Game of Tropes:

Exploring the *Placebo Effect* in Computational Creativity

Tony Veale
School of Computer Science and Informatics
University College Dublin, Belfield D4, Ireland.
Tony.Veale@UCD.ie

Abstract

Twitter has proven itself a rich and varied source of language data for linguistic analysis. For Twitter is more than a popular new channel for social interaction in language; in many ways it constitutes a whole new genre of text, as users adapt to its new limitations (140 character messages) *and* to its novel conventions such as retweeting and hash-tagging. But Twitter presents an opportunity of another kind to computationally-minded researchers of language, a *generative* opportunity to study how algorithmic systems might exploit linguistic tropes to compose novel, concise and re-tweetable texts of their own. This paper evaluates one such system, a *Twitterbot* named *@MetaphorMagnet* that packages its own metaphors and ironic observations as pithy tweets. Moreover, we use *@MetaphorMagnet*, and the idea of Twitterbots more generally, to explore the relationship of linguistic containers to their contents, to understand the extent to which human readers fill these containers with their own meanings, to see meaning in the outputs of generative systems where none was ever intended. We evaluate this *placebo* effect by asking human raters to judge the comprehensibility, novelty and aptness of texts tweeted by simple *and* sophisticated Twitterbots.

Tropes: *Containers of Meaning*

A mismatch between a container and its contents can often tell us much more than the content itself, as when a person places the ashes of a deceased relative in a coffee can, or sends a brutal death threat in a *Hallmark* greeting card. The communicative effectiveness of mismatched containers is just one more reason to be skeptical of the *Conduit* metaphor (Reddy, 1979) – which views linguistic constructs as containers of propositional content to be faithfully shuttled between speaker and hearer – as a realistic model of human communication. Language involves more than the faithful transmission of logical propositions between information-hungry agents, and more effective communication – of attitude, expectation and creative intent – can often be achieved by abusing our linguistic containers of meaning than by treating them with the sincerity that the *Conduit* metaphor assumes. Consider the case of verbal irony, in which a speaker deliberately chooses containers that are pragmatically ill-suited to the conveyance of their contents. For instance, the advertising container "*If you only see one [X] this year, make it this one*" assumes that [X] denotes a category of event – such as "*romantic comedy*" or "*movie about superheroes*" – with a surfeit of available members for a listener to choose from. When [X] is bound to the phrase "*comedy about Anne Frank*" or "*musical about Nazis*", the container proves too hollow for its content, and the reader is signaled to the presence of playful irony. Though such a film may well be one-of-a-kind, the ill-fitting container suggests there are good reasons for this singularity that do not speak to X's quality as an artistic event. Yet if carefully chosen, an apparently inappropriate container can communicate a great deal about a speaker's relationship to the content conveyed within, and as much again about the speaker's relationship to their audience.

As more practical limitations are placed on the form of linguistic containers, the more incentive one has to exploit or abuse containers for creative ends. Consider the use of Twitter as a communicative medium: writers are limited to micro-texts of no more than 140 characters to convey both their meaning and their attitude to this meaning. So each micro-text, or *tweet*, becomes more than a container of propositional content: each is a brick in a larger edifice that comprises the writer's online personae and textual aesthetic. Many Twitter users employ irony and metaphor to build this aesthetic and thus build up a loyal audience of followers for their world view. Yet Twitter challenges many of our assumptions about irony and metaphor. Such devices must be carefully modulated if an audience is to perceive a speaker's meaning in the playful (mis)match of a linguistic container to its contents. Failure to do so can have serious repercussions when one is communicating to thousands of followers at once, with tweets that demand concision and leave little room for nuance. It is thus not unusual for even creative tweets to come packaged with an explicit tag such as *#irony*, *#sarcasm* or *#metaphor*.

Metaphor and irony are much-analysed phenomena in social media, but this paper takes a generative approach, to consider the *production* rather than the *analysis* of creative linguistic phenomena in the context of a fully-autonomous computational agent – a *Twitterbot* – that crafts its own metaphorical and ironical tweets from its own knowledge-base of common-sense facts and beliefs. How might such a system exhibit a sense of irony that human users will find worthy of attention, and how might this system craft interesting metaphoric insights from a knowledge-base of everyday facts that are as banal as they are uncontentious? We shall explore the variety of linguistic containers at the disposal of this agent – a real computational system named *@MetaphorMagnet* – to better understand how such containers can be playfully exploited to convey ironic, witty or thought-provoking views on the world. With *@MetaphorMagnet* we aim to show that interesting messages are not crafted from interesting contents, or at least not necessarily so. Rather, effective tweets emerge from an appropriate if non-obvious combination of familiar linguistic containers with unsurprising factual fillers. In support of this view, we shall present an empirical analysis of the assessment of *@MetaphorMagnet*'s uncurated outputs by human judges.

Just as one can often guess the contents of a physical container by its shape, one can often guess the meaning of a linguistic container by its form. We become habituated to familiar containers, and just as we might imagine our own uses for a physical container, we often pour our own meanings into suggestive textual forms. For in language, meaning follows form, and readers will generously infer the presence of meaning in texts that are well-formed and seemingly the product of an intelligent entity, even if this entity is *not* intelligent and any meaning is *not* intentional. Remarkably, Twitter shows that we willingly extend this generosity of interpretation to the outputs of bots that we *know* to be unthinking users of wholly aleatoric methods. Twitterbots exploit this *placebo effect* – wherein a well-formed linguistic container is presumed to convey a well-founded semantic content – by serving up linguistic forms that readers tacitly fill with their own meanings. We aim to empirically demonstrate here that readers do more than willingly suspend their disbelief, and that a well-packaged linguistic form can seduce readers into seeing what is not there: a comprehensible meaning, or at least an intent to be meaningful. We do this by evaluating two metaphor-generating bots side-by-side: a rational, knowledge-based Twitterbot named *@MetaphorMagnet* vs. an aleatoric and largely knowledge-free bot named *@MetaphorMinute*.

Digital Surrealists: *La Règle Du Jeu*

Most Twitterbots are simple, rule-based systems that use stochastic methods to explore a loosely-defined space of texual forms. Such bots are high-concept, low-complexity text-production mechanisms that transplant the aleatoric techniques of surrealist writers – from André Breton to William Burroughs and Brion Gysin – into the realms of digital content, social networking and online publishing. Each embodies a language game with its own generative rules, or what Breton called "*la règle du jeu.*" Yet Breton, Burroughs and Gysin viewed the use of aleatorical rules as merely the first stage of a two-stage creation process: at this first stage, random recombinant methods are used to confect candidate texts in ways that, though unguided by meaning, are also free of the baleful influence of cliché; at the second stage, these candidates are carefully filtered by a human, to select those that are novel and interesting. Most bots implement the first stage and ignore the second, pushing the task of critiquing and filtering candidate texts onto the humans who read and selectively re-tweet them.

Nonetheless, some bots achieve surprising effects with the simplest language tools. Consider *@Pentametron*, a bot that generates accidental poetry by re-tweeting pairs of random tweets of ten syllables apiece (for an iambic pentameter reading) if each ends on a rhyming syllable. When the meaning of each tweet in a couplet coheres with the other, as in "*Pathetic people are everywhere"* | *"Your web-site sucks, @RyanAir*", the sum of tweets produces an emergent meaning that is richer and more resonant than that of either tweet alone. Trending social events such as the *Oscars* or the *Super Bowl* are especially conducive to just this kind of synchronicity, as in this fortuitous pairing: "*So far the @SuperBowl commercials blow*." | "*Not even gonna watch the halftime show.*"

In contrast, a bot named *@MetaphorMinute* wears its aleatoric methods on its sleeve, for its tweets – such as "*a haiku is a tonsil: peachblow yet snail-paced*" – are not so much random metaphors as random metaphor-shaped texts. Using a strategy that stresses quantity over quality, this bot instantiates that standard linguistic container for metaphors – the copula frame *"X is a Y*" – with mostly random word choices every two minutes. Interestingly, its tweets are as likely to provoke a sense of mystification and ersatz profundity as they are total incomprehension. Yet bots such as *@Pentametron* and *@MetaphorMinute* do not generate their texts from the semantic-level up; rather, they manipulate texts at the word-level, and thus lack any sense of the meaning of a tweet, or any rationale for why one tweet might be better – which is to say, more interesting, more apt or more re-tweetable – than others.

The Full-FACE poetry generator of Colton *et al*. (2012) also uses a template-guided version of the cut-up method to mash together semantically-coherent text fragments in a way that – much like *@Pentametron* – obeys certain over-arching constraints on metre and rhyme. These text fragments come from a variety of online sources, ranging from short tweets to long news articles. News stories are a rich source of readymade phrases that convey resonant images, and these can be clipped from a news text using standard NLP techniques, while tweets that use affect-rich language can also be extracted automatically via standard *sentiment analysis* lexica and tools. Thus, a large stock of resonant similes, such as "*blue as a blueberry*" or "*hot as a sauna*" can be extracted from the Web using a search engine (Veale, 2014), since the simile frame "*as X as Y*" is specific enough to query for, and promiscuous enough to match, a rich diversity of typical X:Y associations. These associations can then be recast in a variety of poetic forms to make their clichéd offerings seem fresh again, as in "*Blueberry-blue overalls*" or "*sauna-hot jungle*."

Indeed, the very act of juxtaposing clichés can itself be a creative act, as evidenced both by the success of the cut-up method in general and that of specific cut-ups in particular. Consider William Empson's withering analysis of the persnickety, cliché-hating George Orwell, whom Empson called "*the eagle eye with the flat feet*" (quoted in Ricks [1995:356], who admires Empson's "*audacious compacting of clichés"*). The Full-FACE system is just one of many CC systems that use an autonomous variant of Burroughs and Gysin's cut-up method to integrate tight constraints on form with loose constraints on meaning.

Breton famously stated that "*Je ne veux pas changer la règle du jeu, je veux changer de jeu.*" Twitterbots do not change or transcend their own rules, but different bots do represent different language games with their own rules. So to change the game, a CC developer can simply build a new bot, to exploit a different set of tropes and linguistic containers. It is rare for any one Twitterbot to incorporate a diverse set of tropes and production mechanisms; each typically follows Breton's experimentalist approach to art in its random sampling of a specific space of possibilities. Each bot thus forms its own art installation, to showcase a single generative idea. *@MetaphorMagnet*, the bot at the heart of this paper, represents a departure from this norm, insofar as it exploits a wide range of tropes and rendering strategies, it employs diverse sources of knowledge, and it applies a variety of reasoning styles to generate surprising conclusions from what is otherwise a stock of banal facts. But does this added sophistication – bought at the cost of increased system complexity and knowledge-engineering effort – result in tweets that are seen as more meaningful, novel, apt or retweetable by human users? It is this point that exercises us most in the coming sections.

The Placebo Effect : *Trope-A-Dope*

We humans obtain more mileage than we care to admit from templates, tropes and other "bot" tricks for linguistic creativity. Consider what Matthew McGlone and Jessica Tofighbakhsh (1999) call the *Keats heuristic*, an insight into creative language use that owes as much to Nietzsche ("*we sometimes consider an idea truer simply because it has a metrical form and presents itself with a divine skip and jump*") as to the poet John Keats ("*Beauty is truth, truth beauty*"). McGlone and Tofighbakhsh (2000) show that when presented with uncommon maxims or proverbs with internal rhyme (e.g. "*woes unite foes*"), subjects tend to view these as more insightful about the world than the equivalent paraphrases with no internal rhyme at all (e.g. "*troubles unite enemies*"). While the Keats heuristic is not exactly a license to pun, it is an incentive to rhyme, and to give as much weight (or more still) to superficial aspects of poetry generation as to deep semantics and pragmatics. Indeed, the heuristic is tacitly central to the operation of virtually every computational creativity (CC) approach to poetry generation (e.g. Milic, 1970; Chamberlain & Etter, 1983; Gervás, 2000; Manurung *et al*. 2012; Veale, 2013). If human poets ask questions first and rhyme later, CC systems typically rhyme first and ask questions later, if at all. For if the human jury in the O.J. Simpson trial could be turned against bald facts with the Keatsian "*If the glove don't fit you must acquit*", readers of computer-generated poetry can be persuaded to see deliberate meaning and resonance in any output that has a "*divine skip and jump*."

There is something undeniably special about poetry, whether it is the gentle poetry of William Shakespeare's "*Shall I compare thee to a summer's day*" or the rough poetry of Johnnie Cochrane's "*If the glove don't fit you must acquit*". Milic (1970), an early CC pioneer, argues that while poetry "*is more difficult to write than prose*" it offers other freedoms to writers due to the willingness of readers to "*interpret a poem, no matter how obscure, until he has achieved a satisfactory understanding*." What then of the enigmatic tweets of bots like *@MetaphorMinute*, whose obscurity is a function of random word choice and whose surface forms are not designed to make any sense at all? Milic argues that computer poetry serves a useful role other than its obviously generative one, by alerting us to "*the curious behavior of familiar words in unfamiliar combinations*." Behaviour that makes perfect sense when dealing with the writings of a gifted human poet, such as our tendency to "*interpret an utterance by making what concessions are necessary on the assumption that a writer has something in mind of which the utterance is the sign*", is, argues Milic, "*inappropriate when the speaker is a computer*." Yet Twitterbots benefit from such concessions and assumptions whether or not followers know them to be bots. This *placebo effect* is especially pronounced in the coining of would-be metaphors, leading Milic to note "*how readily we accept metaphor as an alternative to calling a sentence nonsensical*." *@MetaphorMinute* and other aleatoric bots wring maximal value from this insight by devising texts that they themselves cannot distinguish from nonsense. So this begs an important question: are the meanings imposed on a random text by a creative human of comparable value to those conveyed by a bot with its own model of the world and its own insights to tweet?

Building Metaphors : *Theory and Practice*

What might it mean for a bot to have "*something in mind of which [its] utterance is the sign*"? When it comes to metaphor generation, we might expect that our bot would generate its figurative tweets from a conceptual model of the world as it sees it, in a way that accords with a sound theory of *how* and *why* humans actually use metaphor. For the latter, AI offers us a range of models to choose from.

Computational approaches to metaphor divide into four broad classes: the categorial, the corrective. the analogical and the schematic. Categorial approaches view metaphor as a means to reconceptualize one idea by placing it into a taxonomic category strongly associated with another (see Hutton, 1982; Way, 1991; Glucksberg, 1998). Corrective approaches view metaphor as an inherently anomalous deviation from literal language, and strive to *recover* the corresponding literal meaning of any figurative statement that violates its lexico-semantic norms (see Wilks, 1978;

Fass, 1991). The analogical approaches aim to capture the relational parallels that allow our representation of an idea in one domain, the *source*, to be systematically projected onto our mental representation of an idea in another, the *target* (see Gentner *et al.*, 1989; Veale and Keane, 1997). Finally, schematic approaches aim to explain how related linguistic metaphors arise as surface manifestions of deep seated cognitive structures called *Conceptual Metaphors* (Lakoff & Johnson, 1980; Carbonell, 1981; Martin, 1990; Veale & Keane, 1992). Each approach has its own merits, but none offers a complete computational solution. Bots that aim for a general competence in metaphor must thus implement a selective hybrid of multiple approaches. Yet each approach also requires its own source of knowledge. Categorial approaches require a comprehensive taxonomy of flexible categories that can embrace atypical members on demand. Corrective approaches are built on a substrate of literal case-frames onto which deviant usages can be correctively projected. Analogical approaches assume an inventory of graph-theoretic representations of concepts, from which a structure-mapping engine can eke out its sub-graph isomorphisms. Schematic approaches rely on a stock of Conceptual Metaphors (CMs) – such as *Life is a Journey* or *Theories are Buildings* – to unearth the deep structures beneath the surface of diverse linguistic forms.

Though hybrid approaches demand multiple sources of knowledge, there exist public Web services that integrate this knowledge with the appropriate means of using it for metaphor. The *Thesaurus Rex* Web service of Veale & Li (2013) provides a highly divergent system of fine-grained categorizations that allows a 3rd-party client system to e.g. determine that *War* and *Divorce* have each been viewed as kinds of *destructive thing*, *traumatic event* and *severe conflict* in the texts of the Web. The *Metaphor Eyes* Web service of Veale & Li (2011) is a rich source of relational norms – also harvested at scale from Web texts – such as that businesses earn profits and pay taxes, or that religions ban alcohol and believe in reincarnation. The *Metaphor Magnet* service of Veale (2014) offers a rich source of the stereotypical properties and behaviors of familiar ideas, and provides the means to retrieve salient CMs from the Google n-grams (Brants & Franz, 2006) which can then be further elaborated to create novel linguistic metaphors.

@MetaphorMagnet relies on each of these public Web services to generate the conceptual conceits that underpin its figurative tweets. For instance, it uses *Thesaurus Rex* to provide the categorization insights that it then packages as *odd-one-out* lists or as *faux*-dictionary definitions. It uses the *Metaphor Eyes* service to provide the relational structures it needs to perform structure mapping and thus concoct original analogies and dis-analogies. And it uses the *Metaphor Magnet* service to access the stereotypical properties and behaviors of ideas, and to juxtapose these properties via resonant contrasts and norm contraventions. Once the conceptual chassis of a metaphor is constructed in this way, it is then packaged in an apt linguistic form.

Building Strings: *Trope-On-A-Rope*

CMs such as *Life Is A Journey* and *Politics Is A Game* are more than productive deep-structures for the generation of whole families of linguistic metaphors; they also provide the conceptual mappings that shape our habitual thinking about such familiar ideas as *Life*, *Love, Politics* and *War*. Politicians and philosophers exploit conceptual metaphors to frame an issue and shape our expectations; when a CM fails to match our own experience, we reject it and switch to a more apt metaphor. So a metaphor-generating bot can thus create a thought-provoking opposition by pitting one CM against another that advocates a conflicting view of the world. The following tweet from *@MetaphorMagnet* uses this approach to contrast two views on *#Democracy*:

To some voters, democracy is an important cornerstone.
To others, it is a worthless failure.
#Democracy=#Cornerstone #Democracy=#Failure

The CM *Democracy Is A Cornerstone* (of society) is often used to frame political discussions, and can be seen as an specialization of the CM *Society Is A Building*, itself an elaboration of the CM *Organization Is Physical Structure* (see Grady, 1997). Yet the importance of cornerstones to the buildings they anchor finds a sharp contrast in the assertion that *Democracy Is A Failure*. Each of these affective claims is so commonly asserted that they can be found in the Google n-grams, a large database of short fragments of frequent Web texts. The 4-gram "*democracy is a cornerstone*" has a frequency of 91 in the Google n-grams, while the 4-gram "*democracy is a failure*" has a frequency of 165. These n-grams, which suggest potential CMs for *@MetaphorMagnet*, are elaborated with added detail via the *Metaphor Magnet* Web service, which tells the bot that the stereotypical *cornerstone* is *important* and the stereotypical *failure* is *worthless*. The following tweet makes similar use of CMs found in the Google n-grams, but renders the conflict in a different linguistic container:

Remember when tolerance was promoted by crusading liberals? Now, tolerance is violence that only fearful appeasers can avoid.

The bot is guided here by the suggestive Google 3-gram "*Tolerance for Violence*" (frequency=1353), but it does not directly contrast the ideas #Tolerance and #Violence. Instead, it finds a potential analogy in this juxtaposition, between the promoters of #Tolerance (which it renders as *crusading liberals*) and the opponents of #Violence (which it renders as *fearful appeasers*). The choice of stereotypical properties (*crusading* and *fearful*) is driven by the bot's need to create a resonant semantic opposition. The bot omits the hashtags #Tolerance=#Violence from this tweet due to the confines of Twitter's 140-character limit. But it can also choose to render a complex conceit across two successive tweets, as in the following pair:

Remember when research was conducted by prestigious philosophers? #Research=#Fruit #Philosopher=#Insect

Now, research is a fruit eaten only by lowly insects. #Research=#Fruit #Philosopher=#Insect

@*MetaphorMagnet* uses a number of packaging strategies to turn a figurative comparison into an ironic observation, ranging from the use of an explicit *#irony* hashtag (which is commonplace on Twitter) to the use of "scare" quotes to focus on the part of a tweet most deserving of disbelief. The following tweet showcases both of these strategies:

#Irony: *When some chefs prepare* "fresh" *salads the way apothecaries prepare noxious poisons.* #Chef=#Apothecary #Salad=#Poison

Irony offers a concise means of contrasting two points of view: that which is expected and the disappointing reality. By comparing the preparation of salads – the "*healthy*" option on most menus – to the preparion of poisons, this analogy undermines the expectation of healthfulness and suggests that some salads are noxious and chemical-filled. The real world is filled with situations in which naturally antagonistic properties are found in surprising proximity. These situations, if expressed in the right linguistic form, can be elevated to the level of situational irony. Consider, for instance, the following @*MetaphorMagnet* tweet:

#Irony: *When the timers that are found in enjoyable games activate gruesome bombs.* #Enjoyable=#Gruesome

It is important to stress that @*MetaphorMagnet* does not simply fill linguistic templates with related words. Rather, the above tweet is constructed at the knowledge-level, by a bot that intentionally seeks out stereotypical norms that are related (e.g. by a pivotal idea *timer*) yet which can be placed into antagonistic juxtapositions around this pivot. In effect, the goal of the linguistic rendering is to package a knowledge-level conceit – typically a conflict of ideas and properties – in a tweet-sized narrative. For example, the following tweet is rendered as a narrative of change:

To join and travel in a pack: This can turn pretty girls into ugly coyotes. #Girl=#Coyote

Twitter offers unique social affordances that allow a bot to elevate almost any contrast of ideas into a dramatic narrative. Rather than talk of generic liberals or appeasers, a bot can give these straw men real names, or at least invent fake names that look like the real thing and which, as Twitter handles, seem wittily apropos to the views that are espoused. In this way, by imagining its central conceit as a topic of a vigorous debate by real people, a bot can turn an abstract metaphor into a concrete situation with its own colorful participants. Consider the social debate that is made personal in this tweet from @*MetaphorMagnet*:

.@war_poet *says history is a straight line*
.@war_prisoner *says it is a coiled chain*
#History=#Line #History=#Chain

The handles @war_poet and @war_prisoner are invented by @*MetaphorMagnet* to suit, and amplify, the figurative views that they are advanced in the tweet, by using a mix of relational knowledge (from the *Metaphor Eyes* service) and language data (via the Google n-grams). Since poets write poems about the wars that punctuate history, and poems contain lines, the 2-gram "*war poet*" is recognized as an apt handle for an imaginary Twitter user who might advance a view of *history as a line*. In this case the handle @war_poet really does name a real Twitter user, but this only adds to the sense that Twitterbot confections are a new kind of interactive theatre and performance art (see Dewey, 2014). Note that the more profound aspects of this contrast are not appreciated by @*MetaphorMagnet* itself, or at least not yet. For example, the bot does not yet appreciate what it means for history to be a straight line, and while it knows enough to invent the intriguing handle @war_prisoner, neither does it appreciate what it might mean to be a prisoner of history, enslaved in a repeating cycle of war. The placebo effect is not a binary effect: it benefits by degrees, and can benefit knowledge-rich bots just as much as knowledge-free bots. Our bots will always evoke in we humans more than they themselves can ever appreciate, yet this may itself be a key part of CC's allure.

Bot Vs. Bot : *The Metaphor Challenge*

@*MetaphorMagnet* differs from @*MetaphorMinute* in a number of key ways. For one, its mechanics are informed by Lakoff and Johnson's *Conceptual Metaphor Theory* and a range of computational approaches. For another, it draws on considerable semantic and linguistic resources, from a large knowledge-base of conceptual relations and stereotypical beliefs to the linguistic diversity of the Google n-grams. All of @*MetaphorMagnet*'s tweets – all its hits and all its misses – are open to public scrutiny on Twitter. But to empirically evaluate the success of the bot as a knowledge-based, theory-driven producer of novel, meaningful and retweet-worthy metaphors, we turn to the crowdsourcing platform *CrowdFlower*, where we conduct a comparative evaluation of @*MetaphorMagnet* and its closest knowledge-*free* counterpart, @*MetaphorMinute*. The latter, designed by noted bot-maker Darius Kazemi, uses a wholly aleatoric approach to metaphor generation yet has over 500 followers on Twitter that do not mind its *one-every-two-minutes* scattergun approach to generation. @*MetaphorMinute* crafts metaphors by filling a template with nouns and adjectives that are chosen more-or-less at random, to produce inscrutable tweets such as "*a cubit is a headboard: stational yet tongue-obsessed*."

We chose 60 tweets at random from the past outputs of each Twitterbot. CrowdFlower annotators, who were each paid a small sum per judgment, were not informed of the origin of any tweet, but simply told that each was selected from Twitter because of its metaphorical content. We did not want annotators to actively suspend their disbelief by knowingly dealing with bot outputs. Annotators were paid to rate the content of each tweet along three dimensions, *Comprehensibility*, *Novelty* and likely *Retweetability*, and to rate all three dimensions on the same scale: *Very Low* to *Medium Low* to *Medium High* to *Very High*. Ten annotations were solicited for each dimension of each tweet, though the responses of likely scammers (non-engaged annotators) were later removed from the dataset. Tables 1 through 3 present the distributions of mean ratings per tweet, for each dimension and each Twitterbot.

Comprehensibility	***@Metaphor Magnet***	***@Metaphor Minute***
Very Low	11.6%	**23.9%**
Med. Low	13.2%	**22.2%**
Med High	**23.7%**	22.4%
Very High	**51.5%**	31.6%

Table 1. *Relative Comprehensibility of each bot*

So more than half of *@MetaphorMagnet*'s tweets were ranked as having very high comprehensibility, while less than one third of *@MetaphorMinute*'s tweets are so ranked. More surprising, perhaps. is the result that annotators found more than half of *@MetaphorMinute*'s wholly random metaphors to have medium-high to very-high comprehensibility. This Twitterbot's use of abstruse terminology, such as *stational* and *peachblow*, may be a factor here, as might the bot's use of the familiar copula container *X is Y* for its metaphors, which may well seduce annotators into believing that an apparent metaphor really does have a comprehensible meaning, if only one were to expend enough mental energy to actually discern it.

Tweet Novelty	***@Metaphor Magnet***	***@Metaphor Minute***
Very Low	**11.9%**	9.5%
Med. Low	**17.3%**	12.4%
Med High	**21%**	14.9%
Very High	49.8%	**63.2%**

Table 2. *Comparative Novelty of each bot's tweets*

The dimension *Novelty* yields results that are equally surprising. While half of *@MetaphorMagnet*'s metaphors are rated as having very-high novelty in Table 2, almost two-thirds of *@MetaphorMinute*'s tweets are just as highly rated. However, we should not be overly surprised that *@MetaphorMinute*'s bizarre juxtapositions of rare or unusual words, as yielded by its unconstrained use of aleatoric techniques, are seen as more unusual than those word juxtapositions arising from *@MetaphorMagnet*'s controlled use of attested Web n-grams and stereotypical knowledge. As shown by Giora *et al*. (2004), novelty is neither a source of pleasure in itself nor is it a reliable benchmark of creativity. Rather, pleasurability derives from the recognition of *useful* novelty, that is, novelty that can be understood and appreciated relative to the familiar.

Re-Tweetability	***@Metaphor Magnet***	***@Metaphor Minute***
Very Low	15.5%	**41%**
Med. Low	**41.9%**	34.1%
Med High	**27.4%**	15%
Very High	**15.3%**	9.9%

Table 3. *Relative Retweetability of each bot's tweets*

On Twitter, useful exploitation is frequently a matter of social reach. A tweet is novel and useful to the extent that it attracts the attention of Twitter users and is deemed worthy of re-tweeting to others in one's social circle. Our third dimension, *Re-Tweetability*, reflects the likelihood that an annotator would ever consider re-tweeting a given metaphorical tweet to others. Though we ask annotators to speculate here – neither bot has enough followers to perform a robust statistical analysis of actual retweet rates – the results largely conform to our expectations. The results of Table 3 show retweetability to be a matter of novelty *and* comprehensibility, and not just novelty alone. Though annotators are not generous with their *Very-High* ratings for either bot, *@MetaphorMagnet*'s tweets are judged to be considerably more re-tweetable than the largely random offerings of *@MetaphorMinute*.

Comprehensibility and comprehension are two different things: while a Computational Creativity (CC) version of the placebo effect may well foster a belief that a given tweet has a coherent meaning, it cannot actually provide this meaning. Meaning is the product of interpretation, and interpretation is often hard. Milic (1970) notes that in a context that licences a poetic interpretation, such as one in which a reader is told that a particular text is a metaphor, readers are more likely to accept that the text – as inscrutable as it may be – has a metaphorical meaning rather than dismiss it as nonsense. Recall that over 75% of *@MetaphorMagnet*'s tweets and over 50% of *@MetaphorMinute*'s tweets are judged as having *medium-high* to *very-high* comprehensibility. We thus need to look deeper, to determine whether raters can actually back up these judgments with actual meanings.

In a second CrowdFlower experiment, we make raters work harder, to reconstruct a partial tweet by adding the missing information that will make it whole and apt again. That is, we employ a *cloze* test format for this experiment, by removing from each tweet the pair of key qualities that anchor the tweet and make its comparison of ideas seem meaningful and apt. For *@MetaphorMagnet*, for example, we remove the properties *detailed* and *vague* in this tweet:

To some freedom fighters, freedom is a detailed *recipe. To others, it is a* vague *dream*. #Freedom=#Recipe #Freedom=#Dream

For *@MetaphorMinute*, we excise the pair of qualities *hippy* and *revisional* from the following tweet:

a flatfoot is a houseboat: hippy *and* revisional

For each tweet from each bot, we blank out a pair of original qualities as above; this pairing is the answer that is sought from human judges. We also choose 4 distractor pairs for each original pair, by selecting pairs from other tweets from the same bot. As in our first experiment, we chose 60 tweets at random from the past outputs of each bot, and 10 ratings were solicited for each. Annotators were presented with a tweet in which the key properties were blanked out (as above), and given five randomly ordered pairs of possible fillers to choose from. To make the results of the experiment comparable to those of the 1st experiment (Tables 1,2,3), we obtain the mean aptness of each tweet, so that e.g. if 7 out of 10 raters correctly choose the original pairing, then that tweet is deemed to have an aptness of 0.7. We then place these aptness scores into bands, where the *Very Low* band = 0 to 0.25, *Medium Low* = 0.26 to 0.5, *Medium High* = 0.51 to 0.75, and *Very High* = 0.76 to 1. By calculating the distribution of tweets to each band, we can determine e.g. the percentage of tweets from each bot that are put into the *Very High* band.

Our hypothesis is rather straightforward: if tweets are linguistic containers that are carefully crafted to convey a particular meaning, then it should be easier to select the missing pair of qualities that make this meaning whole; if, on the other hand, the tweet is all there is, and its content is chosen mostly at random, then raters will choose the right pairing with no more success than random selection.

The results reported in Table 4 bear out our hypothesis.

Metaphor Aptness	***@Metaphor Magnet***	***@Metaphor Minute***
Very Low	0%	**84%**
Med. Low	**22%**	16%
Med High	**58%**	0%
Very High	**20%**	0%

Table 4. *Relative Aptness of each bot's metaphors*

The placebo effect in CC can lead us to appreciate a bot's tweets as meaningful but cannot tell us what this meaning should be. Though the results above may seem a foregone conclusion, as *@MetaphorMagnet*'s tweets are designed to communicate a fully recoverable meaning while those of *@MetaphorMinute* are not, this is surely what it means to engage in real communication: to design an utterance so that an intended meaning is re-created, in whole or in part, in the mind of an intelligent, receptive audience.

Fake It 'Til You Make It

The *Placebo Effect* benefits *all* Computational Creativity systems, from superficial users of surealistic techniques to sophisticated knowledge-based AI systems. That this is so should come as no surprise, for we humans also benefit from the effects of an active and receptive mind when dealing with other people. Just as a prior belief in the efficacy of a medical intervention can lead us to perceive (and experience) a post-hoc benefit from an otherwise empty treatment, a prior belief in the meaningfulness of a verbal intervention can lead us to perceive (and enjoy) a creative meaning where none was ever intended. When a CC system uses superficial techniques to convey a sense of understanding and profundity with otherwise shallow linguistic forms, as in Weizenbaum's (1965) infamous ELIZA system, the label "*ELIZA Effect"* proves to be an apt one (Hofstadter, 1995). However, we humans are also subject to an *ELIZA effect* of our own, insofar as we often do others the courtesy of assuming their utterances to be freighted with real meaning and creative intent, and will often work hard to uncover that meaning for them.

At one time or another, we have all relied on catch-phrases, clichés, slogans, idioms, canned jokes and other half-empty linguistic containers to suggest to others that we have deeper meanings in mind, or have something more profound to offer, than we actually do. In a famous polemical essay from 1946, George Orwell excoriates speakers of English for their reliance on jargon, foreign words and empty phraseology as a substitute for thoughts of real substance, while Geoff Pullum (2003) upbraids modern speakers for a grating over-reliance on "*multi-use, customizable, instantly recognizable, time-worn, quoted or misquoted phrases or sentences that can be used in an entirely open array of different jokey variants by lazy journalists and writers.*" These "*phrases for lazy writers in kit form*" are not that different from the template-based language games played by superficial Twitterbots, and though we humans fill our templates – such as "*X is the new black*", "*In X no one can hear you scream*" or "*if the Eskimos have N words for snow then Xs surely have as many for Y*" – with lexical fillers that are contextually apt, we employ our templates to be just as provocative, and to imply or to suggest more than we actually mean.

To see machines work with humans in the construction of real figurative meanings, readers are directed to a variant of *@MetaphorMagnet* – a related bot named *@MetaphorMirror* – that tweets its own novel metaphors in response to breaking news events. This bot's metaphors are not offered as informative summaries of the news, but as figurative lenses through which followers can view the news and adopt a novel perspective on human affairs.

Acknowledgements

This research was supported by the EC project *WHIM:*

The What-If Machine. See *http://www.whim-project.eu/*

References

Thorsten Brants and Alex Franz. (2006). Web 1T 5-gram database, Version 1. *Linguistic Data Consortium*.

Jaime G. Carbonell. (1981). Metaphor: An inescapable phenomenon in natural language comprehension. *Report 2404*. Carnegie Mellon Computer Science Dept.

William Chamberlain and Thomas Etter. (1983). *The Police-man's Beard is Half-Constructed: Computer Prose and Poetry*. Warner Books.

Simon Colton, Jacob Goodwin and Tony Veale. (2012). Full-FACE Poetry Generation. In *Proc. of the 3rd International Conference on Computational Creativity*, Dublin, Ireland.

Caitlin Dewey. (2014). What happens when @everyword ends? Intersect, *Washington Post*, May 23rd edition.

Dan Fass. (1991). Met*: a method for discriminating metonymy and metaphor by computer. *Computational Linguistics,* 17(1):49-90.

Dedre Gentner, Brian Falkenhainer and Janice Skorstad. (1989). Metaphor: The Good, The Bad and the Ugly. In *Theoretical Issues in NLP*, Yorick Wilks (Ed.) Hillsdale, NJ: Lawrence Erlbaum Associates.

Pablo Gervás. (2000). Wasp: Evaluation of different strategies for automatic generation of Spanish verse. In *Proc. of the AISB-2000 Symposium on Creative & Cultural Aspects of AI*, 93-100.

Rachel Giora, Ofer Fein, Jonathan Ganzi, Natalie Alkeslassy Levi and Hadas Sabah. (2004).Weapons of Mass Distraction: Optimal Innovation and Pleasure Ratings. *Metaphor and Symbol **19**(2):115-141*.

Sam Glucksberg. (1998). Understanding metaphors. *Current Directions in Psychological Science,* 7:39-43.

Joseph Grady. (1997). Foundations of Meaning: Primary Metaphors and Primary Scenes. University of California.

Douglas Hofstadter. (1995). The Ineradicable Eliza Effect and Its Dangers. *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought* (Preface 4), Basic Books: New York.

James Hutton (*translator)* (1982). *Aristotle's Poetics*. New York, NY: Norton.

George Lakoff and Mark Johnson. (1980). *Metaphors We Live By*. Chicago, Illinois: Chicago University Press.

James H. Martin. (1990). *A Computational Model of Metaphor Interpretation*. Academic Press.

Ruli Manurung, Graeme Ritchie and Henry Thompson. (2012). Using genetic algorithms to create meaningful poetic text. JETAI 24(1):43–64.

Matthew S. McGlone and Jessica Tofighbakhsh. (1999). The Keats heuristic: Rhyme as reason in aphorism interpretation, *Poetics* **26**(4):235-44.

Matthew S. McGlone and Jessica Tofighbakhsh. (2000). Birds of a feather flock conjointly (?): rhyme as reason in. aphorisms. *Psychological Science* **11** (5): 424–428.

Louis T. Milic. (1971). The possible usefulness of computer poetry. *The Computer in Literary and Linguistic Research*, R.A. Wisbey (Ed.), Cambridge, MA.

Geoffrey Pullum. (2003). Phrases For Lazy Writers in Kit Form. *Language Log post*, October 27, 2003.

George Orwell. (1946). Politics and the English language. *Horizon,* 13(76), April issue.

Michael J. Reddy. (1979). *The conduit metaphor: A case of frame conflict in our language about language.* In A. Ortony (Ed.), *Metaphor and Thought, 284–310.* Cambridge University Press.

Christopher B. Ricks, (1980). Clichés. In: L. Michaels and C. Ricks (Eds), *The State of the Language.* University of California Press, Berkeley.

Tony Veale and Mark T. Keane. (1992). Conceptual Scaffolding: A spatially founded meaning representation for metaphor comprehension. Computational Intelligence 8(3):494-519.

Tony Veale and Mark T. Keane. (1997). The Competence of Sub-Optimal Structure Mapping on 'Hard' Analogies. In *Proceedings of IJCAI'97, the 15th International Joint Conference on Artificial Intelligence.* Nagoya, Japan. Morgan Kaufmann.

Tony Veale and Guofu Li. (2011). Creative Introspection and Knowledge Acquisition. In Proc. of AAAI-2011, *The 25th Conference of the Association for the Advancement of Artificial Intelligence.* San Francisco: AAAI Press.

Tony Veale and Guofu Li. (2013). Creating Similarity: Lateral Thinking for Vertical Similarity Judgments. *In Proceedings of ACL 2013, the 51st Annual Meeting of the Assoc. for Computational Linguistics, Sofia, Bulgaria,*

Tony Veale. (2013). Less Rhyme, More Reason: Knowledge-based Poetry Generation with Feeling, Insight and Wit. *In Proc. of ICCC 2013, the 4th Int. Conference on Computational Creativity. Sydney, Australia.*

Tony Veale. (2014). Running With Scissors: Cut-Ups, Boundary Friction and Creative Reuse. In Proceedings of *ICCBR-2014, the 22nd International Conference on Case-Based Reasoning.*

Eileen Cornell Way. (1991). *Knowledge Representation and Metaphor: Studies in Cognitive systems*. Kluwer.

Joseph Weizenbaum. (1966). ELIZA – A Computer Program For the Study of Natural Language Communication Between Man And Machine. *Communications of the ACM* **9** (1): 36–45.

Yorick Wilks. (1978). Making Preferences More Active. *Artificial Intelligence* 11(3):197-223.