Exploiting Readymades in Linguistic Creativity:

A System Demonstration of the *Jigsaw Bard*

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

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

Demonstration System can be viewed at: *http://www.educatedinsolence.com/jigsaw*

Abstract

Large lexical resources, such as corpora and databases of Web ngrams, are a rich source of pre-fabricated phrases that can be reused in many different contexts. However, one must be careful in how these resources are used, and noted writers such as George Orwell have argued that the use of canned phrases encourages sloppy thinking and results in poor communication. Nonetheless, while Orwell prized home-made phrases over the readymade variety, there is a vibrant movement in modern art which shifts artistic creation from the production of novel artifacts to the clever reuse of readymades or *objets trouvés*. We describe here a system that makes creative reuse of the linguistic readymades in the *Google* ngrams. Our system, the *Jigsaw Bard,* thus owes more to Marcel Duchamp than to George Orwell. We demonstrate how textual readymades can be identified and harvested on a large scale, and used to drive a modest form of linguistic creativity.

1 Introduction

In a much-quoted essay from 1946 entitled *Politics and the English Language*, the writer and thinker George Orwell outlines his prescription for halting a perceived decline in the English language. He argues that language and thought form a tight feedback cycle that can be either virtuous or vicious. Lazy language can thus promote lazy thinking, and vice versa. Orwell pours scorn on two particular forms of lazy language: the expedient use of overly familiar metaphors merely because they come quickly to mind, even though they have lost their power to evoke vivid images,; and the use of readymade turns of phrase as substitutes for individually crafted expressions. While a good writer bends words to his meaning, Orwell worries that a lazy writer bends his meaning to convenient words.

Orwell is especially scornful about readymade phrases which, when over-used, “are tacked together like the sections of a prefabricated hen-house.” A writer who operates by “mechanically repeating the familiar phrases” and “gumming together long strips of words which have already been set in order by someone else” has, he argues, “gone some distance toward turning himself into a machine.” Given his derogatory mechanistic view of the use of readymade phrases, Orwell would not be surprised to learn that computers are highly proficient in the large-scale use of familiar phrases, whether acquired from large text corpora or from the Google ngrams (see Brants and Franz, 2006).

Though argued with passion, there are serious holes in Orwell’s logic. If one should “never use a metaphor, simile or other figure of speech which you are used to seeing in print”, how then are familiar metaphors ever to become *dead* metaphors and thereby enrich the language with new terms and new senses? And if one cannot use familiar readymade phrases, how can one make playful – and creative – allusions to the writings of others, or

Proceedings of the ACL-HLT 2011 System Demonstrations, pages 14–19,
Portland, Oregon, USA, 21 June 2011. ©2011 Association for Computational Linguistics

mischievously subvert the conventional wisdom of platitudes and clichés? Orwell's use of the term *readymade* is entirely negative, yet the term is altogether more respectable in the world of modern art, thanks to its use by artists such as Marcel Duchamp. For many artists, a readymade object is not a substitute, but a starting point, for creativity.

Also called an *objet trouvé* or *found object*, a readymade emerges from an artist's encounter with an object whose aesthetic merits are overlooked in its banal, everyday contexts of use; when this object is moved to an explicitly artistic context, such as an art gallery, viewers are better able to appreciate these merits. The artist's insight is to recognize the transformational power of this non-obvious context switch. Perhaps the most famous (and notorious) readymade in the world of art is 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 "assisted readymades" because they allow an artist to remake the act of 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 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, as in Boden (1994). However, a Duchampian agent must instead be exposed to the multitude of potentially inspiring real-world stimuli that a human artist encounters everyday.

Readymades represent a serendipitous 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 urinals, 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 demonstrated in computational form in the following sections. We show how a rich vocabulary of cultural stereotypes can be acquired from the Web, and how this vocabulary 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 a robust basis for a simile-generation system called *The Jigsaw Bard*. The cognitive / linguistic intuitions that underpin the *Bard*'s concept of textual readymades are put to the empirical test in section 5. 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 a linguistic readymade has significant practical merit in the realms of text generation and computational creativity.

2 Linguistic Readymades

Readymades are the result of artistic *appropriation*, in which an object with cultural resonance – an image, a phrase, a quote, a name, 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 cartoon images from the *New Yorker* magazine (see hashtag *#KanyeNew-YorkerTweets*). 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 text fragments that are often highly quotable because they carry some figurative content which can be reused in different contexts.

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 thus explore here a more modest form of readymade – phrases that can be used as evocative image builders in similes – as in:

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

Each phrase can be found in 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). 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 readymade 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..

3 A Vocabulary of Cultural Stereotypes

How does a computer acquire the knowledge that fish, snow, January, bullets and corpses are cultural signifiers of coldness? 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 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.

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 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 the combinations that are found and their frequencies. These frequencies provide 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 one other. We put the intuitions behind this SlipNet to the empirical test in section five.

4 Harvesting Readymades from Corpora

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.

Ironically, Orwell (1946) suggests that lazy writers "shirk" their responsibility to be "scrupulous" in their use of language by "simply throwing [their] mind open and letting the ready-made phrases come crowding in". For Orwell, words just get in the way, and should be kept at arm's length until the writer has first allowed a clear meaning to crystallize. This is dubious advice, as one expects a creative writer to keep an open mind when considering *all* the possibilities that present themselves. Yet Orwell's proscription suggests how a computer should go about the task of harvesting readymades from corpora: by throwing its mind open to the possibility that a given ngram may one day have a second life as a creative readymade in another context, the computer allows the phrases that match some simple image-building criteria to come crowding in, so they can be stored in a database.

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 linguistic stimuli can be provided by the Google 1T database of Web ngrams (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 stereotypes that share an unstated property Adj_A. The property Adj_A serves to index this combination. Example: "as cold as a *robot fish*".

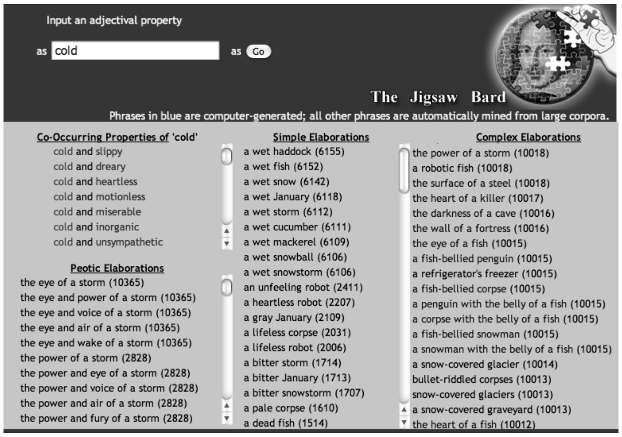
(2) $Noun_{S1}$ $Noun_{S2}$

where both nouns denote stereotypes with salient properties 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 $Noun_S$ denotes a cultural stereotype, and the adjective Adj_A denotes a property that mutually reinforces an unstated but salient property Adj_{SA} of the stereotype. Example: "as cold as a *wet haddock*". The combination is indexed on Adj_{SA}.

More complex 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 description, we focus on 2-gram phrases only.



Phrases in blue are computer-generated; all other phrases are automatically mined from large corpora.

Co-Occurring Properties of 'cold'

cold and slippy
cold and dreary
cold and heartless
cold and motionless
cold and miserable
cold and inorganic
cold and unsympathetic

Peotic Elaborations

the eye of a storm (10365)
the eye and power of a storm (10365)
the eye and voice of a storm (10365)
the eye and air of a storm (10365)
the eye and wake of a storm (10365)
the power of a storm (2828)
the power and eye of a storm (2828)
the power and voice of a storm (2828)
the power and air of a storm (2828)
the power and fury of a storm (2828)

Simple Elaborations

a wet haddock (6155)
a wet fish (6152)
a wet snow (6142)
a wet January (6118)
a wet storm (6112)
a wet cucumber (6111)
a wet mackerel (6109)
a wet snowball (6106)
a wet snowstorm (6106)
an unfeeling robot (2411)
a heartless robot (2207)
a gray January (2109)
a lifeless corpse (2031)
a lifeless robot (2006)
a bitter storm (1714)
a bitter January (1713)
a bitter snowstorm (1707)
a pale corpse (1610)
a dead fish (1514)

Complex Elaborations

the power of a storm (10018)
a robotic fish (10018)
the surface of a steel (10018)
the heart of a killer (10017)
the darkness of a cave (10016)
the wall of a fortress (10016)
the eye of a fish (10015)
a fish-bellied penguin (10015)
a refrigerator's freezer (10015)
a fish-bellied corpse (10015)
a penguin with the belly of a fish (10015)
a corpse with the belly of a fish (10015)
a fish-bellied snowman (10015)
a snowman with the belly of a fish (10015)
a snow-covered glacier (10014)
bullet-riddled corpses (10013)
snow-covered glaciers (10013)
a snow-covered graveyard (10013)
the heart of a fish (10012)

Figure 1. *Screenshot of The Jigsaw Bard, retrieving linguistic readymades for the input property "cold". See http://www.educatedinsolence.com/jigsaw*

Using these patterns, our application – the *Jigsaw Bard* (see Figure 1) – 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 prefabricated 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.

5 Empirical Evaluation

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 2-grams 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 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 to the coherence of complex properties, and to the ability of readymades to accentuate unstated properties).

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, its usage and its denotations to accurately estimate. In this respect, the *Bard* is well served by a large stock of stereotypes and a coherent network of informative properties.

6 Conclusions

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 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, and so, as presented here, is merely a starting point for our continued exploitation of linguistic readymades, rather than 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.

7 Hardware Requirements

The Jigsaw Bard is designed to be a lightweight application that compiles its comprehensive database of readymades in advance. It's run-time demands are low, it has no special hardware requirements, and runs in a standard Web browser.

Acknowledgments

This work was funded in part by Science Foundation Ireland (SFI), via the Centre for Next Generation Localization (CNGL).

References

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

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

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

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

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

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

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

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

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

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

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

Tony Veale and Yanfen Hao. 2007. Making Lexical Ontologies Functional and Context-Sensitive. *In Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics*.

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