The Soul of a New Cliché:
Conventions and Meta-Conventions in the Creative Linguistic Variation of Familiar Forms

Tony Veale
Web Science and Technology Division, KAIST, Yuseong, Daejeon, Korea
Tony.Veale@gmail.com

Abstract

Creativity – whether in humans or machines – is more than a matter of *creation*. To be "creative" implies an ability to do more than invent, but an ability to recognize and appreciate the inventions of others. After all, the ability to recognize surprising value in the efforts of others is the same ability we use to guide our own creative efforts. Solipsistic creativity is rare indeed, and most creativity relies on an audience that is creative enough to value our efforts. Of what value is an ability to e.g. speak ironically if we cannot also understand or appreciate the irony of others? The goal of imbuing computers with creative abilities must thus include a sub-goal of enabling computers to recognize and respond appropriately to the creativity of others. As computers are increasingly used to analyze the burgeoning texts of the world-wide-web, the ability to automatically detect and analyze the linguistic creativity of speakers has become more important than ever. In this paper we consider how speakers engage creatively with cliché, to achieve creative ends through the novel variation of familiar linguistic forms. Our computational analysis of a large collection of linguistic patterns on the Web shows that speakers are surprisingly conservative in their variation strategies, and novelty alone rarely leads to creativity. This conformity can make it easier for computers to detect when speakers are using familiar language in truly original ways.

1 Introduction

Samuel Goldwyn, the co-founder of MGM studios, famously summed up Hollywood's attitude to creativity with the line "Let's have some new clichés". On the face of it, this seems like just another one of Goldwyn's many memorable misstatements (like "include me out!"): after all, it's hard to think of clichés as *new*, or as something that can be invented on demand. Yet, on closer analysis, one can find real insight in Goldwyn's remark. Clichés are considered anathema to the creative process because they represent everything that is conventional and jaded about the status quo. However, clichés become tired thru overwork, and are overworked precisely because they prove themselves so useful in so many different contexts. Few creators set out to create a new cliché, but most would like their efforts to become as much a part of the fabric of our culture as the most tenacious of clichés.

Nonetheless, cliché is generally derided for its baleful effect not just on language, but on thought. George Orwell, in a much-quoted polemic from 1946 (*Politics and the English Language*), poured scorn on two particular forms of clichéd language: the expedient use of familiar metaphors that have lost their power to evoke vivid images; and the use of readymade turns of phrase as substitutes for individually crafted expressions. Rather than bend words to their meanings, Orwell worried that clichés entice lazy writers to bend meanings to their words. He derided the over-use of readymade phrases "tacked together like the sections of a prefabricated henhouse", and fretted that any writer who operates by “mechanically repeating the familiar phrases” is simply “gumming together long strips of words which have already been set in order by someone else”.

Orwell offers a typically monochromatic view of clichés that accentuates the negatives and overlooks the positives. In dismissing clichés as *flyblown*, *jaded* or *over-worked*, he himself succumbs to what Christopher Ricks (1980) calls the language of *cliché-clichés*. The critic William Empson admired Orwell’s eye for cliché, but rejected his proscriptive approach, memorably calling Orwell “the eagle-eye with the flat feet”. As Empson demonstrates with this marvelous combination of two old tropes, clichés are simply lexical resources, like words, and their creative value (or lack thereof) lies entirely in how they are used.

The linguistic Web is home to all the clichés of language, but it is also a space in which speakers feel free to vary these clichés to suit their own needs. New variations evolve quickly on the Web, and new stereotypes arise to anchor these variations in a shared knowledge of popular culture. In this paper we look at one particular pattern of new clichés, the XYZ construct (see Veale, 2012), which allows a speaker to figuratively describe an X from the domain Z in terms of an apt vehicle Y. XYZ metaphors allow speakers to turn a well-known concept into a concise and vivid descriptor, as in “*The telegraph was the internet of the 19th century*.”

We view the linguistic Web as a corpus from which we can harvest a large collection of figurative XYZ instances, and thereby study in microcosm the conventions and meta-conventions of everyday linguistic creativity. For easy retrieval from the Web, we focus on XYZ instances in which the Y field is a proper-named individual. Our analysis of these figurative XYZ instances will allow us to explore the ways in which speakers obey, and sometimes transcend, semantic and pragmatic conventions when seeking to use language creatively.

2 Computers and Creative Linguistic Variation

There is a popular misconception that creativity is a matter of *breaking the rules*. If this were the case, there could be no “fair” creativity in chess, in sports, or in any rule-defined context, yet we see many admirable instances of creativity in each of these contexts. It is more accurate to say that creativity is matter of exploiting or subverting conventions (see Hanks, 1994), and because many of our conventions are so deeply entrenched, we unquestioningly view them as rules (see Veale, 2012). For a computer to be creative, it must have a knowledge of these conventions, not at a hard-wired level where they are coded as rules, but at a knowledge-representation level where they can be modified and manipulated. To attain this level of knowledge, it is best if a computer acquires an understanding of conventions (and their limitations) for itself, rather than have them hard-coded by a programmer. Creative people often exploit an expectation gap between themselves and their audiences: they see orthodoxy as a set of conventional *norms* that can be challenged (Hanks, 1994; Veale & Hao, 2010), while their audiences expect these norms to be obeyed as rules.

Creativity in language is largely a matter of playing with conventions, to establish expectations in the minds of an audience that are then dashed, in ways that surprisingly add rather than subtract meaning. Our analysis here reveals that while creative variation is pervasive in language, speakers tend to be semantically conservative in how they combine their clichés and stereotypes, so that very few expectations are undermined. Such linguistic variation can yield novelty, but yields little in the way of lasting value or creativity.

These findings thus suggest several reasons to be optimistic about the ability of computers to detect and appreciate the creative language of others (and, in turn, to generate language of comparable creativity themselves). The first reason is that creativity is surprisingly dependent on a robust knowledge of cliché and stereotype. Before humans or computers can be creative with language, they first require a firm grasp of the conventional ways in which speakers are *un*creative with language.

The necessary knowledge can be easily harvested from everyday language, as e.g., found in abundance on the Web (see Veale, 2011;2012; Veale & Hao, 2007). The second reason is that the ways in which people playfully combine stereotypes or vary the contents of familiar forms can also be learned from observing how speakers achieve these variations in well-defined contexts, such as e.g. in similes and in figurative XYZ constructions. These are meta-conventions for being creative with language, and these too can be learned from careful observation of human speakers. The third reason is that speakers often mark their attempts at creativity through their use of linguistic support structures. Some structures are as subtle as the addition of a hedge marker like "about" or "not exactly", while others are as overt as the addition of an explanation. These structures help to convey the meaning of a linguistic novelty even when the underlying conceit (such as a humorous metaphor or ironic viewpoint) falls flat.

3 Related Work and Ideas

The linguistic Web is a vast reservoir of opinions and beliefs, as well as a fertile breeding ground for new stereotypes and for new variations on familiar forms. The Web can thus be used as a corpus (see Keller and Lapata, 2007) from which general observations about language can be derived, which may support specific observations about linguistic creativity.

When used as a language corpus, the Web has some noteworthy qualities, beyond the obvious benefits of scale. For one, it is a dynamic corpus, always changing to reflect new additions and new users. When one uses the Web as a corpus, one is using the most up-to-date corpus available. For another, the linguistic Web has a variable resolution. If one finds a phenomenon of interest, one can always go back for more examples, or wait for more to occur naturally. Thus, one can work with a sample of the Web, downloaded locally, and extend this sample dynamically as the need arises. One can also use the linguistic Web as a bridging corpus, to dynamically fill in the gaps in a more conventional corpus.

Of course, care must be taken when using the linguistic Web as a corpus, not least because this is not the principal purpose of the Web, and those who create it and provide access to it have no responsibility to provide the balance that linguists usually require from a corpus. As noted in Kilgarriff (2007), the hit counts provided by search engines for a given query do not have the same authoritativeness as statistics derived from a corpus compiled by linguists. Generally speaking, one should be skeptical of absolute page counts, and work instead with relative measures where possible.

Despite these caveats, the Web is a marvelous source of language users and of language data. When the folklorist Archer Taylor compiled his landmark study of proverbial similes in 1954, his study required the laborious collection of thousands of similes from first-hand sources. When Neal Norrick published his analysis of stock similes in 1986, he relied instead on a set of 366 similes from the 1970 edition of *The Oxford Dictionary of Proverbs*. Rosamund Moon based her 2008 study of similes on a collection of 377 stock phrases that she found in multiple corpora. In contrast, Veale and Hao's 2007 study of stereotype-bearing similes harvested over 12,000 unique simile instances the Web, from which a rich knowledge-representation of concepts and their properties is automatically gleaned. When Veale (2011, 2012) later extended this approach to use a combination of Google n-grams (see Brants and Franz, 2006) and automated queries to the Web (n-grams provide specific hypotheses from which targeted queries are produced), he extended this corpus of similes to over 75,000 unique instances. Clearly, for studying phenomena like creative language use, the linguistic Web is the most comprehensive source of real data that one can use.

For instance, in her 2008 corpus study, Moon argues that "about" has a special role in signaling irony when used to introduce a simile, as in "*about* as crazy as a fox". Yet the number of *about-as*-similes identified in Moon's corpus appears to be too small to support a reliable analysis. Though

Moon's intuition is largely correct, Veale's (2012) analysis of 1000s of *about-as*-similes from the Web shows that "about" is more generally a marker of creative intent, one that can be used non-ironically to sharpen an already overt put-down. The "about" marker reliably signals irony only when a simile would otherwise imply a positive evaluation of a topic. Hao and Veale further show that a simile is only somewhat likely to be ironic when it can also be found on the Web in the *about-as*-form, but it is very likely to be ironic if the *about-as* form is its dominant form on the Web. Hao and Veale (2010) use this observation of Web usage patterns to achieve good results in an automated irony detector for similes. In other words, a computer can exploit a tacit conformity on the part of speakers – e.g., to prefix ironic similes with the hedge "about" – to recognize when those speakers are being playful and/or ironic.

Roncero, Kennedy and Smith (2006) also looked to the Web in their comparative study of similes and metaphors, to see which trope is most likely to be accompanied by an explanation. They found that similes are much more likely than metaphors to need an explanation on the Web. One might imagine that the less conventional the vehicle of a figurative comparison (as in "genes are like *blueprints*"), the more likely it is that the comparison is accompanied by an explanation of its meaning. Yet these researchers also demonstrated a strong correlation between the presence of an explanation and the conventionality of a vehicle. It seems that when one uses a conventional metaphor, such as "time is money", it is with the intent of communicating its accepted meaning. However, when one introduces the hedge "like" (e.g. "time is *like* money"), this signals an intent to stray beyond the accepted meaning, to explore less salient aspects of the vehicle. Because explanations are far from unusual in creative similes, users are free to compose novel comparisons, or to develop novel variations on a conventional simile, without risk of misinterpretation. This suggests a division of labor in creative similes: the simile itself constructs a vivid and memorable image, while the explanation anchors the image in the larger point the user wants to make. Once again, superficial form gives the listener a valuable insight into the speaker's creative intent.

Ironic similes are much less likely to come with explanations, since ironic statements function as jokes, and jokes are undermined by the need for explanation. However, an ironic simile often illustrates a larger point, which may then be elaborated, as in this blog post: "[Robert] Gibbs style as Press Secretary has been *about as soothing as a Gilbert Gottfried monologue*. He's been abrasive, condescending, and elusive in his answers to simple questions put forth to him by the press."

Proper-named individuals like *Gilbert Gottfried* can be suggestive of very specific qualities and behaviors, and can thus allow a speaker to vividly personalize a simile (e.g. Gottfried is famous for his whiney, grating delivery). Veale (2012) found that this strategy is used in a significant number of cases, so that approx. 12% of *about*-similes on the Web use a well-known figure from popular culture that has assumed the status of a popular stereotype, as in "about as lost as *Paris Hilton* in a library" or "about as frustrated as *Stevie Wonder* in an Easter egg hunt". The idea that Paris Hilton is less than bright is a cliché that abounds on the Web, and the uncreative simile "as dumb as Paris Hilton" is found over 5,000 times with Google. Yet the "lost in a library" variant above shows that speakers often vary clichés to achieve novel and creative ends.

Paul Kay (2002) refers to linguistic structures like the *as*- and *about-as*- simile forms as *patterns of coining*. These are schematic structures that can be used for creative ends, but any creativity arises wholly from how the structure is instantiated in each case. As such, patterns of coining are much more than the "phrases for lazy writers in kit form" derided by the linguist Geoff Pullum (2003) – such as "[X] is the new black" – since these phrasal patterns are not schematized allusions to popular phrases and clichés. Nonetheless, these patterns do allow a speaker to coin a phrase that, if popular enough, may become a cliché in its own right. In the following sections, we shall explore the uses of one particular pattern of coining– the figurative XYZ construction – to understand the ways in which speakers vary the contents of familiar forms to achieve novelty and creativity, and even give birth to "new clichés".

4 Birth of a Cliché

The XYZ construction can be used either for literal or figurative purposes. Naturally, literal uses – such as “David Cameron is the prime minister of Britain” – are not generally considered creative, while the creativity of figurative uses depends on a range of different factors, from originality/unexpectedness to concision to semantic and pragmatic incongruity. Moreover, since both literal and figurative uses of the construction have the same XYZ structure, there are few syntactic cues as to whether a particular instance is potentially creative. However, since the Y field of an XYZ should denote an ad-hoc category of entities into which the X is placed, any XYZ in which Y denotes an obvious non-category, such as a proper-named individual, is highly likely to be figurative. Thus, XYZs such as “red meat is the Donald Trump of cancer” and “the potato is the Tom Hanks of the vegetable world” are figurative examples of the form (and more-or-less creative examples at that).

In conceptual terms, a figurative XYZ uses a structural analogy to pinpoint the relative position of X in the Z domain, by noting that Y has the same relative position in its (unstated) domain. For instance, “Tom Hanks” is most famous as an actor, so his unstated domain is likely that of *acting*. Hanks is a prominent and respected member of this domain, where he is often lauded for his versatility. By implication then, this XYZ asks us to view the potato as a prominent and much-loved member of the vegetable world, and to give particular attention to its versatility. Since the Y component must evoke an implicit domain and an implicit rationale for the comparison, speakers understandably use Y’s that have strong stereotypical associations. Hanks is a good choice for Y here since he has become a popular stereotype for versatility, as attested by the range of XYZs in which he occurs on the Web.

The creativity of this example arises from a number of different factors. For one, the comparison of a non-sentient vegetable to a respected actor is semantically and pragmatically incongruous, even if one has a low opinion of actors. Hanks is used in a large number of figurative XYZs on the Web, but many use Hanks to describe another person in the entertainment industry. Instances like “Aamir Khan is the Tom Hanks of Bollywood” score highly for concision, as they concisely allude to the target’s proficiency in multiple genres (romance, drama, comedy) and on multiple levels (acting, producing, directing), but lack the wit that often arises from a clash of mutually-exclusive categories. For another, the XYZ is flattering of potatoes but is disrespectful of Hanks (who is equated with a vegetable). The dual meaning of “vegetable” lends this example a sharpness that is absent from a more common variation on the Web: “ketchup is the Tom Hanks of the food world”.

The comparison also has a riddle-like quality, challenging the listener to decipher and unpack its implicit meaning. Yet the comparison is not so challenging that it requires an explanation. In part, this is due to our strong stereotypical conception of Hanks and in part due to the simplicity of its meaning: only a single property *versatile* is transferred from Y to X. Contrast this with our earlier XYZ (from a Vegan blog): “red meat is the Donald Trump of cancer.” Trump lends a convenient face to the stereotype of the aggressive developer, one who is driven by profit rather than the social good, and one who builds wherever there is money to be made, despite the objections of local residents. The original blog post explains that since red meat has been implicated in the development of many different kinds of cancer, it can also be metaphorically categorized as an aggressive and opportunistic builder of cancers.

One must also take a diachronic view of an XYZ instance, and its variations, to properly assess its level of creativity. The first usage of *Tom Hanks* as a figurative short-hand for versatility has a greater claim to the label “creative” than its 1000th usage. As the stereotype becomes more established – e.g., by repeated usage on the Web – it becomes a less original basis for a creative word choice. While the XYZ form remains a support structure for creative coinages, the partially instantiated skeleton “X is the Tom Hanks of Z” becomes just another instance of Pullum’s “phrases for lazy writers in kit form”. This is the natural and inevitable course of a figurative conceit, and mirrors what Brian Bowdle and

Dedre Gentner (2005) call *the career of metaphor*. These authors suggest that a novel metaphor is originally understood using analogical mapping and other structural reasoning processes, but as the metaphor is re-used and becomes more familiar, it is understood as a simple act of categorization. In other words, the figurative conceit becomes a conventional category, a familiar form rather than a novel departure.

The Web proves to be a fertile ground for spawning new variations of popular XYZ forms. In addition to the use of "X is the *Tom H*anks of Z" to denote versatility in the Z domain, one can find many different variations on the form "X is the *Kenny G* of Z". This latter form tends to be used unflatteringly, and somewhat snobbishly, to describe a target X who has achieved some measure of undeserved fame in domain Z. Variations range from "Eric Clapton is the Kenny G of blues" (a close musician-to-musician mapping) to "Riesling is the Kenny G of wines" (a distant person-to-product mapping). In the *Kenny G* case, the quality that is projected onto X is not easily glossed with a single word (such as "versatility"). In some cases, *G* is invoked as a stereotype for smoothness, sugariness or blandness, but in others he is used to suggest a more complex assessment, along the lines of "X has obvious technical merits, but lacks sufficient depth, purity and distinctiveness to be truly worthy of acclaim".

The first indexed Web use (via Google) of the "X is the *Kenny G* of Z" form dates from 1996/1997, when G was used to describe another jazz musician, Herbie Mann. This is also the first Web use of the *Kenny G* stereotype as it now widely used on the Web. The article motivates the metaphor thus: "the purists, who are very territorial about what they perceive is their music, hate it when anybody sells more and has a bigger public". In 1999, CNN online describes the Price Waterhouse Coopers as the "Kenny G of Accounting Firms", in an article that implies that any firm would be embarrassed by the comparison. By 2000, the stereotype was used to describe writer and critic Susan Sontag as "the Kenny G of literature". In the following years, the popular stereotype of Kenny G is stretched to describe other musicians working in other genres (Blues, R&B, Bebop, Country), instruments (drums, guitar, piano, violin, sitar) and regions (Jamaica, India, The Philippines). While music-domain uses of the stereotype still predominate, the stereotype is broadened in these years to other domains (comedy, film, cuisine).

The *Kenny G* stereotype is a shared but tacit element of popular culture, one that is exploited widely on the Web but explicitly defined nowhere on the Web. At the time of writing. neither *Urban Dictionary* nor *Wikipedia* provide an entry for "Kenny G" that explains its accepted meaning in so many online comparisons. In a sense, the stereotype resides in a distributed fashion in all of its uses on the Web. New variants arise as Web users adapt an existing form to suit their own descriptive needs. The more variants that emerge, the easier it becomes to create new variants, but the bigger the challenge faced by writers who want to do something creative with the stereotype.

5 X is the Y of Z, More or Less

A broad sample of figurative XYZ coinages with a proper-named Y field can be acquired from the Web, though not without using an equally broad range of Web queries. Popular search engines like Google offer the broadest and most up-to-date coverage, but their query languages lack the finesse of dedicated corpus-processing tools. Moreover, commercial search engines may rank their results by popularity, or authoritativeness, or even by their perceived relevance to a paying advertiser, but not by diversity: a simple query like *"* is the * of *"* will not only match a great many phrases that are not figurative XYZs, but will also retrieve a great many duplicates, overlooking interesting one-offs that are coined in little-visited blogs or web-sites. The remedy is to dispatch an enfilade of specific, partially-instantiated queries rather than a single, over-arching query. We should thus identify the most useful Y's for coining XYZs, and formulate queries for the corresponding partially-instantiated patterns, such as "** is the Kenny G of *"*.

The Google database of Web n-grams is a large collection of the most common English word sequences on the Web. Each n-gram contains between 1 and 5 words, is case-specific (allowing for orthographic matching), and has a minimum Web frequency of 40 documents. Looking to the Google 4-grams, we can thus identify all matches for the pattern "the *Fname Lname* of", where *Fname* and *Lname* match any capitalized words that might plausibly be used as the first and last parts of a proper-name. A large list of allowable name elements is harvested from *Wikipedia* for this task.

Likewise, a collection of evocative one-word names (such as *Mozart*, *Einstein* and *Napoleon*) is used to find matches in the Google 3-grams. The combined matches for 3-grams and 4-grams provide instantiations for the proper-named Y component of each XYZ query, allowing us to automatically dispatch a corresponding Google query for each specific Y. In return, Google provides a set of up to 200 matching text snippets for each query, such as the snippet in (2), which contains an extract from the *Sunday Times* that is retrieved for the query "*is the Picasso of":*

> "*Ferran is the Picasso of the modern kitchen*," enthuses Rafael Anson, president of the International Academy of Gastronomy.

An automated filter is then applied to each snippet to identify those that contain well-formed X, Y and Z components, and to exclude those whose X is a pronoun like "he" or "it", as well as those whose Z component does not conform to any of a set of simple patterns (such as a bare noun, like "crime" or "wines", a temporal or geographic specifier, like "European Union" or "20th Century", or a phrase of the form "the Z", "the Z world", "the Z domain", "the Z genre" or "the Z industry"). In all, a series of over 3000 queries harvests more than 60,000 snippets from Google. Subsequent filtering pares this set down considerably, to yield a corpus of 2190 unique XYZs with 668 different Y's.

As expected, the most frequent Y's are all prominent individuals whose propensities and abilities are well established in popular culture and on the Web. The 20 most frequent Y's account for about 10% of all XYZs in the corpus, and the individual-types in this top 20 are broadly representative of the whole, ranging from historical figures to fictional figures to artists, musicians, actors, politicians and sportsmen. These individuals are ciphers for some commonly ascribed properties on the Web: *Benedict Arnold*, for instance, stands for any person who is traitorous or just plain fickle about which side to support; *Rush Limbaugh* can stand for any political loud-mouth with partisan views; and *Chuck Norris*, an expressionless actor who is lampooned relentlessly on the Web, can stand for any person or thing that is equally uncomplicated and uncompromising in its behavior.

Used to anchor 21 XYZs, *Michael Jordan* is the most commonly occurring Y in our corpus. As an athlete who resided at the pinnacle of his sport, Jordan has become a role model for strivers in any sport, and some non-sports besides. Here are the 21 X's that are compared to Jordan in our corpus (with the corresponding Z's in parentheses):

> *Manny Pacquiao* (Philippines), *Andrew Gaze* (NBL), *Chet Snouffer* (boomerang), *Garry Kasparov* (chess), *Mwadi Mabika* (WNBA), *Vince Young* (NFL), *Pádraig Harrington* (golf), *Tiger Woods* (golf), *Randy Couture* (martial arts), *Daryll Pomey* (Philippines), *Tony Hawk* (skateboarding), *Champ Hallett* (wheelchair basketball), *David Berg* (courtroom), *Bronwyn Weber* (cakes), *Michael Chabon* (literary), *the tuna sandwich* (mid-day meal), *Billy Bob Thornton* (movies), *Ralph Appelbaum* (museums), *Allan Bloom* (seminars), *Britney Spears* (pop), *Randall Ross* (rare books)

Some X's are closer to Jordan than others, but most are themselves sportsmen. The closest are basketball players from other leagues (the NBL) or of another gender (e.g., Mwadi Mabika of the

WNBA). Yet Jordan is a model of excellence for any competitive endeavor, and has leadership qualities that even tuna lovers can apparently find inspirational.

The corpus contains 1312 different Z's, once support words like "world", "genre", "domain" and "industry" have been stripped out. However, the top 20 most frequent Z's account for 367 different XYZs, or more than 16% of the whole corpus. We observe significant diversity in this top tier, which includes times (the *21st* and *20th centuries* are most popular), geographic locations (*North, South, East* and *West* are all in the top 20 Z's), sports and sports organizations (the *NBA*, *NFL* and *NHL*), politics and political parties (the *Republicans,* the *GOP,* the *Democratic Party*) and political orientations (*Left* versus *Right*), as well as art, music and gaming. Overall, the most frequent Z, *21st Century*, accounts for 66 different XYZs in the corpus.

6 Semantic Preferences, Ontological Choices

In this Web sampling of XYZs, we see musicians compared to musicians, writers to writers, artists to artists, athletes to athletes and businessmen to businessmen. Putting the most creative cases to one side for a moment, we can see a strong tendency to be ontologically conservative, with X's and Y's seemingly drawn from a somewhat narrow range of ontological categories To quantify the degree to which this intuition characterizes the corpus as a whole, we need to understand the domain-to-domain character of each XYZ.

We therefore annotate each X and Y with one or more domain labels, each denoting a context of human activity. Any annotation scheme should be, *a priori*, sensible and well-motivated, while any observations arising from the resulting analysis should, *a posteriori*, not be attributable to an artifact of the tagging scheme. An exploratory pass through the data leads us to choose a system of 13 distinct domains for annotating the X's and Y's in our Web XYZs: *Politics, Music, Art, ShowBiz, Military, Crime, Business, Religion, Sport, Comedy, Culture, Drama, Science*.

This set does a good job of capturing the diversity of the corpus with an acceptable level of generality. *Business*, for instance, covers the worlds of commerce, finance and industry, while *ShowBiz* covers real individuals who perform on TV, on stage, or in movies. In contrast, *Drama* is used to annotate fictional characters who appear in movies, books or other narrative forms. *Culture* is used to annotate individuals that represent different ethnic or social groups, as well as historical figures that contribute to our understanding of a particular culture.

The relative distribution of these semantic annotations for X's and for Y's is shown in Figure 1. *Politics* is the most popular annotation for both components of our XYZs (22% as a target of description in the X position vs. 18% as a vehicle of description in the Y position). The distribution of other annotations shows just slight variation between X's to Y's.

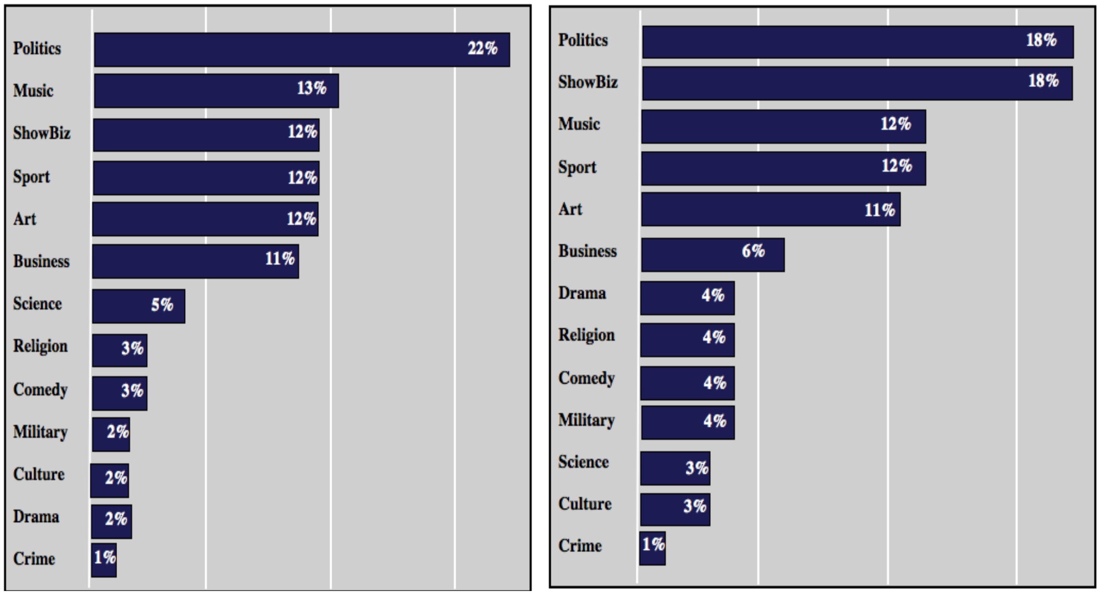


Figure 1. Left: Distribution of domain annotations for the X fields of our Web corpus of XYZs. 21% of XYZs describe a target X from the most popular target domain, *Politics*.
Right: Distribution of domain annotations for the Y fields of our Web corpus of XYZs. 18% of XYZs use a Y vehicle from the most popular vehicle domain, *Politics.*

These figures show that *Politics* is the most popular domain from which to draw both the X (21% of cases) and the Y (18% of cases) in figurative XYZs on the Web. It is worth looking at domain in greater depth. Figure 2 shows the distribution of annotations for X when the Y component of an XYZ is annotated as belonging to the *Politics* domain.

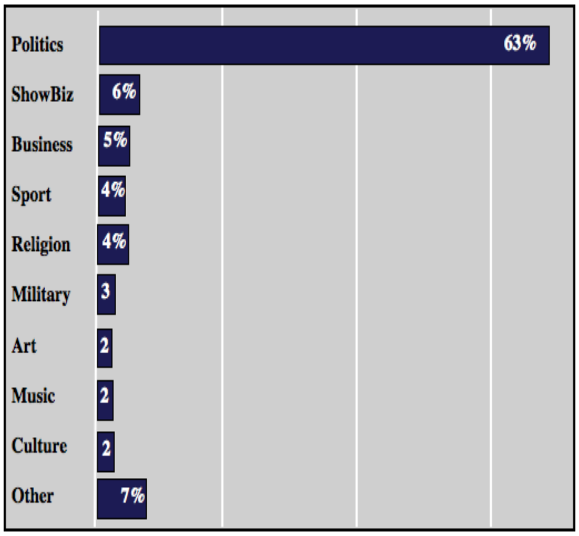
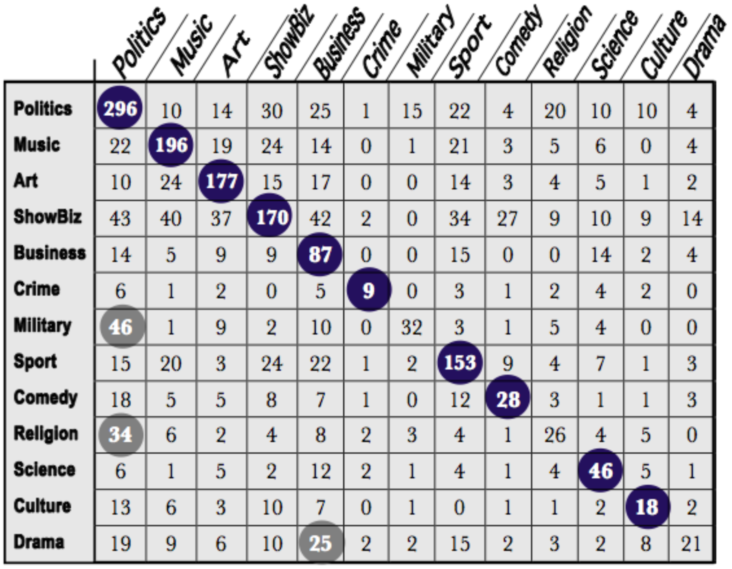


Figure 2. Distribution of domain annotations for the X fields of our Web corpus of XYZs when the Y field is drawn from the domain *Politics.*

Politics appears to be a conceptually incestuous domain, where political vehicles, like *Trotsky* and *Tony Blair*, are predominantly used to describe political targets, like *Newt Gingrich* and *David Cameron*. One can ask whether this apparent domain conservativity – wherein X's and Y's tend to come from the same domain – is a feature of the corpus as a whole, or whether *Politics* is a special case. It is to this topic that we turn our attention next.

7 Conservativity in Linguistic Creativity

The matrix in Figure 3 provides a breakdown of all domain-to-domain mappings in our corpus of XYZs. Each row in Figure 3 corresponds to a different domain as used to annotate a Y, while each column presents the distribution of domains for the X's it is used to describe. Note, for instance, that the intersection of the row *Business* and the column *Sport* contains the number 15: this indicates that, in 15 of the XYZs where the Y component is annotated with *Business*, the X component is annotated with *Sport*. In other words, 15 of the *Business* XYZs in our corpus can be understood as instances of a *Sport-is-a-Business* metaphor.



	Politics	Music	Art	ShowBiz	Business	Crime	Military	Sport	Comedy	Religion	Science	Culture	Drama
Politics	296	10	14	30	25	1	15	22	4	20	10	10	4
Music	22	196	19	24	14	0	1	21	3	5	6	0	4
Art	10	24	177	15	17	0	0	14	3	4	5	1	2
ShowBiz	43	40	37	170	42	2	0	34	27	9	10	9	14
Business	14	5	9	9	87	0	0	15	0	0	14	2	4
Crime	6	1	2	0	5	9	0	3	1	2	4	2	0
Military	46	1	9	2	10	0	32	3	1	5	4	0	0
Sport	15	20	3	24	22	1	2	153	9	4	7	1	3
Comedy	18	5	5	8	7	1	0	12	28	3	1	1	3
Religion	34	6	2	4	8	2	3	4	1	26	4	5	0
Science	6	1	5	2	12	2	1	4	1	4	46	5	1
Culture	13	6	3	10	7	0	1	0	1	1	2	18	2
Drama	19	9	6	10	25	2	2	15	2	3	2	8	21

Figure 3. Mappings of Y domains (rows) to X domains (columns). Each row is a different semantic domain for a Y, each column a different semantic domain for an X, where each cell contains the number of XYZs with this combination of X and Y domains.

The largest value in each row in Figure 3 is highlighted with a dark circle. With the exception of *Politics-as-Military* (46 *Military* metaphors), *Politics-as-Religion* (34 *Religious* metaphors) and *Business-as-Drama* (25 *Drama* metaphors), note how these dark highlights occur mainly on the diagonal, indicating that domain conservativity is a strong convention of figurative uses of the XYZ pattern.

It turns out that these XYZs are also very conservative in the more traditional reading of "conservative", for they are remarkably male-centric too, with very few female concepts on either side of the equation. Figure 4 shows the breakdown of XYZs by gender on an X-to-Y basis.

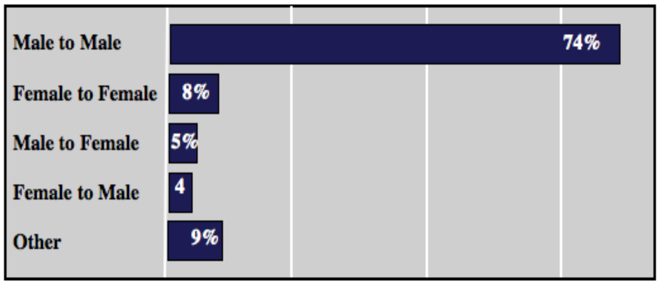


Figure 4. Breakdown of Y-to-X mappings by gender of the mapped entities. '*Other'* denotes expressions where either the source or the target has no obvious gender.

Though just 9% of our figurative XYZ samples involve cross-gender mappings, there are more females in these comparisons than there are in the pure female-to-female cases. Figurative comparisons can bridge large semantic gaps between domains, but gender appears to be a bridge too far for most figurative XYZ comparisons.

Most of our XYZ samples involve individuals from the real world, and very few exploit fictional characters in either X or Y fields. However, in XYZs that do draw upon the world of fiction, fictional Y's are much more common than fictional X's. Fictional characters often instantiate heroic (or anti-heroic) archetypes, and this appears to make them well-suited to the Y field of a figurative XYZ.

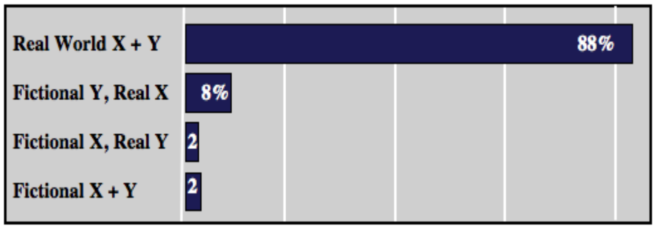
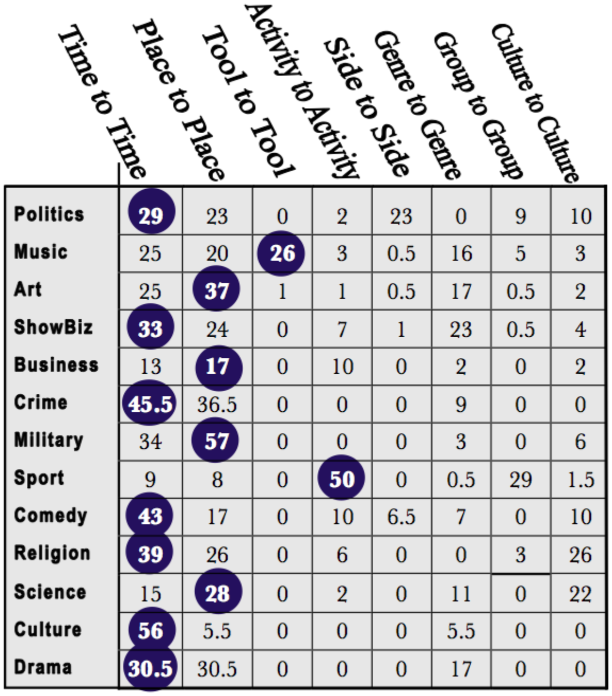


Figure 5. Breakdown of X and Y fields by fictional status. An XYZ comparison of entities in the real world is the norm.

An example XYZ that describes a real person in terms of a fictional entity is "*Jann Wenner is the Charles Foster Kane of the baby boomers*". Wenner, the founder and publisher of *Rolling Stone* magazine, might just as easily be compared to William Randolph Hearst, the real-world newspaper magnate on which *Citizen Kane* was based. More flatteringly, our corpus reveals that Warren Buffet is considered by some to be "*the Sherlock Holmes of the stock market*", while right-wing FOX news host Glen Beck has been called "*the Homer Simpson of the airwaves*". Since the same heroic archetypes continually resurface in popular culture, one fictional character will sometimes be compared to another, as in "*Allan Quatermain is the Indiana Jones of the Victorian age*" and "*Jack Sparrow is the Han Solo of the Caribbean*". However, as shown in Figure 5 above, only a small percentage of Web XYZs employ a fictional X *and* a fictional Y.

If we focus on *intra*-domain XYZ mappings only – such as *music to music*, or *politics to politics* – we can characterize the specific nature of the Z-to-Z mapping in each case. Figure 6 breaks down our sample into the following kinds of mapping: in *Time to Time* mappings, the Z field denotes a time period (e.g. *20th Century*); in *Place to Place*, Z denotes a geographical location (e.g., *The Philippines*); in *Tool to Tool*, Z denotes a tool or implement (like *guitar*); in *Activity to Activity*, Z denotes the kind of activity at which one can excel (like *basketball*, or *internet commerce*); in *Side to Side*, Z denotes either the *left* or the *right* of the political spectrum; in *Genre to Genre*, Z denotes a domain of creative expression (like *Jazz, Modernism*, *Comedy*); in *Group to Group*, Z denotes an organization or coherent group (like *the NBA*, *the NFL*, *the GOP*); and in *Culture to Culture* mappings, Z denotes a human cultural grouping (like *Islam*, *the W*est). Thus, 26% of music domain XYZs use a *Tool to Tool* mapping (e.g., *John Mayer is the Kenny G of the guitar*).



	Time to Time	Place to Place	Tool to Tool	Activity to Activity	Side to Side	Genre to Genre	Group to Group	Culture to Culture
Politics	29	23	0	2	23	0	9	10
Music	25	20	26	3	0.5	16	5	3
Art	25	37	1	1	0.5	17	0.5	2
ShowBiz	33	24	0	7	1	23	0.5	4
Business	13	17	0	10	0	2	0	2
Crime	45.5	36.5	0	0	0	9	0	0
Military	34	57	0	0	0	3	0	6
Sport	9	8	0	50	0	0.5	29	1.5
Comedy	43	17	0	10	6.5	7	0	10
Religion	39	26	0	6	0	0	3	26
Science	15	28	0	2	0	11	0	22
Culture	56	5.5	0	0	0	5.5	0	0
Drama	30.5	30.5	0	0	0	17	0	0

Figure 6. Percentage breakdown of Z-mappings in our sample of figurative XYZs from the Web.

As shown in Figure 6 above, 50% of sporting-domain XYZs in our Web sample use an *Activity-to-Activity* mapping, as in "*Garry Kasparov is the Michael Jordan of chess"*. Many political XYZs (23%) use a *Side-to-Side* mapping, as in "*Newt Gingrich is the Trotsky of the Hard Right"*, while *Place* is significant in both *Art* (37%) and *Science* (28%) XYZs. However, the dimension that most significantly cuts across all kinds of XYZ mappings is *Time*.

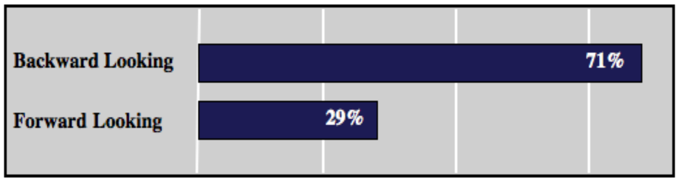


Figure 7. Breakdown of Time-to-Time comparisons according to whether they are backward-looking (X lives after Y) or forward-looking (Y lives after X).

Drilling deeper, we observe that our sample XYZs exhibit a remarkable conservativity when it comes to time. For the most part, we use the past as a lens through which we view the present, so that we often compare a contemporary individual to a historical entity, as in "*Rupert Murdock is the William Randolph Hearst of the 21st Century*". While one can also understand a historical figure by comparison to a contemporary individual, as in "*Jefferson is the Trotsky of the 18th Century*" or "*Russ Meyer is the Tarantino of the 70's*", the forward-looking comparison (in which the Y is more contemporary than the X) is far from being the norm. Overall, there is a very strong preference in XYZs for the backward-looking comparison, in which a contemporary individual is compared to a similar individual from the past. As shown in Figure 7 above, future-looking comparisons (such as "*Lillie Langtry was the Lindsay Lohan of the late 19th century*" and "*Scipio Africanus, the Tommy Franks of the Roman legions*") account for less than one third of all *Time-to-Time* mappings in our corpus of Web XYZs. It seems that even when we strive to be creative in the way we use stereotypes, we still cling to established conventions when it comes to domain, gender and time.

8 Discussion

Clichés come in many different linguistic forms. We have focused here on those forms that have obvious structural characteristics that make them easy to recognize in text, and thus easy to harvest from the linguistic Web. In this way, we can take regular core samples from the Web by looking for patterns like "*as X as Y*", "*about as X as Y*" and "*X is the Y of Z*". For instance, the pattern "*as X as Y*" allows us to explore the range and diversity of stereotypes that exemplify the qualities we care about most. This diversity – of lack thereof – can be surprising. For instance, similes harvested from the Web make extensive use of animal stereotypes, which are more often used to describe people than they are to describe other animals.

The "*about as X as Y*" pattern allows us to target more creative uses of language on the Web. The descriptions harvested with this pattern tend to be more linguistically complex, and are more likely to be unique one-offs. As demonstrated in Veale and Hao (2010), one can reliably determine whether an *about*-simile is intended ironically if one can determine the affect (positive or negative) of the attributed quality. By tracking uses of the "*about as X as Y*" pattern on the Web, we can quantify the relative balance of novelty, re-use and variation (as well as a tendency toward irony) in the linguistic structures that support creativity.

Every pattern of coining involves a creative tradeoff. We use a familiar form (like "*X is the Y of Z*") to carry a meaning that we know our audience knows how to unpack, and so we trade a freedom of form for the ability to convey a nuanced meaning in a compact package. While the form is familiar, the meaning can still surprise. The compactness of the package makes it a convenient form that others are more likely to mimic and re-use, especially on the Web. Yet, while patterns like "*X is the Y of Z*" impose few semantic constraints on the values of X, Y or Z, users intuitively preserve certain norms when coining their own XYZs. Thus, we have seen that figurative XYZs are more likely to compare entities from the same domain of experience, to compare individuals of the same gender, to compare entities in the real world to other real-world entities, and are more likely to compare entities to those that went before than to those that came after.

Content-wise, speakers appear to draw the elements of their creative comparisons from an unnecessarily limited pool of candidates. There is a strong preference to compare males to males (74% of our XYZ sample) while only 17% of our sample contains a female in either position, X or Y. Just over a fifth of our sample XYZs describe a political figure, while just under a fifth use a political figure to describe someone else (who has a 63% chance of also being a political figure). Though the 2,190 examples in our corpus employ a collective total of 668 different individuals in the Y position, the 20 most frequent Y's account for just over 11% of the whole corpus:

> *Michael Jordan* (21), *Chuck Norris* (20), *Donald Trump* (14), *Dick Cheney* (13), *Barry Bonds* (13), *Babe Ruth* (13), *Ann Coulter* (13), *Rush Limbaugh* (12), *Karl Rove* (12), *Barack Obama* (12), *Picasso* (11), *Julia Roberts* (11), *Indiana Jones* (11),

George Clooney (11), *Benedict Arnold* (11), *Ansel Adams* (11), *Walt Disney* (10), *Thomas Edison* (10), *Paganini* (10), *Moses* (10)

But language, like our shared set of popular stereotypes, is constantly evolving. Since our harvesting process is driven by the Google n-grams, we have focused here on stereotypes that were prominent on the Web when the n-grams were first released 6 years ago. Topical comparisons require topical stereotypes, or stereotypes that have already stood the test of time (like *Babe Ruth, Picasso, Benedict Arnold, Moses, Paganini, Walt Disney, Thomas Edison* and *Moses*: 8 of the top 20 Y's in our Web corpus). New stereotypes may develop slowly, then flourish. Thus, in the last 10 years, *Steve Jobs* has shifted from being the occasional X of a figurative XYZ comparison to being a popular Y for such comparisons, with the most dramatic increase in *Steve Jobs* XYZs occurring in the months following his death in 2011.

Language is a combination of the normative and the exceptional (Hanks, 1994). Norms are established collectively, or imported from without (e.g., from society and culture), while creative exceptions knowingly exploit, stretch and subvert these norms. Creative exceptions, if popular enough and mimicked often enough, can even become new norms in their own right. Creative coinages represent just one small facet of language. Nonetheless, novel coinages illustrate the processes of re-use, variation, subversion and re-invention in language on a scale that is easy (or eas*ier*) to study in quantitative terms.

This kind of *re-use with variation* happens all the time in language, but the Web in particular fosters and incentivizes this kind of creative display, both by presenting us with a constant stream of linguistic creativity and by challenging us to respond in kind. Indeed, the Web represents a highly competitive environment for language: as it becomes easier for everyone to speak to the world, it becomes harder for any one individual to be heard. Content producers must differentiate themselves if they are to compete for readers, followers and eyeballs. Linguistic creativity is an obvious value proposition for content producers, whose novel use of words can signal a novel take on the world. Samuel Goldwyn's famous line, "Let's have some new clichés", is as appropriate in the age of the Web as it was in the golden age of Hollywood. We strive to create new and memorable ways of expressing the same shared feelings and attitudes – new clichés, in other words – so that we can forge stronger and more effective connections with each other.

This is the fundamental irony of linguistic creativity: not only does it arise from our need for, and novel treatment of, clichés and stereotypes, it is also responsible for the creation of new clichés and new stereotypes. Consider that English does not provide an adjective for concisely evoking the feeling that something is technically brilliant yet lacking in credibility; the nouns "geek" and "nerd" come close, as do the adjectives "geeky" and "nerdy", but not close enough for all descriptive purposes. Fortunately, popular culture provides salient individuals that exemplify this complex property, so these individuals can be used to fill the newly identified gap in our lexicons. By packaging these well-known individuals into a figurative XYZ comparison, we can dynamically inflate the individual into a whole new category of people and things. Thus, for example, the individual *Kenny G* becomes the category "*the Kenny Gs of the world*", which can be stretched to include not just musicians, but anything we feel to be technically brilliant yet lacking in credibility. But as this category becomes over-populated, creative speakers will naturally seek out new ways of saying the same thing, and spawn new stereotypes and new clichés in the process.

9 Concluding Thoughts

In language, as in other forms of social display, we use creativity to help us to stand out from the crowd. If we cannot always be creative ourselves, we can occasionally re-use the creativity of others, either directly – by simply echoing another's witty remark or turn of phrase – or playfully, by putting a novel twist on a familiar form or on another's characteristic use of language.

Turing (1950) conceived of his famous test – a form of imitation game – as a test of linguistic interaction. In effect, the linguistic Turing Test is a *language imitation* game, where the computer attempts to communicate like a real human, and attempts to respond to human utterances as would a real human. But creativity is the *sine qua non* of language, and though we might conceive of a computer passing the Turing Test without demonstrating a creative fluency with language, it is harder to image such a computer being able to fool an interrogator that possesses this fluency. We expect other humans to react appropriately to our attempts at creativity: to laugh at our jokes, to admire our conceits – when they work – and to dismiss them as lame when they don't. Linguistic intelligence requires social intelligence, and it is this kind of intelligence – more so than IQ – that is assessed by the linguistic Turing Test.

Shortly before Turing wrote his 1950 paper on computers imitating humans, George Orwell fretted in 1946 that unthinking humans were being seduced by a cliché-ridden language to behave more like computers. Thus, any writer with a propensity to gum strips of prefabricated language together had already "gone some distance toward turning himself into a machine". When we talk of computers imitating humans, or of exhibiting human-like abilities of communication, we mean more than computers being able to imitate humans who are acting like machines in their unthinking use of language. Yet this is typically how we conceive of "chatbot"-like computer programs that are designed to pass modern-day instantiations of the Test: such programs, in Orwell's words, work by "tacking [phrases] together like the sections of a prefabricated henhouse".

Language that contains no trace of creativity is artless and tired, but language that contains too much creativity can be excessively artful and tiring. Humans aim for a happy medium between these poles. As we have shown in our analysis here, humans strive for novelty by finding new ways to instantiate familiar patterns of coining, but they are surprisingly conservative in how they choose to instantiate these patterns. Language is a layered system of conventions, and it seems we follow conventions even when we strive to creatively engage with conventions. An appreciation of these meta-conventions – which can be gleaned from a large-scale computational analysis of various support-structures for linguistic creativity in corpora and on the Web – can allow a computer to recognize another's efforts to meaningfully play with convention, and to perhaps find creative value in these playful variations. We are not referring to Wildean levels of linguistic creativity here, or even Orwellian levels for that matter: only when computers are masters of convention, rather than hard-coded slaves of convention, will they be able to imitate humans in their everyday creativity with language.

Acknowledgements

This research was supported by the WCU (World Class University) program under the National Research Foundation of Korea, and funded by the Ministry of Education, Science and Technology of Korea (Project No: R31-30007).

References

Bowdle, B. and Gentner, D. (2005). The Career of Metaphor. *Psychological Review*, *112*, 193-216.

Brants, Y. and Franz, A. (2006). Web 1T 5-gram Version 1. *Linguistic Data Consortium*.

Empson, W. (1980). Quoted in Ricks, (1980).

Hanks, P. (1994). Linguistic Norms and Pragmatic Exploitations, Or Why Lexicographers need Prototype Theory, and Vice Versa. In F. Kiefer, G. Kiss, and J. Pajzs (Eds.) *Papers in Computational Lexicography: Complex-1994*. Hungarian Academy of Sciences, Budapest.

Hao, Y. and Veale, Y. (2010). An Ironic Fist in a Velvet Glove: Creative Mis-Representation in the Construction of Ironic Similes. Minds and Machines, 20(4):483-488.

Kay, P. (2002). Patterns of Coining. *Unpublished manuscript*.

Keller, F. and Lapata, M. (2003). Using the Web to Obtain Frequencies for Unseen Bigrams. *Computational Linguistics*, *29*, 3:459-484.

Kilgarriff, A. (2007). Googleology is Bad Science. *Computational Linguistics, 33*, 1:147-151.

Moon, R. (2008). Conventionalized as-similes in English: A problem case. *International Journal of Corpus Linguistics 13*, 1:3-37.

Norrick, N. (1986). *Stock Similes. Journal of Literary Semantics XV*, 1:39-52.

Orwell, G. (1946). Politics And The English Language. *Horizon 13*, 76:252-265.

Pullum, G. (2003). Phrases for Lazy Writers in Kit Form. *Language Log, http://itre.cis.upenn.edu/~myl/ languagelog/archives/000061.html*.

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

Roncero, T., Kennedy, J. M. and Smyth, R. (2006). Similes on the internet have explanations. *Psychonomic Bulletin and Review*, *13*, 1:74-77.

Taylor, A. (1954). Proverbial Comparisons and Similes from California. *Folklore Studies 3. Berkeley: University of California Press*.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind* **LIX** (236): 433–460.

Veale, T. (2011). Creative Language Retrieval: A Robust Hybrid of Information Retrieval and Linguistic Creativity. *Proceedings of ACL'2011, the 49th Annual Meeting of the Association of Computational Linguistics*.

Veale, T. (2012). *Exploding the Creativity Myth: The Computational Foundations of Linguistic Creativity.* London, UK: Continuum/Bloomsbury.

Veale, T. and Hao, Y. (2007). Learning to Understand Figurative Language: From Similes to Metaphors to Irony. *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*, Nashville, USA.