Great Expectations and EPIC Fails:

A Computational Perspective on Irony and Sarcasm

Tony Veale,
School of Computer Science, University College Dublin, Ireland.

Abstract

Statistical models do an excellent job of capturing the natural rhythms and flows of language, especially if trained on large corpora of usage data. Although they concern themselves with surface forms, not meanings, language models that are trained at web-scale can still assimilate a great deal of native intuition and world knowledge into their statistics. However, phenomena such as irony and sarcasm need more than superficial fluency: they also require intent. Users of irony must immerse themselves in the rhythms of a language whilst also directing its flow from outside, so as to both use *and* abuse the expectations of their audience. This chapter considers the demands that irony and sarcasm place on an artificial user of language, whether to grasp the pragmatic insincerity of human speakers, or to playfully speak with a forked tongue of its own. To consider the relative merits of alternate approaches, we must first operationalize irony in algorithmic terms, and define an objective measure of communicative success for ironic machines.

1. Does Not Compute, Literally

The tensions between the creativity of the individual and the expectations of an audience have rarely been so astutely captured as in this quip from Woody Allen: "I want to forge in the smithy of my soul the uncreated conscience of my race. And then see if I can get them mass-produced in plastic." Creativity is a personal attempt to package one's subjective experiences in a form that acknowledges the expectations of others, and Allen opts for the two-act package of a standard joke: the first is a borrowing, from Joyce's *A Portrait of the Artist as a Young Man*, with a noble if grandiose sentiment, while the second subverts the first with its ignoble desire for financial success. The pairing amplifies the tension in Joyce's metaphor, but tips it from the worthy and sublime into the unworthy and ridiculous, so as to also move us from pathos into bathos. This tells us something of the close kinship between metaphors and jokes, and also something of the popular view of the role of machines in human creativity. Nonetheless, in addition to telegraphing a note of skepticism about machine generation at industrial scale, Allen's joke highlights a triad of qualities that scholars associate with ironic language and thought.

The first of these is *echoic mention* (Sperber and Wilson, 1981; Sperber, 1984; Kreuz and Glucksberg, 1995). Allen's mention of a much-quoted line from Joyce's novel of artistic striving evokes a web of expectations in the reader: that Allen is a Joycean, that he shares the same profound aims (or flatters himself that he does), and that his own strivings can be compared to those of the novel's protagonist, Stephen Dedelus. An explicit echoing of content will also implicitly echo its initial context of use, and so revive those expectations that were active in its first outing. Allen now sets about dashing those grandiloquent expectations via opposition, another quality that scholars have long associated with irony (Grice, 1975:53; Garmendia, 2018:17-41). Here the opposition works at multiple levels, the most striking being the semantic variety that anchors script-based theories of jokes (Raskin, 1985). We see a personal mission statement – one that overtly mentions the "soul" – contrasted with a soulless desire for impersonal profits, while the "plastic" of the second script nudges the "forge" of the first into its other sense, *to fake*. But Allen's add-on is not a natural continuation of Joyce's line, and the clash of literary styles yields what is called "register" humor (Attardo, 2009). This tonal

inconsistency suggests an additional contrast of personalities, and it is here that the third hallmark of irony is most apparent: playful pretense (Clark and Gerrig, 1984; Kumon-Nakamura, Glucksberg and Brown, 1995). Allen's pretense works on multiple levels: that his ignoble aims are compatible with those of Dedalus; that he is injudicious enough to not spy a contradiction; and that this take on his goals is indeed accurate, and not just a playful overstatement of his true beliefs.

Despite the best efforts of irony theorists to tease apart these three stances, on echoing, opposition and pretense, Allen's quip shows how entangled they remain. Pretense involves a disguise of sorts, so we often wrap our views in echoes of the propositional content, or tacit propositional attitudes, of another. Ironic speakers are chameleons that *want* to be seen, as they echo that which they pretend to be. So opposition, in the form of an incongruity between a proposition and its context of use, or between a proposition's content and the attitude that accompanies it, or between the internal parts of the same proposition, is key to separating sincere from ironic uses of echo and pretense. Allen's quoting of Joyce in a context of ugly commercialism yields incongruity of the first kind; sarcasm relies on incongruity of the second kind, in quips such as "Don't worry, we can walk to the curb from here" in a car that is badly parked (e.g., in Allen's film *Annie Hall*); while mocking wit is most evident via internal incongruity of the third kind, as in this line from a woman to a married suitor in a singles bar: "Does your wife know you're single?" Opposition can make an echoic mention or a verbal imposture suspect enough for audiences to scrutinize further, so that it gives up its ironic meaning, while scare-quotes and other marking devices can further add to our suspicions.

If scholars face difficulties in unpicking these interwoven strands, how might a computer fare? In addition to incongruity detection and resolution, a simple quip like Allen's implicates a host of other information-processing needs, from analogy (across propositions and contexts, so e.g., *Allen = Dedalus* and *soul = plastic*) to sentiment analysis (so e.g., Joyce's and Allen's lines are seen to have contrasting emotional valences) to continuity checking (so e.g., Allen's punch line is seen as a stylistic departure from Joyce's setup) to counterfactual reasoning (so that Allen's pretense can be grasped), to say nothing of semantic-level analysis, word-sense disambiguation (e.g., of "forge"), metaphor detection and analysis. Each of these

needs has been extensively studied in the computational literature, and a host of approaches and representations – from symbolic scripts to conceptual graphs to statistical vector spaces and language models – can be used to address them. Each kind of representation gives us a different correlate of incongruity for a machine to look for, reason over, or perhaps even generate in an ironic text of its own. In fact, a machine need not be explicitly programmed to look for one kind or another to appreciate the role of incongruity in irony. In a system that learns to apply the appropriate label to a given text (such as *ironic*), any cluster of features that tends to improve its accuracy will be automatically sifted, weighted and rewarded.

So, while scholars are divided as to the primacy of pretense, echo or opposition in irony, the debate among computationalists is more one-sided. Philosophers of an anti-AI bent have long argued that machines can only truly grasp forms, not the intents that animate them (Searle, 1980), and whatever the truth of this claim, computational models of irony and sarcasm place a singular focus on form. By deconstructing the formal properties of an echo, a machine can "take care of the signifiers and let the signifieds look after themselves" (Chandler, 2002:199). To the extent that pretense is invoked at all, it motivates the insincere use of an echoic mention. For instance, machines detect sarcasm in online posts by looking for formal markers of intent, from explicit emoticons, hashtags and scare-quotes to sentence-internal displays of affective incongruity (e.g., of the form "I love it when <*negative_act*>") to other calcified constructions that mark out a formulaic use of irony or sarcasm. If a provoking context for a given text is available – for instance, the prior post to which the current post is a riposte – then its formal properties can also be considered by a detector. Recent posts by a text's author can additionally be analyzed to provide a psychological context for detection. A distillation of recent posts can indicate, via sentiment analysis, whether an author shows an aggressive attitude or an upbeat mood (see Tausczik and Pennebaker, 2009), and these factors can then shape a detector's view of the author's intent.

As such, a computational model of irony, whether for detection or generation, is not a theory of irony given algorithmic form. For example, a theoretical account must illuminate how irony is both generated and detected, while a computational model usually knows just enough to do one or the other, but not both. Still, we

can learn something about how speakers use, detect and appreciate irony in the evaluation of even simple models. So, while we have sound practical reasons for imbuing machines with an ear for pretense and a flair for linguistic creativity – from richer expressiveness to higher emotional intelligence to a greater tolerance for error – our principal focus here is on what computational models can tell us about humans. As various computational models are surveyed in the sections to come, it will become clear that humans remain a key fixture in the algorithmic loop when it comes to irony and sarcasm, not least because it is the human use and appreciation of these phenomena that we set out to model. After all, we have little or no need for our machines to use these devices when talking to each other.

The next section explores what we mean by incongruity in a creative context. This is a quality that we can model in explicit terms, or allow a detection system to acquire for itself from a sufficiently rich representation of a text. We shall see examples of each approach in section 3, in a brief survey of computational models for detecting sarcasm and irony. Such models rarely need an explicit definition of what it is they are looking for, insofar as their success is measured relative to how well they accord with human judgment. This allows a detector's conception of the phenomenon to be just as nebulous as our folk notion of "*I know it when I see it.*" However, for a system that generates rather than detects, we must give a machine an explicit account of what it is trying to evoke in the minds of its audience. So, in section 4, a simple but explicit model of expectation is presented. Called EPIC, the model captures the property-level expectations that can get upended by an ironic utterance. This allows us to propose a quantifiable view of what it means for such an utterance to achieve its desired effect on an audience. Section 5 then uses this operationalization to quantify the relative value of different framing strategies for machine-generated irony. These strategies range from overt markings to affective incongruity to the explicit echoing of propositional content. In closing, some final thoughts on the viability of a machine sense of irony are then offered in section 6.

2. Tales of the Unexpected

Incongruity arises from a conflict between what we observe and what we expect, and can spur us to revisit our expectations. This clash can take multiple forms. An

easily-controlled form, affective incongruity, arises from a contrast of sentiments, either within a text itself or between a text and its context of use. So, for instance, metaphorical tension is easily stoked by yoking two ideas of opposing sentiment, as when we compare a promise to a prison, love to madness, or religion to a virus. The affective connotation of a word – the degree to which we associate it with a positive or a negative feeling – is a rather shallow basis for signaling a mismatch of ideas, but it may still be enough to alert an audience to our ironic intent. This is a special, and conveniently marked, case of dissociative incongruity, in which a speaker mingles words or ideas that one rarely encounters in the same setting. In semantic approaches to opposition, as in that of Raskin (1985), such incongruity is sparked by the juxtaposition of discrete symbols from different frames that are stipulated to be opposites, such as *life* vs. *death* or *wise* vs. *stupid*. A more diffuse view is offered by statistical models that construct a probability distribution over the words that co-occur with a given term, such as *religion* and *virus*. To estimate the conceptual gulf between two terms, we then quantify the divergence between the probability distributions associated with those terms (Kao *et al.*, 2017).

The most nuanced computational account of expectation in language is offered by statistical language models. Human language is an infinite resource that allows speakers to construct an endless variety of meaningful utterances from a finite vocabulary and grammar, and a language model seeks to distil this infinite reach into a finite series of observations. Such models do not draw a hard line between valid and invalid utterances, but assign a probability of acceptance to any given string. In a generative system, a language model can be used to weigh the relative acceptability of competing formulations of the same meaning, but it can just as easily be used to suggest likely continuations for a given prompt. A model-specific measure, called *perplexity*, also allows us to quantify a shift in predictability as we transition from one kind of text to another, such as from Joyce's literary styling to Allen's informal phrasing. Utterances that are witty in themselves tend to exhibit higher perplexity than their plainer counterparts (Reyes *et al*., 2013), while those that work as witty extensions to other stimuli (such as Allen's quip, or a cartoon caption) tend toward higher familiarity and lower perplexity (Shahaf *et al*., 2015). In the latter cases, the incongruity arises *between* stimuli, not within a single one.

In fact, jokes rely on this gulf between an apparent incongruity and the norm that it distorts. But if jokes can make the familiar seem strange, interpretation re-anchors the strange in the familiar, by recovering the norm so as to quantify its distortion. Since statistical language models capture the regularities of linguistic expression, they can be powerful tools for the analysis and generation of *optimal innovations* that wittily deviate from a norm in small but meaningful ways (Giora *et al*., 2004; Giora, 2018). In a pun, for instance, our aim is to recover the original form of an utterance before a phonetic substitution was made, as in this example: "The dentist was weary after a hard day at the orifice." We can expect a language model to assign a very small probability to the phrase "a hard day at the orifice," but a much higher one to the idiom "a hard day at the office," and it is this shift to the familiar, guided by the phonetic similarity of "orifice" to "office", that allows us to recover the distorted norm. Likewise, for that ironic quip in the singles bar, we expect our model to assign a low probability to "single" following "Does your wife know you're", and a far higher probability to "here." The irony is resolved if "single" is understood as a shorthand that condenses the allegation "Does your wife know you're *here, pretending to be single*?" In this view, the understanding of irony is a generative process that goes well beyond mere detection: it requires language users to generate, and then test, hypotheses as to what it might mean.

Statistical language models acquire a syntagmatic understanding of the words in their training data. Given a prompt like "a knight in shining __", a model can confidently predict the next word to be "armour", and grant only the smallest of likelihoods to "armpits." So, a model that conditions its word probabilities on an extensive prior context can weave stories from the slightest of prompts (Radford *et al*., 2019), but even large, context-attuned models will still treat meaning as a latent variable. Thus, even one that grasps the idiomaticity of "a knight in shining armour" will still fail to appreciate the symbolism of this specific arrangement of signs. Nonetheless, if interrogated in the right way, syntagmatic knowledge can yield semantic insights. Consider the pattern of the "as-as" simile: *<topic> is as <adj> as <vehicle>.* When *<vehicle>* is "a knight in shining armour" or "a knight on a white horse," what are the most likely values of *<adj>*? We expect qualities like *brave*, *chivalrous* and *heroic* to be highly probable, and qualities like *weak*, *wicked* and *dishonorable* to be dismissed as much less likely. In contrast, this situation is

reversed for insincere uses of the pattern, since an attested value of <*adj*> in an ironic simile has little chance of filling this slot in its non-ironic variant. If similes wear their expectations on their sleeves, this is where we should look for them.

In short, sincere similes reveal our expectations of the world while insincere similes dash those expectations for creative ends. By looking specifically to a rich source of similes, a machine can acquire the expectations that underpin our use of words, and also learn how to ironically dash those expectations for itself. So, what is needed is a means of harvesting expectation-laden similes on a vast scale, as well as the means of discerning sincere comparisons – such as "as sharp as a scalpel" – from insincere ironic ones – such as "as sharp as a bowling ball." We return to those means, and their role in supporting the EPIC model, in section 4.

3. Detecting Sarcasm and Irony in Texts

Detection is essentially a labeling task, and is, as such, a highly reductive process. An ironic signal is not detected directly, but rather inferred on the basis of other, more discernible or quantifiable features of a text, such as sentiment, perplexity, lexical and structural ambiguity, or anything else that implies a text is not what it seems to be on the surface. Since sincerity is a matter of intent, machines instead set out to quantify a text's amenability to an insincere reading. This is where the fine academic distinction between irony and sarcasm becomes an arbitrary line: a machine that learns to label will use any features that improve its performance – that is, its agreement with human gold-standard labels for the same texts – but some features will naturally offer more purchase on sarcasm than irony.

In their SASI system, Tsur, Davidov and Rappoport (2010) focused on sarcasm in online product reviews. E-commerce sites that solicit customer feedback make a compelling case for granting machines a grasp of playful insincerity, since the commercial value of reviews resides in their trustworthiness. Surface sentiment is often misleading when reviewers set out to be ironic or sarcastic, but a ground truth may be offered through a different channel. SASI looks at reviews that have both a star rating and a textual content, and seeks out those for which the rating runs counter to the sentiment of the text. From those that have been labeled as sarcastic, SASI extracts word sequences for which a formulaic template can be

generalized. To build its templates, SASI replaces high-frequency function words (HFW), low-frequency content words (CW) and proper names (PN) with a place holder, and assesses how often the template matches sarcastic *and* non-sarcastic reviews. Only those that are frequent and discerning are used as features by the detector. A candidate text is characterized by how well it matches each template, and the results yield a fingerprint that is matched to gold-standard exemplars. A candidate is labeled "sarcastic" if its fingerprint finds more compelling matches among SASI's inventory of sarcastic examples than among its non-sarcastic stock, after weighting any matches to account for the relative imbalance of these stocks.

If sarcasm is a loud sigh and irony a sly wink, then certain easily quantifiable aspects of a text lend themselves more to the former than the latter. For instance, a text that is incongruous in itself, either because it conveys mixed emotions or it subverts its own expectations, is a dramatic gesture that demands our attention. Riloff *et al*. (2013) attune their sarcasm detector to the contrast of sentiment and circumstance that is the hallmark of a sarcastic response. By harvesting tweets that are explicitly tagged with the marker #*sarcasm* – the Twitter equivalent of a dramatic sigh – and which introduce a situation with a positive sentiment (e.g., "I love it when"), a machine can acquire examples of those states of affairs that tend to provoke a sarcastic reply (e.g., "friends forget to call me"). By looking for other overtly sarcastic tweets that allude to similar situations, the detector broadens its understanding of the feelings they inspire, before using these new insights to enrich its store of exemplars, and so on, in a cyclical bootstrapping process. Once it models how negative situations pair with positive emotions in sarcastic tweets, a detector can then perceive insincerity in tweets that lack the marker #*sarcasm*.

Reyes, Rosso and Veale (2013) tackle irony detection within the wider context of wit identification. To separate clever one-liners from mundane sentences, they extract a host of weakly discriminating features from each text, so that its wit can be assessed with a multivariate model. In addition to measures of perplexity and ambiguity, both lexical and syntactic, gapped sequences (or *skip-grams*) are used to acquire the most formulaic constructions (e.g., "a fine <*gap*> indeed"), while a psycholinguistic lexicon offers scores for word sentiment (e.g., *love* = 3.0, *hate* = 1.2), arousal (e.g., *enjoy* = 2.2, *avoid* = 1.7) and imageability (e.g., *snow* = 3, *doubt*

= 1.0). Generalizing skip-grams with sentiment scores then produces templates with affective wildcards, such in "thanks for the *<negative>*." Additional features include *pointedness* (scoring the use of capitalized words and attention-grabbing punctuation), *counterfactuality* (scoring the use of words like "however" or "yet") and *temporal imbalance* (scoring the use of mixed tenses in the same short text). Since irony is far less frequent than non-irony in a typical text, Reyes, Rosso and Veale trained and tested their detector on both balanced (50:50) and unbalanced (30:70) splits of the ironic and non-ironic data sets. While the detector achieved .7 to .75 accuracy on balanced splits, it achieved .75 to .8 on unbalanced splits.

One of the first neural-network models of sarcasm detection was proposed by Ghosh and Veale (2016). Their network architecture is a now-typical assemblage of standard layer types, mapping from a numeric, vector-space embedding of the input (a tweet) to a binary decision as to whether this text is sarcastic or not. In an approach they view as "fracking" for sarcasm, the network extracts and then successively generalizes over those features of the input that help it to make the right call. The gold standard tweets on which the model is trained are harvested using a variety of overt markings as retrieval cues, and then labeled by human raters as to their sarcastic intent. Yet, as noted in Ghosh *et al*. (2015), this intent is often difficult to discern after the fact by independent judges, especially when the context of the tweet is no longer apparent. For instance, some tweets contain enough internal incongruity for the hashtag #*yeahright* to convey sarcasm, while for others is it merely suggestive of exasperation. To remedy this lack, Ghosh and Veale (2017) augmented their network with two new sources of context for each input: the tweet to which the current text is a reply (that provocation is likewise mapped into a numeric vector); and a psychological analysis of the author's most recent tweets, to characterize their mood on 11 dimensions, ranging from upbeat and personable to angry and arrogant (Tausczik and Pennebaker, 2009). In this way, the network can grasp more of a tweet's context that its human raters can.

To remedy this disparity, Ghosh and Veale introduced another innovation, by asking the author of each text in their training and test sets to annotate their own tweets for sarcasm. In this new setup, an automated bot would coarsely pre-filter tweets from certain "magnets" for sarcasm – politicians, comics and other public

figures that use, or attract, sarcasm – before contacting their authors directly, to solicit a *yay* or a *nay* as to their sarcastic intent. The result is a data set in which the authors themselves have interrogated their own intentions to label each text.

The augmented network contains additional long/short-term memory (LSTM) layers to find relevant, if incongruous, connections between different parts of the input representation, or between this input and its two new sources of context. The general architecture of the neural network model is illustrated in Figure 1.

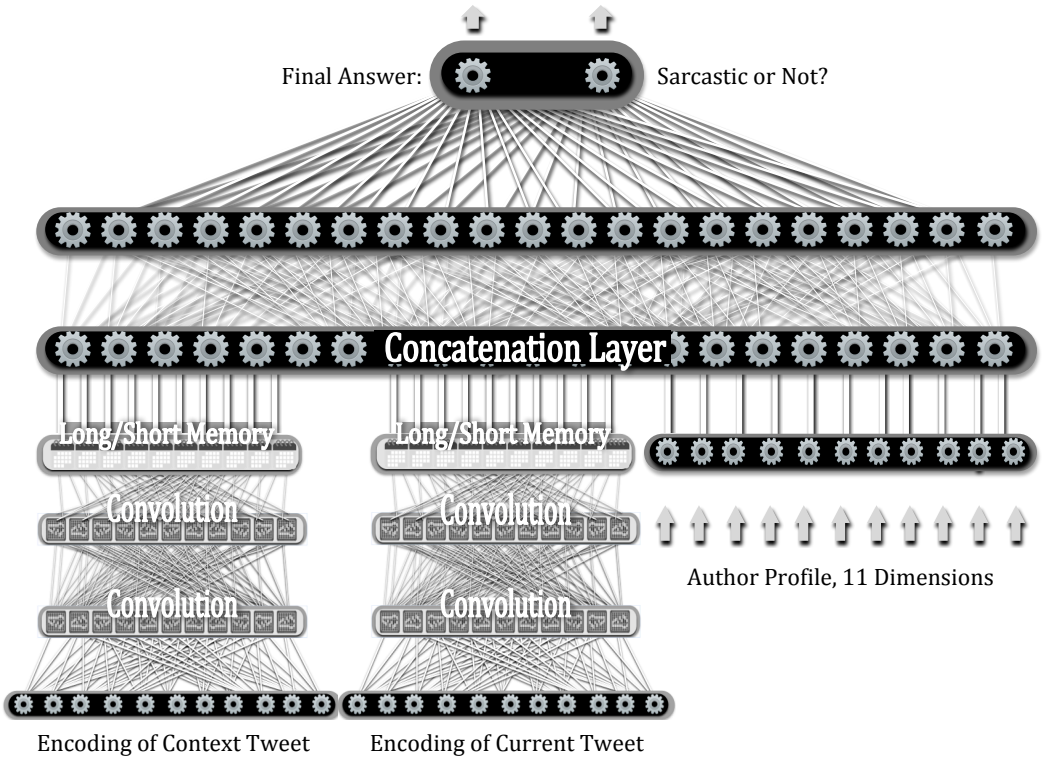

Figure 1. A neural model for sarcasm detection (Ghosh and Veale, 2017).

When a network is given the prior set-up to a potentially sarcastic tweet, it is better able to discern its author's intent, as it can now find something akin to an incitement for an insincere response. By also giving it a coarse sense of who that author is, psychologically speaking, the network can motivate their sarcasm as an extension of their current mood. Thus, depending on the precise configuration of the network – these architectures encourage a great deal of fiddling – a prior context tweet is worth an extra 6% or so of additional accuracy. The recent mood of the author, as evident in their recent tweets, also adds about 6% in combined precision and recall, again depending on the network configuration. Nonetheless,

the benefits of each kind of context are not additive. When a network is given the prior context tweet *and* a profile of the author, one sees much the same bump in performance as when it is given just one or the other. So, while the detector can get what it needs from either context source, it does not seem to need them both.

Neural models are designed to generalize over the most relevant properties of their input representations. The richer the representation, the more nuanced the generalizations it can support. Sarcasm lacks subtlety when compared to irony, yet each requires a text representation that marries fine semantic and pragmatic distinctions to the contingent facts of the world. Consider the simile "as welcome as a CNN reporter at a Trump rally." Only usage data that captures the fractious relationship between the parties can allow a detector to sense the incongruity of the pairing, and label it as *ironic* or *sarcastic* as the task demands (and not apply the same label to "as welcome as a FOX reporter at a Trump rally"). This need to echo the external world pushes our computational models toward an ever more web-driven, data-intensive and context-aware view of word meaning (Potamias, Siolas and Stafylopatis, 2020). However, a task-specific corpus that is annotated for sarcasm or irony cannot reflect all of the contingencies and tensions on which creative insincerity hinges, so it is vital that those nuances are already baked into the word representations that shape the inputs to the model. A possible solution is "*transfer learning*," in which word representations are acquired at web-scale to solve a more generic problem, such as gap-filling in cloze tests, and then later re-deployed in the service of tasks such as sarcasm detection (Zhang *et al*., 2018).

4. What To Expect When You're Expecting

What gets baked into these representations are the many expectations that guide our use of words across contexts. Consider, for example, the expectations evoked by "chair" in "I bought a chair," and how thoroughly they are dashed in the joke "I bought a chair for my mother-in-law, but she refuses to plug it in." Sarcasm and irony, or wit more generally, exploit our shared expectations of words and their default senses, to send us first one way and then another. In particular, as noted by Naseem *et al*. (2020), ironic and sarcastic utterances tend to imbue certain words with a sentiment that runs counter to their default settings in the mental

lexicon. So, for instance, we expect the sentiment of "chair" to fall from a default neutral rating to a low negative rating when it is read as "electric chair." If the job of a detector is to recognize this shift, it is the job of a generator to first cause it.

Neural and other statistical models distribute their expectations across a great many internal variables that resist easy interrogation. However, as we have seen in the context of similes, certain constructions lay bare their presuppositions in a way that facilitates large-scale harvesting (Hearst, 1992). Linguistic expectations take many forms, but our focus here is on the properties we consider typical of a concept like *chair* or, indeed, *electric chair*, and how they are subverted by irony. These properties, mined from web similes, form the basis of the EPIC model that is discussed next, and so provide grist to the subversive mill of irony generation.

4.1. EPIC Successes

What is proposed is a model of property-oriented expectation, named EPIC, in which an expectation (E) predicts a property (P) of an instance (I) of concept (C). Take the concept C denoted by "party." An instance I, such as a birthday party, carries with it a set of expectations {E} of the typical properties {P} of a party (C): for example, I is expected to be fun, entertaining and social. But an expectation E fails if the expected property P cannot be asserted of I, and fails ostentatiously if we can instead assert its opposite, not-P. Yet even if E fails in a more subtle way, the task of the ironist is to exaggerate the truth for wit's sake. In this way, a failed expectation E_1 of I_1 predicting P can echo another expectation E_2 of a non-salient concept C_2 that predicts not-P. Just as parties should to be fun and entertaining, we often expect lectures to be dull and boring. So, in failing to be fun, a party I_1 satisfies another expectation that students warily predict of their lectures. But by matching a failed expectation for P to an non-salient expectation for not-P, an ironist dramatizes the degree to which I_1, an instance of C_1, is non-P, by instead pretending that I_1 is an instance of C_2 that predicts not-P. In this way, EPIC builds on all three of the established pillars of irony: echoing, pretense and opposition.

EPIC's expectations are acquired from the similes that make them explicit. For example, a simile of the form "as <P> as <I>" makes plain the expectation that I, an instance of the implied concept C, should have the property P. Harvesting can

work forwards, from properties to instances, or backwards, from instances to the properties that we expect of them, or back *and* forth in a bootstrapping process that first acquires instances for properties, and then other properties for those instances, and so on until the model has a surfeit of both (Veale and Li, 2009).

Counter-expectations – for instance, that P is *not* expected of instances of C – can be acquired or generated in a number of ways. A dictionary of antonyms can be used to generate the counter-expectation not-P_2 from an expectation P_1 when it is known that P_2 is an antonym of P_1, such as *poor* for *rich* or dull for *exciting*. If P_2 is expected and P_1 is evoked, or vice versa, the result is a semantic opposition in the mold of Raskin (1985). While we can engineer these oppositions, they also emerge as a natural byproduct of the simile harvesting process. As noted in Veale (2013), 15% to 20% of the similes gathered from the web are ironic, which is to say, a simile purporting to exemplify P instead exemplifies not-P (for example, "as welcome as a skunk at a garden party"). Hao and Veale (2010) thus present a classifier for automatically determining the sincerity of each harvested simile. Its most discriminating criteria include: whether the simile is hedged with a marker of semantic imprecision, such as "about," or whether the majority of all instances of that simile on the web are so marked; whether the property P that is asserted by the simile has a positive sentiment (since irony criticizes, we expect P to have a positive affect in ironic similes); and whether there is evidence that a similar assertion has been made using a construction that is less conducive to irony (for instance, if "sophisticated foods such as caviar" is attested in a corpus, we can be confident that "as sophisticated as caviar" is a sincere comparison).

Those similes classified as sincere provide expectations {E} of properties {P} for a given instance of concept C, while those deemed insincere provide counter-expectations of C's properties – or, if you like, expectations of C's anti-properties. For the most part, the similes of the former are simpler and less florid than those of the latter, and rely more on simple nouns (e.g., "as sharp as a knife," "as bent as a banana") than on complex constructions with emergent properties (e.g., "as conspicuous as a fart at a chili festival"). Fishelov (1990) denotes the former as NP similes, for *non-poetic*, and the latter as P similes, for *poetic*, even if the poetry is a savage kind of vernacular wit. In all, our harvesting efforts reap over 75,000

property expectations across almost 10,000 concepts, and over 10,000 counter-expectations, such as that zombies at dinner parties are sophisticated while cops in donut shops are observant. When made explicit in this way, these expectations can fuel the automatic production of utterances that intentionally thwart them.

4.2. Failing Gracefully

The thwarted expectation in which an ironic utterance is rooted can take many forms. EPIC assumes that an expectation concerns a property P of concept C, but failure has indirect effects too. An expectation P of C_1 may also concern a related concept C_2 via the relation $<C_2\ R\ C_1>$. An ironist can thus compare C_1 to a C_3 for which not-P is expected, on the basis of the analogous relation $<C_4\ R\ C_3>$ and the analogy $C_1{:}C_2{::}C_3{:}C_4$. Since C_1 and C_3 are not so much compared as contrasted on the basis of a conflict between P and not-P, this juxtaposition is more disanalogy than analogy. In the prior example of parties and lectures. the disappointment of a failed event can be conveyed with irony using the following disanalogy:

> Some hosts arrange "entertaining" parties the way
> some presenters arrange boring lectures.

We can now appreciate the role of the shared relation R (in this case, *arrange*): it focuses the ironic charge of the disanalogy toward those who arrange the parties that fall so short of our EPIC expectations, in much the same way that explosives experts shape their charges to explode in a particular direction. By wrapping the expected property “entertaining” in ostentatious scare-quotes, the charge seems to echo a lie, a failed prediction that the speaker now repeats with disdain. This “echoic mention” of an injudicious prediction (Sperber & Wilson, 1981; Kreuz & Glucksberg, 1989) offers a way of elevating the veiled criticism of irony into open mockery. Indeed, if the criticism were expressed on Twitter, one might go so far as to append the hashtag #*irony*, as if to say “Isn’t it ironic when …” The relative merits of these various strategies – disanalogy, scare-quotes and overt tagging – for conveying an ironic viewpoint are a subject of evaluation in the next section.

5. Managing expectations with Irony

The success of an ironic utterance hinges on its capacity to highlight the failure of a reasonable expectation. As some are more successful in this regard than others, we need a gradated yardstick of success that goes beyond the binary. For while EPIC predicates success on the inference of not-P in a context that implies P, it does not subscribe to a simple irony-as-opposition view. Instead, it assumes that irony is successful when audiences shift their expectations of C from P toward not-P either in whole or in part. A successful ironic utterance may well leave audiences with the feeling that instances of C occupy a middle-ground between P and not-P that conforms to neither extreme; for example, that "many parties that promise entertainment are only ever entertaining to the people that host them."

While we cannot measure mixed feelings like this, Veale and Valitutti (2017) propose a convenient proxy: if P is a positive property, so that not-P is a negative one, then an ironic statement will only be successful to the extent that audiences downshift their mean rating of P's positivity in the context of the irony. So we can expect, for instance, that the mean positivity of "entertaining" in a null context – such as in a dictionary setting – is higher than its mean rating in the context of a disanalogy that lends the word a halo of disappointment. We can thus see irony as a means of granting words a different valence in context than they possess by default, either in the mental lexicon or a printed dictionary (Naseem *et al*., 2020). This is perhaps the smallest echo that verbal irony can achieve: to echo the sense of a single word, but not its sentiment, to attach a new feeling to a familiar idea. Larger echoes still can be found in analogies that, in the vein of Gentner (1983), reflect the systematic relationships of one conceptual structure in another, while allowing the entities governed by those relations to vary across structures, much as the relation *arrange* is echoed in the mapping of *host:party* to *lecturer:lecture*.

In any case, a computer now has several strategies for communicating ironic intent: relational analogy, scare-quoting, and explicit opposition. Note, however, that opposition works hand-in-glove with analogy, since it is the latter that sets up an explicit clash of opposites in the first place. To use the terms of Attardo *et al*. (2002), analogy is the logical mechanism, or LM, that engineers the conflict. All three strategies work together in the following machine-generated tweet:

When “cultured” gentlemen pursue ladies
the way feral predators pursue prey.

Valitutti and Veale (2015) sought to unpick opposition from analogy by crowd-sourcing an evaluation of machine-generated analogies like this. Each rater was explicitly asked to rate the likelihood of ironic intent in each utterance, and was exposed to variants that either included, or excluded, the contrastive properties; when they were included, they saw variants with or without scare-quotes. Raters used their own judgment as to what constitutes irony, but the results show that a mix of analogy and scare-quotes has a statistically significant effect ($p < .001$). A weaker significance ($p < .002$) is reported for the benefit of explicit contrast, but this is regardless of whether or not scare-quotes are also employed.

It is the property that enables the greatest downshift, and conveys the deepest disappointment, that is wrapped in scare-quotes. The quotes alert an audience to a pretense that works on multiple levels. For example, the speaker here pretends to use the focal word “cultured” in its default sense; certain men pretend to be cultured when, at heart, they are not; and polite society pretends that men who act in a predatory fashion are still deserving of labels like “cultured.” This allows us to tease apart the relative contribution of scare-quoting and analogy to the successful communication of an ironic stance, by measuring the mean affective downshift that each causes in the audience’s perception of the focal property.

A computational approach to metaphor generation (see Veale, 2018) is used to produce 80 distinct utterances, all with the same structure as the “cultured” and “entertaining” examples above. Each uses as its focal property a word P with a positive sentiment, and an expectation E that P will attach to all instances I of concept C. The extent to which an utterance shifts P from a high positivity rating to one closer to that of not-P is the extent to which it succeeds at dashing the expectation E and conveying its ironic intent. Different choices of P are picked at random from a dictionary of affect (Whissell, 1989) if they satisfy four criteria: P has, by default, a high positive sentiment; an antonym or near-antonym of P can be found in a dictionary of antonyms; at least one expectation E attaches to P in the EPIC model; and a contrastive analogy can be built around E that puts P and its antonym in direct analogical alignment, thereby opposing P and not-P.

5.1. Framing an ironic observation

A crowd-sourced evaluation of the resulting 80 analogies has been conducted via the crowdsourcing platform *Figure Eight* (née *CrowdFlower*). Anonymous judges were recruited for the task, and each was paid a small sum to rate the positivity of the focal word in each utterance to which they were exposed. Each judge saw just one linguistic framing (of a possible four variants) for a sampling of the 80 analogies. These four structural variants are defined and labeled as follows:

BASE + QUOTE + COMP: This triad is illustrated above for the analogy *gentleman as predator*. A base expectation E (e.g., that *gentlemen are cultured*) is placed in a relational context (e.g., *cultured gentlemen pursue ladies*), and then compared to a context with an opposing expectation for the same relation (e.g., *feral predators pursue prey*). The focal property of E is wrapped in scare-quotes (e.g., *"cultured"*).

BASE + COMP: This variant omits the quotes but retains the contrastive analogy, as in: *When cultured gentlemen pursue ladies the feral way predators pursue prey.*

BASE + QUOTE: This variant places the base expectation in its relational setting, keeps the quotes and omits the analogy, as in: *"Cultured" gentlemen pursue ladies*.

BASE: This variant omits all but the base expectation in its relational setting, as in: *Cultured gentlemen pursue ladies*.

Ten judges are recruited to rate each item, that is, each variant of each analogy. Having read the presented variant, a judge rates the positivity of the focal word as they perceive it in context. Ratings are elicited on a six-point scale running from -3 (very negative) to + 3 (very positive). To force raters off the fence, 0 is disallowed as a response. These scores are normalized to the range -1.0 to +1.0, and aggregated to yield a mean positivity score for each variant of each analogy. By further averaging these means across different analogies with the same form, we can estimate the mean positivity per structural variant, as shown in Table I.

Table I: Mean positivity per variant, with the likelihood of a positive rating.

Structural Variant	*Mean Positivity*	*Positive Likelihood*
BASE	0.51 (*SD* 0.38)	0.91 (*SD* 0.15)

BASE+QUOTE	0.41 (*SD* 0.46)	0.82 (*SD* 0.13)
BASE+COMP	0.29 (*SD* 0.49)	0.75 (*SD* 0.15)
BASE+QUOTE+COMP	0.20 (*SD* 0.54)	0.64 (*SD* 0.16)

As raters are forced to make a choice that leans positive or negative, the second column in Table I reports the likelihood that a rater will offer a positive rating for a given variant (Valitutti and Veale, 2015). We expect the BASE variant to be the least ironizing of contexts for a focal word, although it is still possible that certain pairings – such as "gentleman" with "pursue" – will evoke a frisson of opposition. Nonetheless, those word choices are a constant across all variants of the same analogy, so the relative positivity scores of different variants can still provide us with a reliable estimate of the resulting downshift. It is clear from Table I that certain structural variants cause a more significant downshift that others. Yet, as also reported by Valitutti and Veale, the differences in mean positivity between any two variant types are statistically significant at the $p < .001$ level.

For instance, the *BASE+QUOTE+COMP* tried causes a mean average downshift of .31 relative to the *BASE* formulation alone, and adds an extra downshift of .21 to *BASE+QUOTE*. In contrast, the addition of scare-quotes adds to the downshift provided by contrastive analogy alone (*BASE+COMP*) by just .09 on average. It is also clear from these results that scare-quotes (*QUOTE*) and analogical contrast (*COMP*) are additive effects when framing an ironic intent: the downshift gained when using both together is at least as great as the sum of their individual shifts.

5.2. Telegraphing an ironic intent

If irony is an insincere echo that raises the suspicions of its audience, scare-quotes let us target just one part of that echo for additional scrutiny. But we can go further still, and openly declare our intent to be ironic, such as by using the hashtag #*irony* on Twitter. This strategy, which is almost exclusively confined to social media, is not as counter-productive as it seems. Irony is not without risk in settings that encourage spontaneity *and* blame, where small errors of judgment can spread quickly and stoke the ire of strangers. Viewed from this angle, #*irony*

is a preemptive marker of the "I was only joking" variety, delivered before rather than after the fact. Moreover, while the marker dilutes the ambiguity of verbal irony, it is quite in keeping with situational irony, and can be read as a shorthand for the expression "Isn't it ironic." As noted in Reyes, Rosso and Veale (2013), the line between verbal and situational irony is often a very fine one in social media, especially if our aim is to highlight a situation in which others seem hypocritical. And so it is with many of our machine-generated examples, as in the following:

#Irony: When some activists promote "enduring" principles
the way trendsetters promote temporary fads.

Valitutti and Veale (2015) set out to quantify the effect of the explicit #*irony* tag on machine-generated ironic utterances. Since contrastive analogy is the strategy that achieves the greatest individual downshifting effect, they took *BASE+COMP* as the foundation for their inquiries. In a second crowd-sourcing experiment that mirrors the first, raters were asked to assess the positivity of the focal word in a set of utterances that conform to either the *BASE+COMP*, *BASE+COMP+QUOTE, BASE+COMP+HASH* or *BASE+COMP+QUOTE+HASH* variants, where *HASH* denotes the affixing of the hashtag #*irony* to the front of the utterance, as shown above.

Their findings suggest that overt tagging of this kind does not actually serve to promote an ironic reading, at least when irony is operationalized as a downshift in perceived positivity. As shown in Table II, the addition of an #*irony* tag does result in a small downshift, either when it is used alone (.05) or in combination with scare-quotes (.16), but only the latter is significant at the $p < .001$ level.

Table II: Mean downshift w.r.t. *BASE+COMP*, and its statistical significance.

Structural Variant	*Mean Downshift*	*Significance*
BASE+COMP *+ QUOTE*	0.18	$p < .001$
BASE+COMP *+ HASH*	0.05	$p > .05$
BASE+COMP *+ HASH+QUOTE*	0.16	$p < .001$

What can explain this? Although raters are still asked to assess the positivity of a

specific focal word, they may implicitly shift their attention to the #*irony* marker when it is present. If the perceived sentiment of this marker is downshifted, so that it takes on the negative sentiment of a sarcastic, utterance-final *not*, then the focal word can be taken at face value while the whole utterance is read as ironic. In any case, the ultimate irony is that #*irony* is not a productive marker of irony.

6. A Giant Sarcastic Machine: What a *Great* Idea!

A contrastive analogy that sours into disanalogy yields a first-order ironic echo. The propositional content from which the echo is constructed must be aligned just so, to create an ironic effect, with perhaps a little help from scare-quotes and other signaling devices. But this content is not itself ironic. Rather, irony emerges from the insincere echoing of the sincere expectations of the EPIC model. Recall, however, that when harvesting those expectations from the web, a machine may also gather a large number of counter-expectations from ironic comparisons. If a machine now reuses an expression that conveys one of these contrary positions, and does so for ironic purposes, what it produces is a second-order ironic echo.

This echo of an echo should preserve the content words of the original simile, since the irony is sparked by the vivid mental images that they collectively paint. However, the function words and other syntactic elements that turn this content into a simile can be replaced, to create a different kind of linguistic container that houses the same ironic spark. A machine need not grasp the nuances of the irony, or "get" the joke, to perform this transformation, which is wit-preserving rather than wit-generating. Veale (2019) uses this *echo-of-an-echo* approach to imbue a book-recommending Twitterbot with an ironic sense of humour. The bot, named @*ReadMeLikeABot*, offers personalized book recommendations on the basis of a user's latest tweets, and uses much the same approach to personality-profiling as Ghosh and Veale (2017) to obtain an 11-dimensional picture of a user's recent mood. From this picture it then derives one or more characterizations of the user, such as "laid-back," "neurotic" or "sophisticated," so as to use those labels as retrieval indices for its book selections. Consider these alternate framings of a recommendation for a user that the bot has characterized as "philosophical":

> Hey @bookreader, I used to be as philosophical as a bowel movement until I read 'Steppenwolfe' by Hermann Hesse on the solitude theme.
>
> Hey @bookreader, given your personality profile I don't know which philosophical book is more you: 'Steppenwolfe' by Hermann Hesse on the solitude theme, or 'The Bowel Movement' by Stephen Tolkien.

Each framing reuses the ironic simile "as philosophical as a bowel movement." In the first, it is reused wholesale, function words and all, so it remains a simile. In the second, its content is repurposed as a nonce book title with a fictional author of its own (a random cut-up of real authors). The bot has several strategies, and several reasons, for creating these second-order echoes. For example, it turns the simile "as welcome as a skunk at a garden party" into a faux-mystery novel, "The Affair of The Skunk at The Garden Party," and invents a new author to go with it, "Agatha Chandler," by cutting up the names of famous mystery writers. While the title echoes the original simile, it also playfully echoes the commercial norms of the publishing industry. These pretend books, a kind of ironic pretense in its own right, allow the bot to make fun of the real books with which they are juxtaposed. As the bot's recommendations are public, it seeks to avoid repetition when it can. It does not recommend the same book twice to the same user, nor does it make the same recommendation to different users on the same day. When this leaves it with nothing new to recommend, the bot falls back on satire and offers one of its nonce creations instead. So the bot leans most on irony when it is forced to, and only once it has already demonstrated its bona fides as a book recommender.

The inherent pretense of irony is well-suited to this kind of glib satirizing, and ironic similes can be massaged into many different linguistic forms. For instance, Veale (2021) shows how a domain with entrenched linguistic norms, such as the world of brew pubs and hipster beers, can be satirized by pouring ironic similes into its most standardized containers. The names of traditional English pubs, for instance, conform to the construction "The X and Y," as in "The Dog and Pony" and "The Drop and Bucket," and many ironic similes can be adapted accordingly. Consider, for example, "The Elephant and Tutu," "The Fart and Elevator," "The Fish and Tree" and "The Corpse and Disco." In the same way, beer names can be repurposed from vivid simile vehicles such as "Rabid Dog", "Mediaeval Ordeal,"

“Hungry Snake” and “Kosher Pork.” The process is a facile one, and the machine fails to grasp the strange logic of each pairing, but it does not need to. It simply needs to detect irony in one linguistic container so it can transfer it into another.

Computational approaches to irony, metaphor, joking and linguistic creativity more generally show that detection is always easier than interpretation, and that purposeful generation is harder still. Statistical language models trained on vast amounts of raw text fare much better at anomaly detection than at meaningful anomaly generation. They are natural pedants, but they are not natural wits. In this respect, they conform to William Empson’s critique of George Orwell: “the eagle-eye with the flat feet.” Consider the statistical language model employed by Google in its Gmail service. The model, trained on a large corpus of email texts, is so attuned to the norms of email-writing that it can offer plausible completions for half-written sentences, giving algorithmic form to a fear expressed by Orwell (1946) that “ready-made phrases … will construct your sentences for you – even think your thoughts for you.” But the model also detects unlikely deviations from the norm, to identify those parts of a text that may need to be rewritten. To the pedant, wit can seem like an error in need of fixing. To the wit, an apparent error can carry a deliberate meaning, but this insight comes only from interpretation.

The following extract from an email to a colleague illustrates the distinction:

> BTW, we went to see "Tenet" last night. Our brains are still bent out of shape. But we plan to see it again last week, so we'll understand it eventually.

The joke, such as it is, requires prior knowledge of the time-travel movie “Tenet,” in which the cod-science principle of *entropy reversal* allows people and objects to move backwards through time. Without this knowledge, the choice of “last” in the phrase “see it again last week” will seem like an obvious error, and so Gmail dutifully underlines it in blue. Theories of humour are necessarily generic, but jokes always hinge on the specifics. While the machine excels at recognizing the deviation, and at suggesting a fix (“next” for “last”), it lacks the specific insight to understand it not as a mistake but as an exploitation of the norm (Hanks, 2013).

The choice of “last” here is deliberately ironic: I wanted the recipient to read it as “last” (for my time-travelling pretense) *and* “next” (for my actual intent). It is

hardly surprising that Gmail fails to appreciate the deliberateness of the choice, yet this does not preclude statistical language models from playing an important role in detecting and appreciating wit. They may be pedants, but an attention to detail and an ingrained sense of orthodoxy are key ingredients in the enterprise of humour. But a statistical model cannot serve as a complete solution in itself. Insofar as irony requires us to assume a dual perspective on a situation, to see it for what it is and for what was expected of it, we shall need multiple competing models to construct this conceptual parallax. Statistical models like Gmail's are wide-ranging savants regarding the natural rhythms of language, and they tacitly incorporate many, if not all, of the expectations that are overtly captured in EPIC. Bosselut *et al*. (2019) show how a large, pre-trained language model can be used to generate not just text completions, but new entries in an EPIC-like database of common sense knowledge. However, even a statistical model steeped in implicit wisdom still needs a view from the outside to make sense of a perceived failure of expectation. This external view may be provided by another statistical model, albeit one trained on different data to imbue it with a more comic sensibility – imagine one model leaning to James Joyce and another to Woody Allen, say – or it might come from an assemblage of distinct systems, each focusing on a specific kind or theory of humour, or on a different logical mechanism for its production.

In fact, a heterogeneous assembly of complementary systems may be best able to tackle a phenomenon that looks like a pretense to some, an echoic mention to others, and simple opposition to everyone else. In the composite view, these are not rival theories but complementary modules, lending their individual voices to a collective vote as to whether a given utterance should be understood as ironic. The ends typically justify the engineering means in computational modelling, so a computational approach to irony is no substitute for an actual theory of irony, especially since data-driven approaches tend to trade explainability for accuracy and robustness. However, explainability can yet emerge from how the individual pieces are all stitched together, in a theory-led fashion, to realize an engineering solution that enables the detection, interpretation and generation – and perhaps even appreciation – of this intriguing, if vexing, facet of human communication.

References

Attardo, Salvatore, Christian F. Hempelmann and Sara Di Maio. "Script oppositions and logical mechanisms: Modeling incongruities and their resolutions." *Humor: International Journal of Humor Research* 15, no. 1 (2002): 3-46.

Attardo, Salvatore. "Register-based humor." In *The Primer of Humor Research*, edited by Victor Raskin. Berlin: Mouton de Gruyter (2009):230-253.

Bosselut, Antoine, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz and Yejin Choi. "COMET: Common-sense Transformers for Automatic Knowledge Graph Construction." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy (2019):4762–79.

Chander, Daniel. *Semiotics: The Basics*. London, UK: Routledge, 2020.

Clark, Herbert H., and Richard J. Gerrig. "On the pretense theory of irony." *Journal of Experimental Psychology: General* 113, no. 1 (1984): 121-126.

Fishelov, David. "Poetic and Non-Poetic Simile: Structure, Semantics, Rhetoric." *Poetics Today* 14, no. 1 (Spring, 1993): 1-23.

Garmendia, Joana. Irony as Opposition. In *Irony* (Key Topics in Semantics and Pragmatics. Cambridge: Cambridge University Press, (2009):17-41.

Gentner, Dedre. "Structure-mapping: A Theoretical Framework." *Cognitive Science* 7, no. 2 (1983):155–170.

Ghosh, Aniruddha, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden and Antonio Reyes. "Semeval-2015 task 11: Sentiment analysis of figurative language in twitter." Proceedings of the 9th international workshop on semantic evaluation, Denver, Colorado (2015):470-478.

Ghosh, Aniruddha, and Tony Veale. "Fracking Sarcasm with Neural Networks." In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, edited by Alexandra Balahur, Erik van der Goot, Piek Vossen, and Andres Montoyo, 161-168. San Diego, California, 2016.

Ghosh, Aniruddha, and Tony Veale. "Magnets for Sarcasm: Making Sarcasm Detection Timely, Contextual and Very Personal." *Proceedings of the Conference on*

Empirical Methods in Natural Language Processing, edited by Martha Palmer Rebecca Hwa, and Sebastian Riedel, 493–502. Copenhagen, Denmark, 2017.

Giora, Rachel, Ofer Fein, Ann Kronrod, Idit Elnatan, Noa Shuval and Adi Zur. "Weapons of Mass Distraction: Optimal Innovation and Pleasure Ratings." *Metaphor and Symbol* 19, no. 2 (2004): 115-141.

Giora, Rachel. "Lying, Irony, and Default Interpretation." In *The Oxford Handbook of Lying*, edited by Jorg Meibauer. London: Oxford University Press, 2018.

Grice, H. Paul. "Logic and conversation." In *Syntax and Semantics 3: Speech Acts*, edited by Peter Cole and Jerry L. Morgan, 183-198. Cambridge, MA: Academic Press, 1975.

Hanks, Patrick. Lexical Analysis: Norms and Exploitations. Cambridge, MA: MIT Press, 2013.

Hao, Yanfen, and Tony Veale. "An Ironic Fist in a Velvet Glove: Creative Mis-Representation in the Construction of Ironic Similes." *Minds and Machines* 20, no. 4 (2010): 483-88.

Hearst, Marti A. "Automatic Acquisition of Hyponyms from Large Text Corpora." Proceedings of the 15th International Conference on Computational Linguistics, edited by Christian Boitet, 539-545. Nantes, France, 1992.

Kao, Justine T., Roger Levy, and Noah D. Goodman. "A Computational Model of Linguistic Humor in Puns." *Cognitive Science* 40, no. 5 (2016): 1270-85.

Kreuz, Roger J., and Sam Glucksberg. "How to be sarcastic: The echoic reminder theory of verbal irony." *Journal of Experimental Psychology: General* 118, no. 4 (1989): 374-386.

Kumon-Nakamura, Sachi, Sam Glucksberg, and Mary Brown. "How about another piece of pie: The Allusional Pretense Theory of Discourse Irony." *Journal of Experimental Psychology: General* 124, no. 1 (1995): 3-21.

Naseem, Usman, Imran Razzak, Peter Eklund and Katarzyna Musial. "Towards Improved Deep Contextual Embedding for the identification of Irony and Sarcasm." Proceedings of the International Joint Conference on Neural Networks *(IJCNN)*, Glasgow, UK, 2020, 1-7.

Orwell, George. “Politics and The English Language.” *Horizon* 13, no. 76 (April 1946).

Liu, Liyuan, Jennifer L. Priestley, Yiyun Zhou, Herman E. Ray and Meng Han. "A2Text-Net: A Novel Deep Neural Network for Sarcasm Detection." Proceedings of the 1st IEEE International Conference on Cognitive Machine Intelligence, Los Angeles, CA, 2019, 118-126.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskeve. “Language Models are Unsupervised Multitask Learners.” *Open AI Technical Paper*, 2019.

Raskin, Victor. *Semantic Mechanisms of Humor*. Dordrecht: D. Reidel, 1985.

Reyes, Antonio, Rosso, Paolo and Veale, Tony. “A multidimensional approach for detecting irony in Twitter.” *Language Resources & Evaluation* 47 (2013):239–268.

Riloff, Ellen, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert and Ruihong Huang. “Sarcasm as Contrast between a Positive Sentiment and Negative Situation.” Proceedings of the Conference on Empirical Methods in Natural Language Processing, edited by D. Yarowsky, T. Baldwin, A, Korhonen, K. Livescu, and S. Bethard, 704–714. Seattle, Washington, 2013.

Searle, John. “Minds, Brains and Programs.” *Behavioral and Brain Sciences* 3, no. 3 (1980): 417–457.

Shahaf, Dafna, Eric Horvitz, and Robert Mankoff. “Inside Jokes: Identifying Humorous Cartoon Captions.” Proceedings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, edited by T. Joachims, G. Webb, D. Margineantu and G. Williams, 1065-1074. Sydney, Australia, 2015.

Sperber, Dan and Deirdre Wilson. “Irony and the Use-Mention Distinction.” In *Radical Pragmatics*, edited by Peter Cole, 295-318. New York: Academic Press, 1981.

Sperber, Dan. “Verbal Irony: Pretense or Echoic Mention?” *Journal of Experimental Psychology: Genera*l 133, no. 1 (1984): 130-136.

Tausczik, Yla R., and James W. Pennebaker. “The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods.” *Journal of Language and Social*

Psychology 29, no. 1 (2009): 24-54.

Tsur, Oren, Dmity Davidov and Ari Rappoport. "ICWSM – A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews." Proceedings of the 4th International AAAI Conference on Weblogs and Social Media, edited by W. Cohen and S. Gosling, 161-169. Washington, D.C., 2010.

Valitutti, Alessandro and Tony Veale. "Inducing an ironic effect in automated tweets." Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII), edited by B. Schuller, 153-159. Xi'an, China, 2015.

Veale, Tony. "Humorous Similes." *Humor: The International Journal of Humor Research* 21, no. 1 (2013): 3-21.

Veale, Tony. "The 'default' in our stars: Signposting non-defaultness in ironic discourse." *Metaphor and Symbol* 33, no. 3 (2018): 175-184.

Veale, Tony. "Read Me Like A Book: Lessons in Affective, Topical and Personalized Computational Creativity." Proceedings of the 10th International Conference on Computational Creativity, edited by K. Grace, M. Cook, D. Ventura, and M.L. Maher, 25-32. Charlotte, North Carolina, 2019.

Veale, Tony. *Your Wit Is My Command: Building AIs With A Sense Of Humor*. Cambridge, MA: MIT Press, 2021.

Veale, Tony and Guofu Li. "Growing finely-discriminating taxonomies from seeds of varying quality and size." Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, edited by C. Gardent and J. Nivre, 835-842, Athens, Greece, 2009.

Veale, Tony, and Alessandro Valitutti. "Sparks Will Fly: Engineering Creative Script Conflicts." *Connection Science* 29, no. 4 (2017): 332-349.

Whissell, Cynthia. "The dictionary of affect in language." In *Emotion: Theory and research*, edited by Robert Plutchik and Henry Kellerman, 113-131. San Diego: Harcourt Brace, 1989.

Zhang, Shiwei, Xiuzhen Zhang, Jeffrey Chan and Paolo Rosso. "Irony detection via sentiment-based transfer learning." *Journal of Information Processing and Management* 56, no. 5 (2018):1633–1644.