingly replace one phrase with another; rarely do they delete phrases, and nearly never introduce new phrases. Reversing the direction of the editing process, this implies that the most straightforward way of producing satire from a serious head-

line is to replace a trailing noun phrase with another noun phrase.

One may, however, not just replace any noun phrase with any other noun phrase; rather, the corresponding scripts need to be opposed along one of a few dimensions essential to the human condition and typically pitting "good" vs. "bad". Also, the two opposing scripts need to be connected via certain subtle mechanisms, and we pointed out false analogy as one prominent mechanism. These findings echo the predictions made by the prevailing theory of humor. We now summarize this theory and discuss our results in its context.

Relation to Semantic-Script Theory of Humor. As mentioned (Sec. 6), the most influential theory of verbal humor has been Raskin's (1985) *Semantic-Script Theory of Humor,* which posits a twofold necessary condition for humorous text: (1) the text must be compatible with two different *semantic scripts* (simply put, a semantic script is a concept together with its commonsense links to other concepts); and (2) the two scripts must be opposed to each other along one of a small number of dimensions.

The second criterion is key: the mere existence of two parallel compatible scripts is insufficient for humor, since this is also the case in plain, non-humorous ambiguity. Rather, one of the two scripts must be possible, the other, impossible; one, normal, the other, abnormal; or one, actual, the other, non-actual. These oppositions are abstract, and Raskin (1985, p. 127) gives several more concrete classes of opposition, which closely mirror the dimensions we empirically find in our aligned pairs (Fig. 4(a)). Our results thus confirm the theory empirically. But the advantages of our methodology go beyond, by letting us quantify the prevalence of each opposition. In addition to the concrete oppositions of Fig. 4(a), we also counted how pairs distribute over the above 3 abstract oppositions, finding that most satirical headlines are of type *possible/impossible* (64%), followed by *normal/abnormal* (28%), and finally *actual/non-actual* (8%).

In typical jokes, one of the two scripts (the so-called *bona fide* interpretation) seems more likely given the text, so it is in the foreground of attention. But in the punchline it becomes clear that the *bona fide* interpretation cannot be true, causing initial confusion in the audience, followed by a search for a more appropriate interpretation, and finally surprise or relief when the actually intended, non–*bona fide* script is discovered. To enable this process on the recipient side, the theory posits that the two scripts be connected in specific ways, via the so-called **logical mechanism,** which resolves the tension between the two opposed scripts.

Attardo (2001, p. 27) gives a comprehensive list of 27 logical mechanisms. While our analysis (Sec. 5) revealed that one mechanism—*false analogy*—dominates in satirical headlines, several others also occur: e.g., in *figure–ground reversal,* the real problem (the "figure") is left implicit, while an unimportant side effect (the "ground") moves into the focus of attention (e.g., H12 in Fig. 4(a): waterboarding, like baths, does waste water, but the real problem is ethical, not ecological). Another common mechanism—*cratylism*—plays with the assumption prevalent in puns that phonetic implies semantic similarity (e.g., H11 in Fig. 4(a)).

Satire is a form of art, and the examples just cited highlight that it is often the creative combination of several mechanisms that makes a headline truly funny. Beyond the bare mechanism, the precise wording matters, too: e.g., either *16th Lassie* or *17th Lassie* would suffice to make H6 in Fig. 4(a) funny, but the combination *16th or 17th Lassie* is wittier, as it implies not only that Lassie has been played by many dogs, but also that people do not care about them, thus reinforcing the *human/animal* opposition.

We conclude that, while satirical headlines—as opposed to typical jokes—offer little space for complex narratives, they still behave according to theories of humor. Our contributions, however, go beyond validating these theories: the aligned corpus lets us quantify the prevalence of syntactic and semantic effects at play and reveals that the dominant logical mechanism in satirical headlines is false analogy.

Satirical-headline generation. This points to a way of generating satirical headlines by implementing the false-analogy template of Fig. 4(c): pick an entity x (e.g., Pepsi) and a central property $P(x)$ of x (e.g., "Pepsi is a popular drink"); then pick another entity x' for which $P(x')$ also holds, but which is opposed to x along one of the axes of Fig. 4(a) (e.g., Bordeaux wine, which is in a *high/low stature* [*sublime/mundane*] opposition to Pepsi); and finally generate a headline $H(x')$ based on $P(x')$ (e.g., *2018 Bordeaux vintage benefits from outstanding grape harvest*) which cannot be seriously formulated for x instead x', due to the opposition, yielding the satirical $H(x)$ (e.g., *2018 Pepsi vintage benefits from outstanding high-fructose corn harvest,* where we analogously replaced *grape* with *high-fructose corn,* cf. Sec. 5). The subtitle of the present paper was also generated this way.

Most humans are unaware of the logical templates underlying satire, while machines have difficulties finding entity pairs opposed in specific ways and formulating pithy headline text. We hence see promise in a hybrid system for coupling the respective strengths of humans and machines, where the machine guides the human through the template instantiation process while relying on the human for operations such as finding appropriate entities for substitution etc.

Human perception of satirical vs. serious news. Recall that in task 2 (Sec. 2), players also rate unmodified satirical and serious headlines g with respect to how likely they consider them to be serious. Table 1 shows that, although players are generally good at distinguishing satire from real news, they do make mistakes: 10% of serious headlines are consistently misclassified as satirical (e.g., *Schlitz returns, drums up nostalgic drinkers*), and 8% of satirical headlines, as serious (e.g., *Baltimore looking for safer city to host Super Bowl parade*). Studying these misunderstood headlines can yield interesting insights into how readers process news, especially in an age where "fake news" is becoming a ubiquitous scourge. We leave this analysis for future work.

Beyond humor. The mechanism underlying *Unfun.me* defines a general procedure for identifying the essential portion of a text that causes the text to have a certain property.

In our case, this property is humor, but when asking players instead to remove the rudeness, sexism, euphemism, hyperbole, etc., from a given piece of text, we obtain a scalable way of collecting fine-grained supervised examples for better understanding these ways of speaking linguistically.

8 Conclusion

Humor is key to human cognition and holds questions and promise for advancing artificial intelligence. We focus on the humorous genre of satirical news headlines and present *Unfun.me*, an online game for collecting pairs of satirical and similar-but-serious-looking headlines, which precisely reveal the humor-carrying words and the semantic structure in satirical news headlines. We hope that future work will build on these initial results, as well as on the dataset that we publish with this paper (West and Horvitz 2019), in order to make further progress on understanding satire and, more generally, the role of humor in intelligence.

References

Andrist, S.; Bohus, D.; Yu, Z.; and Horvitz, E. 2016. Are you messing with me? Querying about the sincerity of interactions in the open world. In *Proc. ACM/IEEE International Conference on Human–Robot Interaction (HRI)*.

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):293–348.

Attardo, S. 2001. *Humorous Texts: A Semantic and Pragmatic Analysis*. Walter de Gruyter.

Berger, A. L.; Pietra, V. J. D.; and Pietra, S. A. D. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics* 22(1):39–71.

Binsted, K., and Ritchie, G. 1997. Computational rules for generating punning riddles. *HUMOR: International Journal of Humor Research* 10(1):25–76.

Binsted, K.; Nijholt, A.; Stock, O.; Strapparava, C.; Ritchie, G.; Manurung, R.; Pain, H.; Waller, A.; and O'Mara, D. 2006. Computational humor. *IEEE Intelligent Systems* 21(2):59–69.

Burfoot, C., and Baldwin, T. 2009. Automatic satire detection: Are you having a laugh? In *Proc. Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL–IJCNLP)*.

Davidov, D.; Tsur, O.; and Rappoport, A. 2010. Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In *Proc. Conference on Computational Natural Language Learning (CoNLL)*.

Filatova, E. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *Proc. Language Resources and Evaluation Conference (LREC)*.

Glass, I. 2008. Tough room. Act one: Make 'em laff. *This American Life* 348. `https://www.thisamericanlife.org/348/tough-room/act-one`.

Gneiting, T., and Raftery, A. E. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102(477):359–378.

Goldwasser, D., and Zhang, X. 2016. Understanding satirical articles using common-sense. *Transactions of the Association of Computational Linguistics* 4(1):537–549.

González-Ibánez, R.; Muresan, S.; and Wacholder, N. 2011. Identifying sarcasm in Twitter: A closer look. In *Proc. Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL–HLT)*.

Khodak, M.; Saunshi, N.; and Vodrahalli, K. 2018. A large self-annotated corpus for sarcasm. In *Proc. Language Resources and Evaluation Conference (LREC)*.

Labutov, I., and Lipson, H. 2012. Humor as circuits in semantic networks. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*.

Martin, R. A. 2010. *The psychology of humor: An integrative approach*. Elsevier.

Mihalcea, R., and Pulman, S. 2007. Characterizing humour: An exploration of features in humorous texts. In *Proc. International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*.

Navarro, G. 2001. A guided tour to approximate string matching. *ACM Computing Surveys* 33(1):31–88.

Raskin, V. 1985. *Semantic Mechanisms of Humor*. Reidel.

Raskin, V., ed. 2008. *The Primer of Humor Research*. Mouton de Gruyter.

Reyes, A.; Rosso, P.; and Veale, T. 2013. A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation* 47(1):239–268.

Ritchie, G.; Manurung, R.; Pain, H.; Waller, A.; Black, R.; and O'Mara, D. 2007. A practical application of computational humour. In *Proc. International Joint Conference on Computational Creativity (ICCC)*.

Shahaf, D.; Horvitz, E.; and Mankoff, R. 2015. Inside jokes: Identifying humorous cartoon captions. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Stock, O., and Strapparava, C. 2006. Laughing with HAHAcronym, a computational humor system. In *Proc. National Conference on Artificial Intelligence (AAAI)*.

Tinholt, H. W., and Nijholt, A. 2007. Computational humour: Utilizing cross-reference ambiguity for conversational jokes. In *Proc. International Workshop on Fuzzy Logic and Applications (WILF)*.

von Ahn, L., and Dabbish, L. 2008. Designing games with a purpose. *Communications of the ACM* 51(8):58–67.

Wallace, B. C.; Choe, D. K.; and Charniak, E. 2015. Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment. In *Proc. Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL–IJCNLP)*.

West, R., and Horvitz, E. 2019. Github project repository. `https://github.com/epfl-dlab/unfun`.