Distributed Divergent Creativity:

Computational Creative Agents at Web Scale

Tony Veale, Guofu Li

School of Computer Science and Informatics, University College Dublin

Contact author: Tony.Veale@UCD.ie

Abstract

Divergence is a multi-faceted capability of multi-faceted creative individuals. It may be exhibited to different degrees, and along different dimensions, from one individual to another. The same may be true of computational creative agents: such systems may do more than exhibit differing levels of divergence: they may also implement the mechanics of divergence in very different ways. We argue that creative capabilities such as divergence are best viewed as cognitive *services* that may be called upon by cognitive agents to complete tasks in ways that may be deemed "original", or to generate products that may be deemed "creative". We further argue that in a computational embodiment of such an agent, cognitive services are best realized as modular, distributed Web services which hide the complexities of their particular implementations and which can be discovered, re-used and composed as desired by other Web-aware systems with diverse creative needs of their own. We describe the workings of one such reusable service for generating divergent categorizations on demand, and show how this service can be composed with others to support the generation and rendering of novel metaphors in an autonomous *Twitterbot* system.

Keywords: *creativity, divergence, similarity, web services, metaphor, Twitterbots*

1. Introduction

Habitual language aims to make essentialists of us all. The very existence of a word like "*creativity*", and the confidence with which we bandy it about, suggests that the underlying concept is just as simple and coherent. In fact, this simple word hides a philosophical morass, for creativity manifests itself in varying guises across diverse contexts and goals, and may be the product of different capabilities in different individuals. The adjective "creative" is ontologically promiscuous, and can equally be used to describe the producer of a creative act, the process that constitutes the act, or the product that results from the act. Even a relatively well-understood dimension of creative production such as *divergence* – which Baer (1999:753) defines as "*a kind of thinking often associated with creativity which involves the generation of varied, original or unusual ideas in response to an open-ended question or task*" – has complex dimensions of its own, and creative thinkers in diverse domains may be divergent in different ways (see Guilford, 1967). Some exhibit *fluency*, a capability for generating many alternate solutions or views on a topic. Others exhibit *flexibility*, a capability for generating substantially different kinds of solutions or perspectives. Some are more *original* than others, exhibiting a capability to generate solutions or perspectives that are rare or historically novel. Finally, some are also more *elaborate* than others, able to generate specific solutions that are rich in practical detail. Few creators score highly on all of these dimensions, while most score highly on just one or two. Psychometricians can measure creative aptitude in part by measuring a capacity for divergent thinking. For instance, the *unusual uses test* asks subjects to suggest atypical uses for a familiar object such as a brick or coffee can, where answers are collated and scored on each of these four dimensions (Torrance, 1980).

If human creators are divergent to different degrees on different dimensions, it

makes sense to assume that computational creators will exhibit the same multiplicity. Such systems may favor one dimension of divergence over others, or work best when connected in agent assemblies that carefully balance their complementary capabilities, as in a successful human brainstorming session. This is not simply a question of emphasis or parameter-setting: different computational agents may be divergent in different ways because they employ very different implementations of divergent behaviour. For divergence and creativity are not synonymous. Rather, divergence is a cognitive *service* that one may call upon when seeking to achieve creative ends, such as when solving a vexing problem or generating an artifact that is novel and useful. Just as software services are best designed as black boxes that hide the complexity of their inner workings so as to maximize both their interoperability and ease of reuse, creative computational services can likewise be conceived as modular capabilities that can be fluidly combined, composed and reused in pursuit of a larger creative goal.

To maximize their modularity, discoverability and reuse, creative computational systems can operate at Web scale, where they can recruit other, distributed creative Web services as needed to provide different creative capabilities, such as *divergence,* or indeed, competing implementations of these general capabilities. In this approach to Web-based of computational creativity, a creative system may do pivotal work of its own as a centralized process, but is just as likely to distribute key aspects of its work to appropriate external services with their own inner workings and aesthetic preferences. We thus explore two different but complementary forms of divergence in this work: *representational* divergence, in which a system finds diverse new ways to categorize and reason about a given idea; and *functional* divergence, in which a system calls upon a diversity of alternate computational services (perhaps modeling a diversity of creative thought processes) to achieve a particular goal.

Erl (2008) defines a service-oriented architecture (or SOA) as "*an architectural model that aims to enhance the efficiency, agility, and productivity of an enterprise by positioning services as the primary means through which solution logic is represented*." The SOA perspective on computationally creative systems allows for divergence in a variety of ways and at a variety of levels. At its simplest, it allows for a single Web service to provide divergent production on demand; such a service might generate alternative perspectives on a familiar concept, unusual uses for a common object, or alternate solutions to a loosely-defined problem. However, SOA also allows for a diversity of such services to compete with each other, so that a single computational system may pick and choose from the outputs of each on a topic-by-topic or a problem-by-problem basis. Moreover, even when only a single relevant service is available, a creative system may still use the outputs of this service in ways that were not anticipated by the original designer of the service. Necessity is the mother of invention, and creative systems are as free to find unusual uses for the outputs of a third-party service as are the humans who sit the *unusual uses test*.

Indeed, the need for divergence often forces builders of computational systems to exploit convergent resources in unusual "off-label" ways. Consider WordNet (Fellbaum, 1998), a much used resource for language processing that is a graph-structured cross between an electronic dictionary and a thesaurus. WordNet organizes words into sets of near synonyms, and associates meanings (and textual glosses) with these *synsets*. In the case of noun meanings, synsets are organized into taxonomies or standard AI IS-A hierarchies. As a static resource, WordNet necessarily represents a convergent world-view. Without its own dynamic reasoning capabilities, it limits itself to stating conventional views, so that e.g. *letter openers* are *tools* (and not potential *weapons*), *mustard* is a *condiment* (and not a possible *pesticide*), and so on.

For all its limitations, WordNet is large and it is free. It is thus used in a diversity of ways by NLP researchers, from a simple list of words in different categories (for checking spelling and solving crosswords), to a source of word-senses for word-sense disambiguation, to a source of expansion terms for the automated elaboration of information retrieval search queries, to a source of lexical and semantic features for building text classifiers. Perhaps its most effective use is as the convergent basis for measures of semantic similarity (see Budanitsky and Hirst (2006) for a review). Such measures use WordNet's taxonomy of noun-senses to assign a plausible similarity score to pairs of words, such as *dog* & *cat* or *car* & *bus*. In short, the public availability of WordNet as a resource, and of tools for calling upon it as a service, has encouraged a wide divergence of uses that its designers could never have anticipated.

In this paper we present a novel Web service for divergent categorization that we call *Thesaurus Rex*. This service may be called upon to provide alternative views on a familiar concept, placing the concept into diverse fine-grained categories that reflect its diverse (and sometimes unusual) uses in the hands of a creative thinker. *Thesaurus Rex* can thus be seen as a necessary complement to the convergent categorization of resources like WordNet. We demonstrate that *Rex* works well with WordNet, and that the combination of the two enables machines to make automated similarity judgments that accord remarkably well with human judgment. Finally, we show how *Rex* can be used in conjunction with other creative Web services, such as the *Metaphor Magnet* service of Veale (2013b), to generate a diversity of novel but apt metaphors that are in turn packaged as pithy tweets by an automated Twitterbot named *@MetaphorMagnet*. We present the workings of *Thesaurus Rex* in the next section, before describing how *Rex* is composed with the *Metaphor Magnet* service to generate diverse metaphors that are nonetheless grounded in intuitive, convergent notions of semantic similarity.

2. Diverging from the Convergent

Every act of creative divergence involves a departure from a convergent norm. The distinction between a convergent world view, which often structures problems so as to allow a single right answer and many wrong answers, and a divergent world view, which finds a contextual validity for many of these "wrong" answers, is captured in Tolstoy's famous opening line from Anna Karenina: "*Happy families are all alike; unhappy families are each unhappy in their own way*." To divergently mine the space of "wrong" answers for creative value, as novelists like Tolstoy do, and to understand what makes these unusual answers novel and interesting, one still needs a sense of the convergent, to compare to and to push against.

We represent a convergent world view at the level of properties and categories. Let $P_c(S) = \{p_1, p_2 \ldots p_n\}$ denote the set of properties that are typically associated with a concept *S*. For example, $P_c(\text{Scientist}) = \{\text{rational,logical,skeptical,}\ldots\}$. Conversely, $P_d(S)$ is the much larger set of properties that a creative individual may associate with *S* in diverse contexts. Interesting values that defy the convergent norm in $P_d(\text{Scientist})$ may include *illogical*, *awkward* and *mad*. Properties determine how we categorize a concept or an individual, so let $H_c(S)$ denote the set of categories, or hypernyms, into which *S* is typically placed. Thus, according to the convergent category system of WordNet, $H_c(\text{Scientist}) = \{\text{Person}\}$. Though WordNet allows for cross-categorization and multiple perspectives on the same word sense, most word senses are categorized under a single direct hypernym, and so $|H_c(S)| = 1$ for most values of *S*. When using WordNet to generate perspectives on a concept *S*, $H_c(S)$ is unlikely to score highly for fluency, flexibility or originality. So let $H_d(S)$ denote the

set of possible categorizations of S that allow us to imagine unusual instances of *S* or non-obvious perspectives on *S*-ness. Popular culture suggests many members for $H_d(\text{Scientist})$, such as *expert*, *geek, nerd, explorer* and *pioneer*, while a creative individual may invent many more that seem apt in a given context. As our properties and categories should work together to yield an integrated world-view, let $P \cdot H_c(S)$ denote the set of fine-grained categorizations of *S* that combine a property from $P_c(S)$ with a category from $H_c(S)$. For example,

$$P \cdot H_c(\text{Scientist}) = \{\text{logical\_person, rational\_person, } \ldots\}$$

Conversely, $P \cdot H_d(S)$ offers a more diverse set of plausible, fine-grained views on *S*, offering e.g. *antisocial_geek* and *arrogant_genius* for $S = \text{Scientist}$. While $P \cdot H_d(S)$ enlarges on $P \cdot H_c(S)$ it should also diverge from $P \cdot H_c(S)$ in meaningful if subversive ways.

Thesaurus Rex is a public Web service that offers, for any familiar concept *S*, a diverse set $P \cdot H_d(S)$ of fine-grained categorizations for *S* that it has previously acquired from the Web. In this section we discuss how *Thesaurus Rex* (or *Rex* for short) builds $P \cdot H_d(S)$ from $H_c(S)$ for many different values of *S*. By crawling the Web in advance, and pre-compiling a divergent system of categories, *Rex* can offer a real-time service to third-party systems that need to take a broader perspective on common ideas. *Rex* can also provide support to systems that aim to generate novel metaphors, by calculating similarity judgments on demand or by finding shared categorizations for very different concepts

To implement $P_c(S)$, *Rex* must look beyond WordNet to identify the cultural associations that everyday language makes evident in constructions such as the simile.

For instance, idiomatic similes tell us that ovens are *hot*, deserts are *dry*, skyscrapers are *tall*, Vikings are *blond*, and so on. Similes assume a convergence of speaker and hearer as to the salient properties of a comparison concept, and thus allow us (or our systems) to identify the affective and cultural resonances of familiar ideas that are needed in applications ranging from the linguistic to the robotic (see Veale, 2012b; Lewandowska-Tomaszczyk, 2014). Harvesting common similes from the Web (such as "*as cunning as a fox*") allows us to construct a convergent world view, while the long tail of Web similes, such as "*as bald as a bowling ball*", suggests associations that are rare but just as meaningful. Veale (2012a,b) describes how similes can be mined at Web scale to yield a large set of stereotypical associations, and we use this knowledge here as the basis of $P_c(S)$. From this starting point, we can also populate $P_d(S)$, $H_d(S)$ and $P \cdot H_d(S)$ from the Web, for many values of *S*, using an iterative bootstrapping process. Given a property $p_1 \in P_c(S)$, we can construct the Web query "p_1 * such as *S* and *", where * denotes a wildcard that can be filled by any word. As described in Kozareva, Riloff and Hovy (2008) and Veale, Li and Hao(2009), p_1 and *S* serve as anchors in this underspecified query. Suppose $p_1 = \text{exotic}$ and $S = \text{snake}$; the query "*exotic * such as snakes and **" finds Web texts that each yield a hypernym for $H_d(S)$, a fine-grained category for $P \cdot H_d(S)$ and an additional member for this category too. Using the Google API, *Rex* can retrieve "*exotic pets such as snakes and reptiles*", from which it extracts *pet* to add to $H_d(\text{Snake})$ and $H_d(\text{Reptile})$, and *exotic_pet* to add to $P \cdot H_d(\text{Snake})$ and $P \cdot H_d(\text{Reptile})$.

These new additions to *Rex*'s divergent category system can, in turn, be used to find related categorizations from the Web. For instance, *Rex* can now formulate the query "*exotic * such as reptiles and **", to learn that bats, amphibians and parrots can

also be classified as exotic pets and as exotic animals. By repeating the process again with newly acquired anchors, *Rex* enlarges its membership of exotic pets to include spiders, turtles and scorpions. By also shifting the position of the wild-card in a query, *Rex* can obtain new properties for $P_d(S)$, learning e.g. that spiders and snakes are *warm-weather pets*, spiders and scorpions are *creepy-crawly pets* and *invertebrate pets*, and spiders and ostriches are *odd pets*. Each cycle of queries uses the knowledge *Rex* has acquired in the previous cycle to anchor its new hypotheses, so that *Rex* can validate and elaborate these hypotheses via the Web. *Rex* thus harnesses the diversity of the Web by viewing all Web results as the output of a single highly divergent voice, rather than a mishmash of many convergent voices that are all talking at once.

We want the contents of $P_d(S)$, $H_d(S)$ and $P \cdot H_d(S)$ to be consistent with those of $P_c(S)$, $H_c(S)$ and $P \cdot H_c(S)$, and thus aim to exclude the more nonsensical voices of the Web. To this end, we filter what we acquire from the Web, to exclude risky categorizations that appear too far from the convergent norm. For each new candidate member $h_i \in H_d(S)$ that *Rex* acquires, it determines whether h_{new} is sufficiently close to a member $h_{old} \in H_c(S)$ to be deemed meaningful. If, according to a standard WordNet-based similarity metric, h_{new} is within a certain similarity threshold from h_{old}, *Rex* adds h_{new} to $H_d(S)$ and also makes the corresponding additions to $P_d(S)$ and $P \cdot H_d(S)$. In this way, *Rex* seeks to safely expand its categorical reach but does not eagerly embrace everything.

Consider the concept *cola*. Starting from $refreshing \in P_c(\text{Cola})$, *Rex* learns from its first series of Web queries that cola is an *effervescent_beverage*, a *sweet_beverage*, a *nonalcoholic_beverage* and more. After a second cycle of bootstrapping, in which this new knowledge is used to pose a new set of queries, *Rex* learns that cola is a

sugary_food, a *fizzy_drink* and a *dark_mixer*. After a third cycle, cola is found to be an *everyday beverage* and a *common drink*. After a fourth, it is also found to be an *irritating food* and an *unhealthy drink*. After the fifth, it is found to be a *stimulating drink*, a *toxic food* and a *corrosive substance*. In all, $refreshing \in P_c(\text{Cola})$ adds 14 fine-grained categorizations to $P \cdot H_d(\text{Cola})$ after 1 cycle, 43 after 2 cycles, 72 after 3 cycles, 93 after 4 cycles, and 102 after 5 cycles. During these bootstrapping cycles, the categorization *refreshing_beverage* also becomes associated with the concepts *champagne*, *lemonade* and *beer*. In the course of five bootstrapping cycles applied to its initial contents of $P_c(S)$ – finding a total of 16,688 simile-derived stereotypical associations for 6,512 concepts – *Rex* populates $P \cdot H_d(S)$ with more than 1,600,000 fine-grained categorizations for over 100,000 atomic and multi-word values of *S*.

These results can be evaluated along the same dimensions and using the same metrics that are applied to the outputs of human divergent thinking. For instance, the fluency of *Rex*'s divergence for a concept *S* is simply estimated as $|P \cdot H_d(S)|$, the raw number of different categorizations it provides for S. Likewise, the flexibility of *Rex* on S can be estimated as $|H_d(S)|$, the number of different generalizations that feed into $P \cdot H_d(S)$. The obviousness of any categorization *p_h* in $P \cdot H_d(S)$ is a function of how many times *p_h* is re-discovered as a categorization for *S* during *Rex*'s Web bootstrapping: for an obvious perspective is a familiar perspective, and will thus be encountered many times in different contexts, while a more original perspective may only be seen once. *Rex* can thus rank the members of $P \cdot H_d(S)$ by originality, that is, by inverse obviousness. The generality of any perspective *p_h* for *S* can be estimated as a function of how many other concepts *p_h* is seen to categorize: large categories are less discriminating, and offer less information about their members. Since

elaborate category definitions (such as *corrosive substance*) are more specific than less elaborate one (such as *substance*), the elaborateness of a categorization can thus be modeled as the *information content* (or IC) of the category into which it places an idea. This in turn can be estimated as a function of the size of the category.

Of course, *Rex* may score highly on all of these dimensions whilst still retrieving nonsensical, noise-ridden or useless categorizations from the Web. As these metrics are designed to punish convergence, they will reward anything that is non-convergent. Ultimately, we need to evaluate the outputs of *Rex* on a task with specific, convergent criteria for success and failure. Convergence and divergence are complementary rather than antagonistic modes of thought (de Bono, 1970), and so the success of a divergent system can often be measured by how well it contributes to a task with a convergent goal. We thus consider the contribution of *Thesaurus Rex* to the estimation of inter-concept similarity. This is a task that requires a cognitive agent to reason about the points of similarity – shared properties and categorizations – linking two ideas, and to fold this diversity into a single numeric measure that should accord well with human judgments. We now present the basis of such a similarity task.

3. Convergent *versus* Divergent Views on Similarity

The numeric measurement of inter-concept similarity is amongst the most robust and practical services one can squeeze from a convergent category system like WordNet. WordNet differs from conventional print dictionaries by organizing lexical units and their senses – the union of which might be termed a *lexicalized concept* – into sense hierarchies (at least for nouns and verbs) in which general categories are successively divided into increasingly informative sub-categories or instance-level ideas. This allows a computational system to gauge the overlap in information content (or IC),

and thus meaning, of two word senses or lexicalized concepts. One need only identify the deepest point in the taxonomy at which this content starts to diverge. This point of divergence is often called the LCS*,* or *least common subsumer*, of two concepts (see Pederson, Patwardhan and Michelizzi, 2004). Since sub-categories add new properties to those they inherit from their parents – Aristotle called these properties the *differentia* that stop a category system from collapsing into itself – the depth of a lexicalized concept in a taxonomy is an intuitive proxy for its information content. The measurement of similarity becomes a convergent process *par excellence* when it becomes a search for the most informative LCS in a steadily narrowing hierarchy.

Wu and Palmer (1994) use the depth of a lexicalized concept in the WordNet hierarchy as a proxy for IC, and thereby estimate the similarity of two concepts as twice the depth of their LCS divided by the sum of their individual depths. Leacock and Chodorow (1998) instead use the length of the shortest path between two concepts as a proxy for the conceptual distance between them. To connect any two ideas in a hierarchical system, one must vertically ascend the hierarchy from one concept, change direction at a potential LCS, and then descend the hierarchy to reach the second concept. (Aristotle was also first to suggest this approach in his *Poetics*). Leacock and Chodorow normalize the length of this path by dividing its size (in nodes) by twice the depth of the deepest concept in the hierarchy; the latter is an upper bound on the distance between any two concepts in the hierarchy. Negating the log of this normalized length yields a corresponding similarity score. While the role of an LCS is merely implied in Leacock and Chodorow's use of a *shortest path,* the LCS is pivotal nonetheless, and like that of Wu & Palmer, the approach uses an essentially vertical reasoning process to converge upon a single "best" generalization.

Depth is a convenient proxy for information content, but more nuanced proxies can

yield more rounded similarity measures. Resnick (1995) draws on information theory to define the IC of a lexicalized concept as the negative log likelihood of its occurrence in a corpus, either explicitly (via a direct mention) or by presupposition (via a mention of any of its sub-categories or instances). Since the likelihood of a general category occurring in a corpus is higher than that of any of its sub-categories or instances, such categories are more predictable, and less informative, than rarer categories whose occurrences are less predictable and thus more informative. The negative log likelihood of the most informative LCS of two lexicalized concepts offers a reliable estimate of the amount of information shared by these concepts, and thus a good estimate of their similarity. Lin (1998) combines the intuitions behind Resnick's metric and that of Wu and Palmer to estimate the similarity of two lexicalized concepts as an IC ratio: twice the IC of their LCS divided by the sum of their own ICs. Jiang and Conrath (1997) consider the converse notion of *dissimilarity*, noting that two lexicalized concepts are dissimilar to the extent that each contains information not shared by the other. If the IC of their most informative LCS is a good measure of what they *do* share, then the sum of their individual ICs, minus twice the content of their most informative LCS, is a reliable estimate of their dissimilarity.

Seco, Veale and Hayes (2006) present a minor innovation, showing how Resnick's notion of information content can be estimated without using an external corpus. Rather, when using Resnick's metric (or that of Lin, or Jiang and Conrath) for measuring the similarity of lexicalized concepts in WordNet, one can use the category structure of WordNet itself to estimate IC scores. Typically, the more general a concept, the more descendants it will possess. Seco, Veale and Hayes thus estimate the IC of a lexicalized concept as the log of the sum of all its unique descendants (both direct and indirect), divided by the log of the total number of concepts in the

entire hierarchy. Not only is this *intrinsic* view of information content convenient to use, without recourse to an external corpus, but it offers a better estimate of information content than its extrinsic, corpus-based alternatives, as measured relative to mean human ratings for the 30 word-pairs in the Miller & Charles (1991) test set.

A similarity measure can draw on other sources of information besides WordNet's category structures. One might eke out additional information from WordNet's textual glosses, as in Lesk (1986), or use category structures other than those offered by WordNet. Looking beyond WordNet, entries in the online encyclopedia Wikipedia are not only connected by a dense topology of lateral links, but they are also organized by a rich hierarchy of overlapping categories. Strube and Ponzetto (2006) show how Wikipedia can support measures of similarity and relatedness that better approximate human judgments than many WordNet-based measures. Nonetheless, WordNet can be a valuable component of a hybrid measure, and Agirre, Alfonseca, Hall, Kravalova, Pasca and Soroa (2009) use an SVM (support vector machine) to combine features from WordNet with raw text features harvested from the Web. Their best measure achieves a **0.93** correlation with human judgments on the Miller & Charles test set.

A fine-grained hierarchy permits fine-grained similarity judgments, and though WordNet is a popular platform for similarity measurement, its sense hierarchies are not especially fine-grained. However, we can make WordNet subtler and less single-mindedly convergent by automatically adding the fine-grained categories that are implied by its structure and its glosses. Veale (2003) describes a means of estimating $P_c(S)$ and $P \cdot H_d(S)$ for WordNet, by combining an adjectival property P that is found in the glosses of two lexicalized concepts S_1 and S_2 at the same depth in WordNet with the LCS of these concepts, H, to yield a new fine-grained category *P_H*. For example, the adjective "supreme" in the glosses of Odin and Zeus and

Jupiter, combined with their LCS *deity*, suggests *Supreme-deity* as a shared fine-grained categorization, while "1st" + *letter* allows *1st-letter* to be added to both $P \cdot H_d(\text{Alpha})$ and $P \cdot H_d(\text{Aleph})$. When applied to a broad spectrum of senses in WordNet, the effect is to elaborate its under-developed categories by giving them informative new sub-categories. Elaboration is one dimension of creative divergence identified by Guilford (1967), and Veale (2003) shows that the "lifting" of new fine-grained categories from WordNet glosses turns underspecified categories like *Deity* or *Letter* into highly structured conceptual systems that can be analogically mapped to each other with far greater precision. Thus, Veale (2003) shows how the members of the category *Greek-deity* can be accurately mapped to the members of *Roman-deity* or *Norse-deity*, while Greek letters can be accurately mapped to Hebrew letters.

4. Integrating Convergent *and* Divergent Knowledge Sources

How might the categorizations of a highly divergent system like *Thesaurus Rex* be married to the conservative, hand-crafted categories of WordNet, when the former is loosely encyclopaedic in nature and the latter aims for the rigor of a tightly-controlled dictionary? Simply, we assume that any categorization P_H_x offered by *Rex* for a word/concept can be added to WordNet if H_x can be mapped to some hypernym of some sense of C_y in WordNet. For instance, the perspective *corrosive_substance* can be added to WordNet for $C_y = \text{cola}$ by finding the specific hypernym of *cola* that corresponds to $H_x = \text{beverage}$. This process has the added benefit of disambiguating the words concerned, as both *cola* and *substance* are only tied to the specific WordNet senses of these words that share a hypernymic relationship. Only C_y and H_x pairs that can be mapped to specific hypernym relationships are mapped in this way, so only the diverse categorizations of *Rex* that fit into WordNet's world view are carried across.

A new categorization such as *corrosive_substance* is inserted as a hyponym of the correct sense of *substance* and as a hypernym of the correct sense of *cola*. Once *Rex*'s categorizations for a given word, or pair of words, are imported into WordNet in this way, any of the standard WordNet similarity measures can be applied to the pair.

Similarity is a measure of information overlap: how many categories do two ideas share, and how informative are those categories? This is not a question we can answer with a specific number, as categories may vary from person to person. Nonetheless, we can expect the most informative categories to contribute most to a consensus sense of similarity, where the information content (IC) of category *H* is formalized as in (1):

$$IC_{wn}(H) = -\log \frac{size_{wn}(H)}{\sum_{h \in WN} size_{wn}(h)} \quad (1)$$

Here $size_{wn}(H)$ is the number of lexicalized concepts in WordNet that claim *H* as a direct or indirect hypernym. We proper-named instances from estimates of category size, as WordNet contains an uneven sampling of such entities across its categories. The formulation in (1) is used to estimate the IC of a WordNet category relative to WordNet's other categories. When measuring the IC of a fine-grained categorization *p_h* suggested by *Rex*, we consider the relative size of all the fine-grained categories in *Rex*, as shown in the *Rex*-specific variation of (2):

$$IC_{rex}(P\_H) = -\log \frac{size_{rex}(P\_H)}{\sum_{h \in REX} size_{rex}(p\_h)} \quad (2)$$

Note that $size_{rex}(P\_H)$ denotes the size of the *Rex* category *P_H* after it and its members are successfully added to WordNet, so that $size_{rex}(P\_H)$ counts the number of lexicalized concepts in WordNet for which *P_H* can be added as a new

hypernym. The denominator in (2) sums over the size of all the fine-grained categories that can be successfully transplanted from *Rex* to WordNet, after they are added to WordNet. The information content of a *Rex* category *P_H* can be estimated relative to *Rex* using (2), while the information content of its *H* component can be measured relative to WordNet using (1). We thus calculate the geometric mean of both measures, in (3), to obtain the information content of a category *P_H* relative to both *Rex* and WordNet. We use the geometric rather than the arithmetic mean here, so that the combined IC_{wnrex} score is high only when the IC_{rex} and IC_{wn} scores are high, so that a good result reflects a convergence of these complementary resources.

$$IC_{wnrex}(P\_H) = \sqrt{IC_{rex}(P\_H) \times IC_{wn}(H)} \quad (3)$$

Any pair of concepts C_1 and C_2 that we wish to compare may share a variety of direct or indirect hypernyms in WordNet, and each will have its own information content as estimated by (1), (2) or (3). Rather than pick a single hypernym as an LCS, we instead build a feature vector for C_1 and C_2, in which each of the WordNet hypernyms of C_1 or C_2 corresponds to its own real-valued dimension in both vectors. If a concept C_1 or C_2 has a WordNet hypernym H_x then one dimension of its vector representation will encode a numeric measure of the information content of H_x, so that each hypernym contributes to a vector according to its specificity. If only one of C_1 or C_2 claims H_x as a WordNet hypernym, then the value associated with H_x in the other vector is 0.

When comparing C_1 and C_2 we may import additional categories $P_i\_H_j$ from *Thesaurus Rex* into WordNet. Each $P_i\_H_j$ is inserted under the appropriate sense of H_j in WordNet, so the dimension corresponding to H_j in each vector needs to reflect the addition of this new information. A vector for a concept *C* comprises *n* numeric

values $H_0 \ldots H_n$ for the n pooled WordNet hypernyms $H_0 \ldots H_n$ of C_1 and C_2. The value of the dimension for each H_j in the vector representation of C_i is given by (4):

$$\underset{i\in\{1,2\}}{C_i}[H_j] = \max\left(IC_{wn}\left(H_j\right), \max\left(IC_{wnrex}\left(P\_H_j\right)\right)\right) \tag{4}$$

So the most informative *Rex* category P_H_j will influence the dimension for H_j only if P_H_j brings more information from *Rex* than that already present in WordNet's view of H_j. With vectors for C_1 and C_2, we can now estimate the similarity of C_1 and C_2 as the cosine of the angle between these vectors. Identical vectors yield a similarity score of 1.0, while orthogonal vectors will yield a minimal similarity score of 0.

5. The Role of Creative Divergence in Similarity Judgments

5.1. Evaluating Coverage and Quality

Recall that *Rex* associates with a topic S a set $P \cdot H_d(S)$ of fine-grained categorizations of the form $P_i\_H_j$. To estimate both the coverage and quality of these categories, we replicate the experimental setup of Almuhareb & Poesio (2005), who use information extraction from the Web to acquire attribute values for different terms, and who use clustering over these values to group concepts into intuitive categories. Almuhareb & Poesio evaluate the quality of the resulting category clusters by comparing them to the categories of a hand-crafted gold-standard: WordNet.

Almuhareb & Poesio created a balanced set of 402 nouns from 21 semantic classes in WordNet. We denote this noun set here as AP402. They then acquired attribute values for these nouns (such as *hot* for coffee, *red* for car, etc.) using the Web query "(a | an | the) * S (is | was)" to find corresponding P_i values for each S. Those authors

did not seek to acquire a set of hypernyms $H_d(S)$ for each *S*, nor did they try to link the acquired attribute values to a parent category (H_j) in the taxonomy (they did, however, seek matching attributes for these values, such as *Temperature* for *hot*, but that aspect is not relevant here). They acquired 94,989 attribute values in all for the 402 nouns in AP402. These values were then used as features of the corresponding nouns in a clustering experiment, using the CLUTO system of Karypis (2002). By using attribute values as a basis for partitioning AP402 into 21 different categories, Almuhareb & Poesio attempted to reconstruct the original 21 WordNet categories from which AP402 is drawn. The more accurate the match to the original WordNet clustering, the more reliably these attribute values can be used as a representation of conceptual structure. In a first attempt, they achieved 56.7% clustering accuracy against the original human-assigned categories of WordNet. By using a noise-filter to remove almost half of their Web-harvested attribute values, they raised cluster accuracy to 62.7%. Specifically, they achieve a cluster purity of 0.627 and a cluster entropy of 0.338 using 51,345 features to describe the 402 nouns in AP402.

We replicate the same experiment for *Thesaurus Rex* using the same AP402 noun set, and assess the clustering accuracy (again using WordNet as a gold-standard) after each bootstrapping cycle. We use only the P_i part of each category $P_i\_H_j$ in $P \cdot H_d(S)$ as a feature for the clustering process. This avoids circularity, since the H_j parts were previously filtered against WordNet and only the P_i parts are truly independent of WordNet. The different values of $P_i\_H_j$ in $P \cdot H_d(S)$ were acquired from the Web by *Thesaurus Rex* using an iterative bootstrapping process that was allowed to run for five iterations. It is instructive then to consider the clustering accuracy of *Thesaurus Rex* on AP402 after each iteration of the bootstrapping process. Table 1 presents the

cluster purity achieved after each iteration, along with the number of distinct P_i fields used as features in the clustering process. There are words in AP402 that are not encountered during bootstrapping, and the coverage column shows the percentage of AP402 (where 100% = all 402 nouns) that was actually clustered at each iteration.

Table 1. *Clustering accuracy on the AP402 noun test-set after each iteration of bootstrapping on the Web. Purity represents the average conceptual homogeneity of each cluster. A purity of 1.0 is achieved only when each cluster only contains terms/concepts from the same WordNet category.*

Cycle	Entropy	Purity	# Features	Coverage
1st	.254	.716	837	59%
2nd	.280	.712	1338	73%
3rd	.289	.693	1944	79%
4th	.313	.660	2312	82%
5th	.157	.843	2614	82%

AP402 includes some low-frequency words, such as *casuarina, cinchona* and *dodecahedron*, and Almuhareb and Poesio note that one third have a frequency of just 5 to 100 occurrences in the British National Corpus. Looking to the coverage column of each table, we thus see that there are words in AP402 for which no categorizations at all can be acquired in 5 cycles of Web bootstrapping. Test words for which it fails to find a any categorization include *yesteryear, nonce* (very rare), *salient* (typically an adjective), and *airstream* (not typically a solid compound). The coverage of the categorizations acquired by *Thesaurus Rex* for AP402 tops out at 82% after 5 cycles. Notice that as more features are acquired and coverage increases (albeit in smaller

increments) with each successive cycle until cycle 4, we see concomitant decreases in cluster purity. Our intuition here is that as new terms are added to the knowledge-base, their relative newness makes them more difficult to categorize. However, when coverage tops out at 82% during the 4th cycle, those new features added in the 5th cycle inform the categorization of terms added in previous cycles, and cause purity to rise accordingly. Had space and time limitations not prevented us from running a 6th cycle of bootstrapping, we might thus have seen purity rise even further. Most striking of all is the representational concision of the diverse categorizations that are acquired in these 5 cycles. *Thesaurus Rex* yields a high cluster accuracy (purity = 0.843) using a pool of just 2614 fine discriminators, while Poesio and Almuhareb use 51,345 features even after their feature-set has been filtered for noise.

5.2. Evaluating Inter-concept Similarity

We evaluate *Thesaurus Rex* as a similarity service by estimating how closely its judgments correlate with those of human judges on the 30-pair word set of Miller & Charles (henceforth M&C), who aggregated the judgments of human raters into mean ratings for these pairs. We evaluate three variants of *Rex* on M&C: ***Rex*-full**, which enriches WordNet with as many of *Rex*'s categories as it can coherently import; ***Rex*-wn**, which uses only WordNet categories to drive its cosine similarity metric, with no input at all from *Rex*; and ***Rex*-conv**, which enriches WordNet with relatively convergent categorizations from *Thesaurus Rex* that are encountered at least five times during Web bootstrapping. While *Corrosive-substance* is a common and thus convergent view for *acid* that is encountered repeatedly during bootstrapping, this view is found just once or twice for *cola* or *juice*. The frequency of a fine-grained perspective for a topic roughly approximates what Ortony (1979) calls *salience,* insofar as a frequently used perspective is likely to seem more obvious to speakers,

and a more salient basis of comparison, than one that is rare and infrequently used.

Table 2 lists coefficients of correlation (Pearson's ***r***) with mean human ratings for a range of WordNet-based metrics. Table 2 includes the hybrid *WordNet+Web+SVM* metric of Agirre, Alfonseca, Hall, Kravalova, Pasca and Soroa (2009) – who report a correlation of **.93** – and the Mutual-Information-based PMI*max* metric of Han, Finin, McNamee, Joshi and Yesha (2012). The latter achieves good results for 27 of the 30 M&C pairs by enriching a PMI metric with an automatically generated thesaurus. Yet while informative, this automatic thesaurus is not organized as an explanatory system of fine-grained categories as it is in *Thesaurus Rex*. While *Rex* provides a practical numeric estimate of the similarity for two lexicalized ideas, it also offers practical category-grounded explanations as to why they are similar, e.g. by showing that *cola* & *acid* are not just *substances*, they are *acidic substances* and *corrosive substances*.

Table 2. *Product-moment correlations (Pearson's **r**) with mean human ratings on all 30 word pairs of the Miller & Charles similarity data-set.*

Similarity metric	***Pearson's r***	***Similarity metric***	***Pearson's r***
Wu & Palmer'94*	.74	Seco *et al.* '06*	.84
Resnick '95*	.77	Agirre *et al.* '09	**.93**
Leacock/Chod'98*	.82	Han *et al.* '12	.856
Lin '98*	.80	***Rex*-wn**	.84
Jiang/Conrath '97*	-.81 **	***Rex*-full**	.89
Li *et al*. '03	.89	***Rex*-conv**	**.93**

** As re-evaluated by Seco et al. (2006) for all 30 pairs ** As a distance mesasure, Jiang/Conrath is used here as a reverse measure of similarity, hence the minus sign.*

5.2.1. Analysis

***Rex*-wn** does no better than the metric of Seco, Veale and Hayes (2006) on the M&C dataset, suggesting that *Rex*'s vectors of IC-weighted WordNet hypernyms are no more discerning than a single convergent LCS. However, *Rex*'s vectors do permit it to encode its own fine-grained perspectives, allowing ***Rex*-full** to achieve a comparable correlation – **0.89** – to that of Li, Bandar and McLean (2003). These perspectives are sometimes idiosyncratic and may not generalize across independent judges, while the mean ratings of M&C are the stuff of consensus, not individual creativity. Outside the realm of metaphor it often makes sense to align our judgments with those of others.

By limiting its use of *Thesaurus Rex* to the perspectives that other judges are most likely to use, ***Rex*-conv** obtains a correlation of .**93** with mean human ratings on all 30 M&C pairs. This result is comparable to that reported by Agirre, Alfonseca, Hall, Kravalova, Pasca and Soroa (2009), who use SVM-based learning to combine the judgments of two metrics, one based on WordNet and another on the analysis of the Web contexts of both input terms. However, *Rex* has the greater capacity for insight, as it augments the structured category system of WordNet with structured, finer-grained categories of its own. Because *Rex* makes selective use of the products of divergent thinking, this selectivity also yields concise explanations for its judgments.

6. Diverse Services, Converging in One Client

Metaphor Magnet (Veale, 2013a) is another Web service – or rather a constellation of related services – that provides creative linguistic expressions on demand. The service operates on the principles of *Creative Information Retrieval* (Veale, 2011; Veale, 2013b) to retrieve apt linguistic forms from the Google N-Grams database of (Brants and Franz, 2006) which it then re-purposes as metaphors and other kinds of figurative

expression. *Metaphor Magnet* views a corpus such as the Google n-grams as a vast lexicalized idea space, in which a large number of well-formed phrases with plausible semantic interpretations float in a much larger sea of Web noise. Shorn of their original intent, these phrases can be imbued with new purpose and meaning in a new metaphorical context. Consider the Google 2-gram "*robot fish*": the most likely use of the phrase on the Web is to denote a class of fish-like submersible robot, but knowing that both fish and robots are stereotypically *cold*, *Metaphor Magnet* repurposes this linguistic readymade as a metaphoric vehicle for emotional coldness, as in the simile "*as cold as a robot fish*". *Metaphor Magnet* taps into the same knowledge-base of stereotypical properties (from Veale 2012a) as *Thesaurus Rex*, allowing these properties to serve as a lingua franca for any interactions between the two services.

Metaphor Magnet also mines familiar copula metaphors of the form "*X* is a *Y*" from Google's Web n-grams. The Web is rich in copula metaphors such as "*politics is war*" and "*crime is a disease*" which *Metaphor Magnet* can use as building blocks for its own original compositions. Consider its strategy of counterpuntal composition, where two competing metaphors for the same topic are framed so as to highlight the sharp difference in viewpoint offered by each. For example, *Metaphor Magnet* finds the potential metaphors "*compassion is a virtue*" and "*compassion is a weakness*" in the Google 4-grams for the topic *compassion*, and uses the antonymy of *treasured* and *overlooked* to note a strong affective contrast between *virtue* (a word with a strong positive connotation) and *weakness* (one with strong negative connotation). A metaphor-generating bot on Twitter, named *@MetaphorMagnet,* interacts with the *Metaphor Magnet* service as a client, to obtain apt metaphors for its topic of the hour. So e.g. given conflicting views on *compassion*, the bot imposed its own linguistic rendering (in 140 characters), to broadcast this pithy tweet of September 21, 2014:

To some humanitarians, compassion is a treasured virtue. To others, it is an overlooked weakness. #Compassion=#Virtue #Compassion=#Weakness

If it can find Google 2-grams that link the source and target ideas of a metaphor, in which the head of the 2-gram denotes a kind of person, *@MetaphorMagnet* reuses these 2-grams as the Twitter handles of fictional Twitter users, to which it can ascribe two conflicting views. Consider the use of *the 2-gram debt investor* in this tweet:

.@suffering_saint says poverty is a cherished blessing

@debt_investor says it is an overlooked risk. #Poverty=#Blessing #Poverty=#Risk

So *@MetaphorMagnet*, like any user of Twitter, strives for a diversity of outputs and viewpoints while converging around an identifiable aesthetic. In the case of *@MetaphorMagnet*, its aesthetic is that if a cynical curmudgeon that strives to showcase the negative in every situation, not least because measures of lexical sentiment make such an aesthetic a computationally achievable one. To achieve these divergent and convergent aims, *@MetaphorMagnet* is a client of both the *Metaphor Magnet* and *Thesaurus Rex* Web services. For unlike other metaphor-generating Twitterbots, such as the *Metaphor-a-Minute* bot (handle: *@metaphor-minute*), which fill templates with mostly random combinations of words, *@MetaphorMagnet* only generates metaphors that it itself can appreciate. The *Metaphor Magnet* service provides a shallow semantic understanding, by selecting readymade n-grams in ways that exploit its knowledge of antonymy (derived in large part from WordNet) and lexical affect (i.e., +/- real-valued sentiment scores, as acquired in Veale 2012b), of verbs and their case frames, and of nouns and the roles they can fill. For its part, *Thesaurus Rex* enables *@MetaphorMagnet* to select metaphors that pair a topic with a vehicle for which WordNet-based similarity metrics report a low score, yet for which *Rex* suggests one or more fine-grained categorizations to unite the two.

For example, the Web 3-gram "*war and divorce*" suggests to *@MetaphorMagnet* that war and divorce are sufficiently associated in the popular imagination to support a metaphor betweent the two. In turn, the *Metaphor Magnet* service provides a variety of phrases to elaborate the metaphor *Divorce is War*, and suggests corpus-guided interpretrations of the pairing (e.g. the Google 2-grams reveal that, like wars, divorces can be described as *ugly*, *major*, *serious* and even *bloody*). Its related sub-services also suggests poetic framings, in the guise of similes, superlatives, rhetorical questions and other tropes. Yet because *Metaphor Magnet* is a robust Web service, it will provide such outputs for even tenuous metaphorical pairings, like the unmetaphorical *War is Conflict*. So *@MetaphorMagnet* also calls upon the *Thesaurus Rex* service to ensure that *War* and *Divorce* are dissimilar enough to be metaphorically interesting, yet mediated by one or more specific, fine-grained categorizations that bring them closer together. Figure 1 shows the shared categories returned by *Thesaurus Rex* for the input *divorce & war*:

Place Figure 1 about HERE

Figure 1. The output of Thesaurus Rex for the input "divorce & war". The size of each fine-grained perspective reflects its Web frequency (and thus salience) for "war".

*Thesaurus Re*x allows *@MetaphorMagnet* to see that two dissimilar concepts (whose WordNet LCS is *event*) are connected in ways that highlight key properties of war, such as major, traumatic and devastating. War is thus a good vehicle for divorce.

@MetaphorMagnet is selective in the metaphors it generates, combining the inputs of two different Web services to identify the ideas, pairings and viewpoints that may yield a memorable tweet. It then uses a diversity of linguistic forms to package these pairings and viewpoints in pithily provocative ways. Yet creative metaphor generators

like *@MetaphorMagnet* can fall into the same "uncanny valleys" (Mori, 2005) as generators of photorealistic imagery: as such systems rely more on their own semantic models to generate their own outputs from first principles, and rely less on the wholesale reuse of human texts, the more likely it is their artifice will shine through. As a result, complex creative systems may seem less natural and appealing than much simpler toy systems that lack any understanding of what they generate, but which rely instead on superficial trickery to attract attention. It remains to be seen as to when, or how, creative systems will climb out of the uncanny valleys into which their hidden complexity pitches them. For now, we evaluate *@MetaphorMagnet* against a simple but popular baseline: the *@Metaphor-Minute* Twitterbot, which fills its linguistic templates with random word suggestions from the *Wordnik* Web service to produce outputs that are often bizarre but very novel. Table 3 shows the distribution of mean human ratings for randomly chosen tweets from *@MetaphorMagnet* and *@Metaphor-Minute*, as elicited from volunteers on the crowd-sourcing platform Crowdflower.

Table 3. *Distribution of mean human ratings for @MetaphorMagnet metaphors w.r.t comprehensibility, novelty and retweetability. @Metaphor-Minute baseline in [].*

Rating	***Comprehensibility***	***Aptness***	***Novelty***	***Retweetability***
Very Low	11.6% [23.9%]	0% [84%]	11.9% [9.5%]	15.49% [40.94%]
Med. Low	13.2% [22.2%]	22% [16%]	17.3% [12.4%]	41.88% [34.14%]
Med. High	23.7% [22.4%]	58% [0%]	21% [14.9%]	27.36% [15.04%]
Very High	**51.5%** [31.6%]	**20**% [0%]	49.8% [**63.2%**]	**15.27%** [9.88%]

We chose 60 tweets at random from the outputs of each bot. CrowdFlower annotators were not informed of the origin of any tweet, but simply told that each was collected

from Twitter because of its metaphorical content. For each tweet, annotators were asked to rate its metaphor along three dimensions, *Comprehensibility*, *Novelty* and likely *Retweetability*, and to rate all three dimensions on the same scale, ranging from *Very Low* to *Medium Low* to *Medium High* to *Very High*. CrowdFlower was used to solicit ten annotations per tweet (and thus, per dimension), though scammers (non-engaged judges) were later removed from this pool. Table 3 shows the distributions of mean ratings per tweet, along each of these three dimensions for each Twitterbot.

To ensure that raters really did assign a meaning to each tweet they rated as comprehensible, we also conducted a *cloze* test on CrowdFlower, where other judges were presented with the same tweets, but from which a key pair of qualities had been blanked out. For instance, for the *@MetaphorMinute* tweet "*a doorbell is a sportsman: fleetwide and infraclavicular*" the qualities *fleetwide* and *infraclavicular* was replaced with blanks, while for the @MetaphorMagnet tweet "*To some voters, democracy is an important cornerstone. To others, it is a worthless failure*" the qualities *important* and *worthless* were replaced with blanks. In each case, raters were presented with the removed pair of qualities and four distractor pairs taken from other tweets from the *same* bot. If 75% or more of raters for a tweet were able to choose the correct pair to re-fill the blanks, the tweet was deemed to have *Very High* Aptness; if 25% or less were able to do so, it was deemed to have *Very Low* Aptness; the two intermediate quartiles were mapped to Medium Low and Medium High Aptness.

Though *@MetaphorMagnet* obtains more favorable ratings for each dimension (except novelty), the differences – though statistically significant at the $p < .001$ level – are not as large as the differences in complexity and depth of the two systems. So systems like *@MetaphorMagnet* must further evolve in divergence *and* convergence: more divergence is needed to produce surprising, provocative and novel outputs, and

more convergence is needed to make these surprises truly meaningful to an audience.

7. Conclusions

We have argued for the merits of viewing creativity as a constellation of related capabilities rather than as a constellation of connected mechanisms, or indeed, as a single *one-size-fits-all* mechanism. Capabilities may be expressed to different degrees in different individuals, and may not share the same cognitive or neurological basis in all who express them. When it comes to modeling these capabilities on a computer, to build computational systems that exhibit some measure of human-level creativity, it follows that different machines may implement a creative capability in different ways and express this capability in differing degrees and with varying *unusual* uses. Just as an electricity grid can pool power supplies from diverse sources – wind, hydro, fossil and nuclear – into a seamless meta-service, a service-oriented architecture (SOA) of distributed creative Web services can allow the clients of these services to be creative in their own right, without having to implement everything themselves, or indeed, having to understand the internal workings of any remote services they may recruit.

This distributed view of machine creativity is lightweight and theory-neutral: a service for a human capability such as divergent categorization may use a cognitively plausible approach, or even be grounded in a model that is neurologically inspired, or it may employ an approach that pools the viewpoints and knowledge of many different humans, each of which may be convergent in their own ways. *Thesaurus Rex* elects for the latter, and though it does not offer a psychological model of individual human divergence, it yields results on an extrinsic similarity task that closely accord with human similarity judgments. It is this extrinsic quality that allows *Rex* to be used to good effect with other creative Web services, such as *Metaphor Magnet*, to support

the workings of an autonomous creative generation system like *@MetaphorMagnet*. These services have since been used by third-party developers to build bots of their own, such as *@AppreciationBot* (a generator of figurative critiques in response to a tweets about museum pieces by another bot named *@MuseumBot*) and *@HueHueBot* (which invents and tweets figurative new names for the RGB colours tweeted by a bot named *@EveryColorBot*). We cannot anticipate all the uses to which a generic service will be put, but if a service is truly reusable then others will use it in diverse and unpredictable ways, to suit their own needs rather than those anticipated by its creator. Ultimately, the individual services in a SOA for Web creativity will be judged on their robustness, on their coverage, on the naturalness of their outputs and on their ability to plug-and-play with others in unanticipated ways. A Web service such as *Thesaurus Rex* will thus be judged not just on its own divergent abilities, but on its contribution to the divergence of diverse clients in their varied application contexts.

Thesaurus Rex can be accessed as an interactive application on the Web at *http://boundinanutshell.com/therex2*, where instructions on how to use it as a remote XML-producing service are also available. *Metaphor Magnet* can be accessed as an application at http://ngrams.ucd.ie/metaphor-magnet-acl/ while *@MetaphorMagnet* is the handle of the autonomous Twitterbot that exploits both of these services.

References

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

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca and Aitor Soroa. 2009. Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. *In Proceedings of NAACL '09, The 2009 Annual Conference of the North American Chapter of the Association for Computational*

Linguistics, pp. 19—27.

Abdulrahman Almuhareb and Massimo Poesio. 2005. Concept Learning and Categorization from the Web. In *Proceedings of the Annual Meeting of the Cognitive* Science *Society*, Italy, July.

John Baer. Gender Differences. In Mark A. Runco, Steven R. Pritzker (Eds.), *Encyclopedia of Creativity*, Volume I. New York: Academic Press.

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

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics,* 32(1):13-47.

Edward de Bono. 1970. *Lateral thinking: creativity step by step*. New York: Harper & Row.

Thomas Erl. 2008. *SOA: Principles of Service Design.* Prentice Hall.

Christiane Fellbaum (ed.). 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

J. Paul Guilford. 1967. *The Nature of Human Intelligence*. New York: McGraw Hill.

Lushan Han, Tim Finin, Paul McNamee, Anupam Joshi and Yelena Yesha. 2012. Improving Word Similarity by Augmenting PMI with Estimates of Word Polysemy. *IEEE Transactions on Data and Knowledge Engineering* (13 Feb. 2012).

Jay Y. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th International Conference on Research in Computational Linguistics*, pp. 19-33.

George Karypis. 2002. CLUTO: A clustering toolkit. *Technical Report 02-017*,

University of Minnesota. http://www-users.cs.umn.edu/~karypis/cluto/.

Zornitsa Kozareva, Eileen Riloff and Eduard Hovy. 2008. Semantic Class Learning from the Web with Hyponym Pattern Linkage Graphs. *In Proc. of the 46th Annual Meeting of the ACL,* pp 1048-1056.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In Fellbaum, C. (ed.), *WordNet: An Electronic Lexical Database,* 265–283.

Yuhua Li, Zuhair A. Bandar and David McLean. 2003. An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 4. 871-882.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th ICML, the International Conference on Machine Learning*, Morgan Kaufmann, San Francisco CA, pp. 296– 304.

Michael Lesk. 1986 Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of ACM SigDoc, ACM*, 24–26.

George A. Miller and Walter. G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes* **6**(1):1-28.

Andrew Ortony. 1979. Beyond literal similarity. *Psychological Review, 86,* 161-180.

Masahiro Mori. 2005. On the Uncanny Valley. In *Proceedings of the Humanoids-2005 workshop: Views of the Uncanny Valley*, Tsukuba, Japan.

Ted Pederson, Siddarth Patwardhan and Jason Michelizzi. 2004. WordNet::Similarity: measuring the relatedness of concepts. In *Proceedings of HLT-NAACL'04*

(Demonstration Papers) the 2004 annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 38-41.

Philip Resnick. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of IJCAI'95, the 14th International Joint Conference on Artificial Intelligence.*

Nuno Seco, Tony Veale and Jer Hayes, 2004. An Intrinsic Information Content Metric for Semantic Similarity in WordNet. In *Proceedings of ECAI'04*, *the European Conference on Artificial Intelligence.*

Michael Strube and Simone Paolo Ponzetto. 2006. WikiRelate! Computing Semantic Relatedness Using Wikipedia. In *Proceedings of AAAI-06, the 2006 Conference of the Association for the Advancement of AI,* pp. 1419–1424.

Ellis P. Torrance. 1980. Growing Up Creatively Gifted: The 22-Year Longitudinal Study. *The Creative Child and Adult Quarterly*, 3, 148-158

Tony Veale. 2003. The analogical thesaurus: An emerging application at the juncture of lexical metaphor and information retrieval. In *Proceedings of IAAI'03, the 15th International Conference on Innovative Applications of AI*, Acupulco, Mexico.

Tony Veale, Guofu Li and Yanfen Hao. 2009. Growing Finely-Discriminating Taxonomies from Seeds of Varying Quality and Size. In *Proceedings of EACL'09, the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Athens, Greece, 835-842.

Tony Veale. 2011. Creative Language Retrieval: A Robust Hybrid of Information Retrieval and Linguistic Creativity. In Proceedings of ACL'2011, the 49th Annual Meeting of the Association for Computational Linguistics, Jeju, South Korea.

Tony Veale. 2012a. Exploding the Creativity Myth: The Computational Foundations

of Linguistic Creativity. *London: Bloomsbury Academic.*

Tony Veale. 2012b. Seeing the Best and Worst of Everything on the Web with a Two-level, Feature-rich Affect Lexicon. In Proceedings of WWW'2012, the 21st World-Wide-Web conference, Lyon, France.

Tony Veale. 2013a. A Service-Oriented Architecture for Computational Creativity. *Journal of Computing Science and Engineering, 7(3):159-167.*

Tony Veale. 2013b. Linguistic Readymades and Creative Reuse. *Transactions of the SDPS: Journal of Integrated Design and Process Science, 17(4):37-51.*

Paul A. Wilson, Barbara Lewandowska-Tomaszczyk. 2014. Affective Robotics: Modelling and Testing Cultural Prototypes. *Cognitive Computation*, Vol. 6, no. 4, pp 814-840.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of ACL'94, 32nd annual meeting of the Association for Computational Linguistics, Las Cruces, New Mexico,*. pp. 133-138.

Figure 1. The output of Thesaurus Rex for the input divorce & war. The size of each fine-grained perspective reflects its Web frequency (and thus salience) for "war". Readers are invite to generate outputs of their own by visiting the Rex service online at *http://boundinanutshell.com/therex2/*