We Can Re-Use It For You Wholesale

Serendipity and *Objets Trouvés* in Linguistic Creativity

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

Abstract

The *objet trouvé* or *found object* has become a staple of modern art, one which demonstrates that artistic creativity is just as likely to arise from serendipitous encounters in the real world as it is from purposeful exploration in the studio. These *readymades*, like Duchamp's *Fountain* (a urinal!) seem banal in their conventional contexts of use, but take on a new resonance and meaning when viewed in an artistic space. This paper considers the linguistic equivalent of the readymade object: a well-formed phrase that takes on a new significance when used in a new context with a potentially new and resonant figurative meaning. We show how linguistic readymades can be recognized and harvested on a large scale, and used to provide a robust and scalable form of creative language generation.

For Best Results, Just Add Water

Computationalists in the classical AI tradition generally prefer to model the creative process as a purposeful exploration of a conceptual space (e.g., see Boden, 1994). This concentration of computational effort makes good engineering sense, but little artistic sense, since much of what we consider to be creative insight occurs not in the studio or the laboratory but through everyday interaction with the real world. Serendipitous discovery is thus unlikely to arise in purposeful explorations, since specific applications have no remit beyond the immediate concerns of their programming (and programmers), and have no other lives to live when not actively pursuing these concerns.

The *objet trouvé* or *found object* is perhaps the most potent example of serendipity in artistic creation. An artist encounters an object with aesthetic merits that are overlooked in its banal, everyday contexts of use, yet when this object is moved to an explicitly artistic context, such as an art gallery, viewers are better able to appreciate these merits. The transformational power of a simple context switch is most famously demonstrated by the case of Marcel Duchamp's *Fountain*, a humble urinal that becomes an elegantly curved piece of sculpture when viewed with the right mindset. Duchamp referred to his *objets trouvés* as "readymades", since they allow us to remake the act of artistic creation as one of pure insight and inspired recognition rather than one of manual craftsmanship (see Taylor, 2009). In computational terms, the Duchampian notion of a readymade allows artistic creativity to be modeled not as a construction problem but as a decision problem. A computational Duchamp need not explore an abstract conceptual space of potential ideas. Rather, the issue instead becomes: how do we expose our Duchampian agent to the multitude of potentially inspiring real-world stimuli that a human artist encounters everyday?

Readymades represent a form of creativity that is poorly served by exploratory models of creativity, such as that of Boden (1994), and better served by the investment models such as the *buy-low-sell-high* theory of Sternberg and Lubart (1995). In this view, creators and artists find unexpected or untapped value in unfashionable objects or ideas that already exist, and quickly move their gaze elsewhere once the public at large come to recognize this value. Duchampian creators invest in everyday objects, just as Duchamp found artistic merit in bottles and combs. From a linguistic perspective, these everyday objects are commonplace words and phrases which, when wrenched from their conventional contexts of use, are free to take on enhanced meanings and provide additional returns to the investor. The realm in which a maker of linguistic readymades operates is not the real world, and not an abstract conceptual space, but the realm of texts: large corpora become rich hunting grounds for investors in linguistic *objets trouvés*.

This proposal is realized in a computational form in the following sections. A rich vocabulary of cultural stereotypes is acquired from the web, and it is this vocabulary that facilitates the implementation of a decision procedure for recognizing potential readymades in large corpora – in this case, the Google database of web ngrams (Brants and Franz, 2006). This decision procedure provides the basis for a robust web application called *The Jigsaw Bard*, and the cognitive insights that underpin *The Bard*'s conception of linguistic readymades are then put to the empirical test using statistical analysis. While readymades remain a contentious notion in the public's appreciation of artistic creativity – despite Duchamp's *Fountain* being considered one of the most influential artworks of the 20th century – we

shall show that the notion of linguistic readymade has significant practical merit in the realms of text generation and computational creativity.

A Modest Proposal

Readymades are the result of artistic *appropriation*, in which an object with cultural resonance – an image, a phrase, a thing – is re-used in a new context with a new meaning. As a fertile source of cultural reference points, language is an equally fertile medium for appropriation. Thus, in the constant swirl of language and culture, movie quotes suggest song lyrics, which in turn suggest movie titles, which suggest book titles, or restaurant names, or the names of racehorses, and so on, and on. The 1996 movie *The Usual Suspects* takes its name from a memorable scene in 1942's *Casablanca*, as does the Woody Allen play and movie *Play it Again Sam*. The 2010 art documentary *Exit Through the Gift Shop*, by graffiti artist Banksy, takes its name from a banal sign sometimes seen in museums and galleries: the sign, suggestive as it is of creeping commercialism, makes the perfect readymade for a film that laments the mediocrity of commercialized art.

Appropriations can also be combined to produce novel mashups; consider, for instance, the use of tweets from rapper Kanye West as alternate captions for cartoons from the New Yorker magazine (see the hashtag *#KanyeNewYorkerTweets*). Hashtags can themselves be linguistic readymades. When free-speech advocates use the hashtag *#IAMSpartacus* to show solidarity with users whose tweets have incurred the wrath of the law, they are appropriating an emotional line from the 1960 film *Spartacus*. Linguistic readymades, then, are well-formed and highly quotable text fragments, that carry some figurative content which can be reused and revitalized in many different contexts. In this spirit, the title of this paper, and all of its section headings, are readymades of varying provenance.

Naturally, a quote like "*round up the usual suspects*" or "*I am Spartacus*" requires a great deal of cultural knowledge to appreciate. Since literal semantics only provides a small part of their meaning, a computer's ability to recognize linguistic readymades is only as good as the cultural knowledge at its disposal. We need to explore a more modest form of readymade, as in the following phrases:

> *a wet haddock*
>
> *snow in January*
>
> *a robot fish*
>
> *a bullet-ridden corpse*

Each phrase can be found in the Google database of web ngrams, and each is likely a literal description of a real object or event – even "robot fish", which describes an autonomous marine vehicle whose movements mimic real fish. But each exhibits figurative potential as well, providing a memorable description of physical or emotional coldness. Whether or not each was ever used in a figurative sense before is not the point: once this potential is recognized, each phrase becomes a reusable linguistic ready-

made for the construction of a vivid figurative comparison, as in "*as cold as a robot fish*". We now consider the building blocks from which these comparisons can be ready-made.

Round Up The Usual Suspects

How does a computer acquire the knowledge that fish, snow, January, bullets and corpses are cultural signifiers of coldness, and that "heartless" robots in particular are cultural signifiers of emotional iciness? Much the same way that humans acquire this knowledge: by attending to the way these signifiers are used by others, especially when they are used in cultural clichés like proverbial similes (e.g., "as cold as a fish").

In fact, folk similes are an important vector in the transmission of cultural knowledge: they point to, and exploit, the shared cultural touchstones that speakers and listeners alike can use to construct and intuit meanings. Taylor (1954) catalogued thousands of proverbial comparisons and similes from California, identifying just as many building blocks in the construction of new phrases and figurative meanings. Only the most common similes can be found in dictionaries, as shown by Norrick (1986), while Moon (2008) demonstrates that large-scale corpus analysis is needed to identify folk similes with a breadth approaching that of Taylor's study. However, Veale and Hao (2007) show that the world-wide-web is the ultimate resource for harvesting similes.

Veale and Hao use the Google API to find many instances of the pattern "*as ADJ as a/an **" on the web, where ADJ is an adjectival property and * is the Google wildcard. WordNet (Fellbaum, 1998) is used to provide a set of over 2,000 different values for ADJ, and the text snippets returned by Google are parsed to extract the basic simile bindings. Once the bindings are annotated to remove noise, as well as frequent uses of irony, this web harvest produces over 12,000 cultural bindings between a noun (such as *fish*, or *robot*) and its most stereotypical properties (such as *cold, wet, stiff, logical, heartless*, etc.). Stereotypical properties are acquired for approx. 4,000 common English nouns. This is a set of building blocks on a larger scale than even that of Taylor, allowing us to build on Veale and Hao (2007) to identify linguistic readymades in their hundreds of thousands in the Google ngrams.

However, to identify readymades as resonant variations on cultural stereotypes, we need a certain fluidity in our treatment of adjectival properties. The phrase "*wet haddock*" is a readymade for coldness because "wet" accentuates the "cold" that we associate with "haddock" (via the web simile "*as cold as a haddock*"). In the words of Hofstadter (1995), we need to build a *SlipNet* of properties whose structure captures the propensity of properties to mutually and coherently reinforce each other, so that phrases which subtly accentuate an unstated property can be recognized. In the vein of Veale and Hao (2007), we use the Google API to harvest the elements of this SlipNet.

Specifically, we hypothesize that the construction "*as ADJ_1 and ADJ_2 as*" shows ADJ_1 and ADJ_2 to be mutually

reinforcing properties, since they can be seen to work together as a single complex property in the context of a single comparison. Thus, using the full complement of adjectival properties used by Veale and Hao (2007), we harvest all instances of the patterns "*as ADJ and * as*" and "*as * and ADJ as*" from Google, noting both the combinations that are found and their relative frequencies. These frequencies provide the link weights for the Hofstadter-style SlipNet that is then constructed. In all, over 180,000 links are harvested, connecting over 2,500 adjectival properties to each other. We put the intuitions behind this SlipNet to the empirical test in a later section.

We Can Remember It For You, Wholesale

In the course of an average day, a creative writer is exposed to a constant barrage of linguistic stimuli, any small portion of which can strike a chord as a potential readymade. In this casual inspiration phase, the observant writer recognizes that a certain combination of words may produce, in another context, a meaning that is more than the sum of its parts. Later, when an apposite phrase is needed to strike a particular note, this combination may be retrieved from memory (or from a trusty notebook), *if* it has been recorded and suitably indexed.

Given a rich vocabulary of cultural stereotypes and their properties, computers are capable of indexing and recalling a considerably larger body of resonant combinations than the average human. The necessary barrage of stimuli can be provided by the Google 1T database of web ngrams – snippets of web text (of one to five words) that occur on the web with a frequency of 40 or higher (Brants and Franz, 2006). Trawling these ngrams, a modestly creative computer can recognize well-formed combinations of cultural elements that might serve as a vivid vehicle of description in a future comparison. For every phrase **P** in the ngrams, where **P** combines stereotype nouns and/or adjectival modifiers, the computer simply poses the following question: is there an unstated property A such that the simile "*as A as P*" is a meaningful and memorable comparison? The property A can be simple, as in "as *dark* as a chocolate espresso", or complex, as in "as *dark and sophisticated* as a chocolate martini". In either case, the phrase **P** is tucked away, and indexed under the property **A** until such time as the computer needs to produce a vivid evocation of **A**.

The following patterns are used to identify potential readymades in the web ngrams:

(1) $Noun_{S1}$ $Noun_{S2}$

where both nouns denote are stereotypes that share an unstated property Adj_A. The property Adj_A serves to index this combination. Example: "as cold as a *robot fish*".

(2) $Noun_{S1}$ $Noun_{S2}$

where both nouns denote stereotypes with salient proper-

ties Adj_{A1} and Adj_{A2} respectively, such that Adj_{A1} and Adj_{A2} are mutually reinforcing. The combination is indexed on $Adj_{A1}+Adj_{A2}$. Example: "as dark and sophisticated as a *chocolate martini*".

(3) Adj_A $Noun_S$

where the noun is a known stereotype, and the adjective is a property that mutually reinforces an unstated, but salient, property of the stereotype. Example: "as cold as a *wet haddock*". The combination is indexed on this property.

Other, syntactically richer structures for P are also possible, as in the phrases "*a lake of tears*" (a melancholy way to accentuate the property "wet") and "*a statue in a library*" (for "silent" and "quiet"). In this current work, we shall focus on 2-gram phrases only.

Using these patterns, our application – the *Jigsaw Bard* – pre-builds a vast collection of figurative similes well in advance of the time it is asked to use or suggest any of them. Each phrase **P** is syntactically well-formed, and because **P** occurs relatively frequently on the web, it is likely to be semantically well-formed as well. Just as Duchamp side-stepped the need to physically originate anything, but instead appropriated pre-built artifacts, the *Bard* likewise side-steps the need for natural-language generation. Each phrase it proposes has the ring of linguistic authenticity; because this authenticity is rooted in another, more literal context, the *Bard* also exhibits its own Duchamp-like (if Duchamp-*lite*) creativity. We now consider the scale of the *Bard*'s generativity, and the quality of its insights.

Use Only The Finest Ingredients

The vastness of the web, captured in the large-scale sample that is the Google ngrams, means the *Jigsaw Bard* finds considerable grist for its mill in the phrases that match (1)…(3). Thus, the most restrictive pattern, pattern (1), harvests approx. 20,000 phrases from the Google 2-grams, for almost a thousand simple properties (indexing an average of 29 phrases under each property, such as "*swan song*" for "beautiful"). Pattern (2) – which allows a blend of stereotypes to be indexed under a complex property – harvests approx. 170,000 phrases from the 2-grams, for approx. 70,000 complex properties (indexing an average of 12 phrases under each, such as "*hospital bed*" for "comfortable and safe"). Pattern (3) – which pairs a stereotype noun with an adjective that draws out a salient property of the stereotype – is similarly productive: it harvests approx. 150,000 readymade phrases for over 2,000 simple properties (indexing an average of 125 phrases per property, as in "*youthful knight*" for "heroic" and "*zealous convert*" for "devout").

The *Jigsaw Bard* is best understood as a creative thesaurus: for any given property (or blend of properties) selected by the user, the *Bard* presents a range of apt similes constructed from linguistic readymades. The numbers above show that, recall-wise, the *Bard* has sufficient coverage to

work robustly as a thesaurus. Quality-wise, users must make their own determinations as to which similes are most suited to their descriptive purposes, yet it is important that suggestions provided by the *Bard* are sensible and well-motivated. As such, we must be empirically satisfied about two key intuitions: first, that salient properties are indeed acquired from the web for our vocabulary of stereotypes (this point relates directly to the aptness of the similes suggested by the *Bard*); and second, that the adjectives connected by the SlipNet really do mutually reinforce each other (this point relates directly to the coherence of complex properties, as well as to the ability of readymades to accentuate an unstated property).

Both intuitions can be tested using Whissell's (1989) dictionary of affect, a psycholinguistic resource used for sentiment analysis that assigns a pleasantness score of between 1.0 (least pleasant) and 3.0 (most pleasant) to over 8,000 commonplace words. We should thus be able to predict the pleasantness of a stereotype noun (like *fish*) using a weighted average of the pleasantness of its salient properties (like *cold*, *slippery*). We should also be able to predict the pleasantness of an adjective using a weighted average of the pleasantness of its adjacent adjectives in the SlipNet. (In each case, weights are provided by relevant web frequencies.) We can use a two-tailed Pearson test ($p < 0.05$) to compare the predictions made in each case to the actual pleasantness scores provided by Whissell's dictionary, and thereby assess the quality of the knowledge used to make the predictions. In the first case, predictions of the pleasantness of stereotype nouns based on the pleasantness of their salient properties (i.e., predicting the pleasantness of Y from the Xs in "*as X as Y*") have a positive correlation of **0.5** with Whissell; conversely, ironic properties yield a negative correlation of **–0.2**. In the second, predictions of the pleasantness of adjectives based on their relations in the SlipNet (i.e., predicting the pleasantness of X from the Ys in "*as X and Y as*") have a positive correlation of **0.7**. Though pleasantness is just one dimension of lexical affect, it is one that requires a broad knowledge of a word and its usage to accurately estimate. In this respect, the *Bard* is well served by a large stock of stereotypes and a coherent network of informative properties.

So Long, And Thanks For All The Robotic Fish

Fishlov (1992) has argued that poetic similes represent a conscious deviation from the norms of non-poetic comparison. His analysis shows that poetic similes are longer and more elaborate, and are more likely to be figurative and to flirt with incongruity. Creative similes do not necessarily use words that are longer, or rarer, or fancier, but use many of the same cultural building blocks as non-creative similes. Armed with a rich vocabulary of building blocks, the *Jigsaw Bard* harvests a great many readymade phrases from the Google ngrams – from the evocative "chocolate martini" to the seemingly incongruous "robot fish" – that can be used to evoke an equally wide range of properties.

This generativity makes the *Bard* scalable and robust. However, any creativity we may attribute to it comes not from the phrases themselves – they are readymades, after all – but from the recognition of the subtle and often complex properties they evoke. The *Bard* exploits a sweet-spot in our understanding of linguistic creativity just as much as Duchamp and his followers exploited a sweet-spot in the public's understanding of art and how it is practiced. But as presented here, the *Bard* is a starting point on our exploitation of linguistic readymades, and not an end in itself. By harvesting more complex syntactic structures, and using more sophisticated techniques for analyzing the figurative potential of these phrases, the *Bard* and its ilk may gradually approach the levels of poeticity discussed by Fishlov. For now, it is sufficient that even simple techniques serve as the basis of a robust and practical thesaurus application.

Exit Through The Gift Shop

A screenshot of *The Jigsaw Bard* is presented in Figure 1 (see overleaf). The application can be accessed online at: *http://www.educatedinsolence.com/jigsaw*

References

Boden, M. 1994. Creativity: A Framework for Research, Behavioural and Brain Sciences 17(3):558-568.

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

Fellbaum, C. (ed.) 2008. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge.

Fishlov, D. 1992. Poetic and Non-Poetic Simile: Structure, Semantics, Rhetoric. *Poetics Today*, 14(1), 1-23.

Hofstadter, D. R. 1995. *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. Basic Books, NY.

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.

Sternberg, R. J. and Lubart, T. I. 1995. *Defying the crowd: Cultivating creativity in a culture of conformity*. Free Press, New York.

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

Taylor, M. R. (2009). *Marcel Duchamp: Étant donnés* (Philadelphia Museum of Art). Yale University Press.

Veale, T. and Hao, Y. 2007. Making Lexical Ontologies Functional and Context-Sensitive. *In proc. of the 46th Ann. Meeting of the Assoc. of Computational Linguistics*.

Whissell, C. 1989. The dictionary of affect in language. In R. Plutchnik & H. Kellerman (eds.) *Emotion: Theory and research*. New York: Harcourt Brace, 113-131.

Figure 1. *Screenshot of The Jigsaw Bard, retrieving linguistic readymades on demand*.