From Conceptual Mash-ups to Bad-ass Blends: A Robust Computational Model of Conceptual Blending

Tony Veale

Abstract Conceptual blending is a complex cognitive phenomenon whose instances range from the humdrum to the pyrotechnical. Most remarkable of all is the ease with which we humans regularly understand and produce complex blends. While this facility will doubtless elude our best efforts at computational modeling for some time to come, there are practical forms of conceptual blending that are amenable to computational exploitation right now. In this chapter we introduce the notion of a *conceptual mash-up*, a robust form of blending that allows a computer to creatively re-use and extend its existing common-sense knowledge of a topic. We show also how a repository of such knowledge can be harvested automatically from the Web, by targeting the casual questions that we pose to ourselves and to others every day. By acquiring its world knowledge from the questions of others, a computer can eventually learn to pose introspective questions of its own, in the service of its own creative *mash-ups*.

1 The Plumbing of Creative Thought

We can think of figurative comparisons as pipes that carry salient information from a source domain to a target domain. Some figurative pipes are thin, such as a simile that transfers just a single property from a source concept onto a target idea (Hao & Veale, 2010). Other pipes are fat, and so convey a good deal more information: think of the resonant metaphors (Veale, 2012; Veale, Shutova, & Klebanov, 2016) and evergreen analogies (Goel, 2017) that yield deeper meanings the more you look at them. By convention, most pipes carry information in one direction only, from the source domain to the target domain. But creativity is no respecter of convention, and creative comparisons are sometimes a two-way affair, in which aspects of the source

Tony Veale
School of Computer Science, University College Dublin, Ireland.
e-mail: tony.veale@gmail.com

and target are thoroughly mixed together in a back-and-forth exchange of ideas at the boundary of seemingly very different domains (Lavrač et al., 2017), to create something utterly new and imaginative. To appreciate the differences between these different kinds of figurative plumbing, consider the following excerpt from the script for the movie Jurassic Park, which captures an exchange between the park's creator, John Hammond, and a wry mathematician, Ian Malcolm, who has been asked to evaluate the park's viability before it is opened to the public. The park of the title is populated with genetically-engineered dinosaurs, so the dialogue takes place against a backdrop of carnivorous mayhem and rampant destruction:

> **John Hammond**: All major theme parks have delays. When they opened Disneyland in 1956, nothing worked!
>
> **Dr. Ian Malcolm**: Yeah, but, John, if The Pirates of the Caribbean breaks down, the pirates don't eat the tourists.

At this point in the movie, nothing is working in Jurassic Park, but nothing worked in 1956 at Disneyland either, and the latter turned out to be a huge financial and cultural success. Hammond thus frames *Disneyland* as a triumph, by focusing on the temporal sequence of events associated with its launch, its initial problems, and its eventual success. With this implicit analogy to Jurassic Park, whose launch has been plagued by unique problems of its own, Hammond predicts that his own troubled venture will follow the same script and achieve the same success. In effect, he sees *Disneyland* and *Jurassic Park* as two overlapping frames (much as in Lavrač et al. (2017)), and wants others to see the overlap too, so they might come to the same conclusions. Malcolm's rejoinder is also intended to be understood in the context of this analogy, but it is much more than an analogy as conceived in e.g. (Falkenhainer, Forbus, & Gentner, 1989; Gentner, 1983; Gentner, Falkenhainer, & Skorstad, 1989; Goel, 2017; Veale & Keane, 1997). It involves mapping, yes, so that *The Pirates of the Caribbean* is aligned with the attractions of *Jurassic Park* and the pirates of the former are mapped to the dinosaurs of the latter. But the salient behaviors of the latter - such as eating people willy-nilly - are also integrated with the protagonists of the former, to generate a counterfactual image of animatronic pirates eating tourists in mouse-earred caps. In the words of Fauconnier and Turner (Fauconnier & Turner, 1994, 2002), Malcolm has created a *blend* and is now *running the blend*: that is, he is conducting a mental simulation to explore the emergent possibilities that were hitherto just latent in the juxtaposition of both conceptual frames.

When the actor and writer Ethan Hawke was asked to write a profile of Kris Kristofferson for Rolling Stone magazine, Hawke had to create an imaginary star of his own to serve as an apt contemporary comparison. For Hawke, Brad Pitt is as meaningful a comparison as one can make, but even Pitt's star power is but a dim bulb to that of Kristofferson when he shone most brightly in the 1970s. To communicate just how impressive the singer-actor-activist would have seemed to an audience in 1979, Hawke assembled the following Frankenstein-monster from the body of Pitt and other assorted star parts:

> "Imagine if Brad Pitt had written a No. 1 single for Amy Winehouse, was considered among the finest songwriters of his generation, had been a Rhodes scholar, a U.S. Army Airborne Ranger, a boxer, a professional helicopter pilot and was as politically outspoken as Sean Penn. That's what a motherfuckin' badass Kris Kristofferson was in 1979."

Pitt comes off poorly in the comparison, but this is precisely the point: no contemporary star comes off well, because in Hawke's view, none has the wattage that Kristofferson had in 1979. The awkwardness of the comparison, and the fancifulness of the blended image, serves as a creative meta-description of Kristofferson's achievements. In effect Hawke is saying, "look to what lengths I must go to find a fair comparison for this man without peer". Notice also how salient information flows in both directions in this comparison. To create a more rounded comparison, Hawke finds it necessary to mix in a few elements from other stars (such as Sean Penn), and to also burnish Pitt's résumé with elements borrowed from Kristofferson himself. Most of this additional structure is imported literally from the target, as when we are asked to imagine Pitt as a boxer or a helicopter pilot. Other structure is imported in the form of an analogy: while Kristofferson wrote songs for Janis Joplin, Pitt is imagined as a writer for her modern counterpart, Amy Winehouse.

This Pitt 2.0 doesn't actually exist, of course. Like Ian Malcolm's view of Jurassic Park *qua* Disneyland, Hawke's description is a conceptual blend that constructs a whole new source concept in its own counterfactual space. Blending is pervasive in modern culture, and can be seen in everything from cartoons to movies to popular fiction, while the elements of a blend can come from any domain of experience, from classic novels to 140-character tweets to individual words. As defined by Fauconnier (1994, 1997) and Fauconnier and Turner (1994, 2002), conceptual blending combines the smoothness of metaphor with the structural complexity and organizing power of analogy. We can think of blending as a cognitive operation in which conceptual ingredients do not flow in a single direction, but are thoroughly stirred together, to create a new structure with its own emergent meaning. Moreover, a blend can itself be used as a component part in larger blends, to create pyrotechnical flourishes of language that dazzle and amaze but rarely over-tax our powers of conceptual analysis. Consider the following snarky comparison, freshly minted for Sam Mendes in the Guardian newspaper after studio bosses had chosen him to direct the 23^{rd} film in the *James Bond* franchise: "*Appearance: Like the painting in George Clooney's attic.*" This is not a simple comparison, but a complex blend that is loaded with figurative meaning, and we require neither a prior mental image of Sam Mendes nor a knowledge of the paintings in Clooney's attic to understand its real meaning. We can be quite certain that the picture in question is not a real picture that Clooney might actually own, whether *A Rake's Progress* or *Dogs Playing Poker*, but an entirely fictional painting that we create on the fly, via Fauconnier and Turner's process of conceptual blending. As Fauconnier and Turner might say, this is a multi-layered blend that must be unpacked in several stages. The blend exploits our familiarity with Oscar Wilde's *The Picture of Dorian Gray*, a morality tale concerning the fate of a handsome but narcissistic young man who pledges his soul so that his painted self might suffer the ravages of time in his stead. Dorian soon discovers that his portrait – the infamous "painting in the attic" – not only

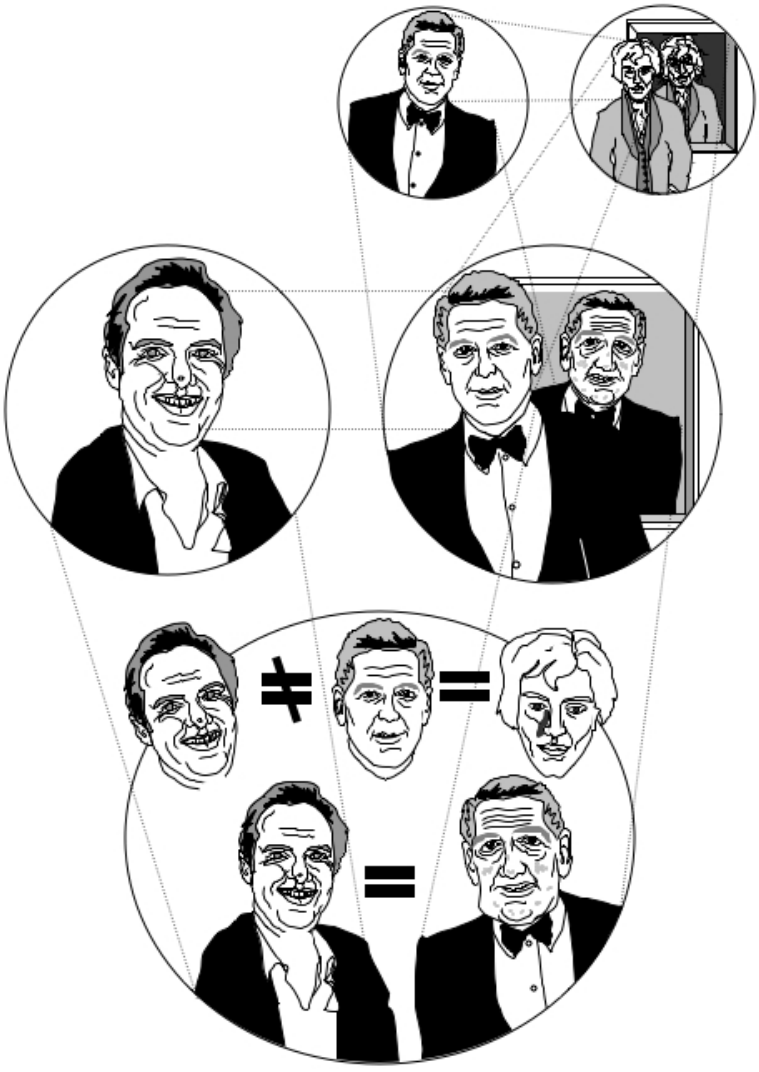

Fig. 1: The *blend-within-a-blend* that underpins The Guardian's comparison of director Sam Mendes to "the picture in George Clooney's attic" (reproduced from (Veale, 2012))

changes to reflect his true age, but also holds a mirror to his inner being. As Dorian descends into moral degeneracy, his painted counterpart suffers a more physical degeneration, for as Dorian's portrait becomes increasingly ugly to behold he himself remains preternaturally youthful.

In the right hands, any cliché can be revitalized in a well-turned phrase, and The Guardian breathes humorous new life into the Dorian Gray cliché-archetype by embedding it within two more topical nested blends. A visual representation of the workings of this blend-within-a-blend is provided in Fig. 1. The inner blend re-imagines Wilde's story with a new leading man, George Clooney, whose matinée-idol good looks make him an apt substitute for the handsome youth of the original tale. Clooney has maintained his status as a Hollywood sex-symbol for almost two decades, and he remains a regular fixture in the pages of celebrity gossip sheets. We find it easy to imagine a slowly decaying portrait in a dark corner of Clooney's attic, and even if the conceit has the tang of sour grapes, this just adds to its snarkily humorous effect. Note that this inner blend is more than a simile, a metaphor or an analogy, and it does more than compare George Clooney to Dorian Gray. Rather, it creates a new version of the morality tale with its very own star. In the world of the blend, Clooney really does have a portrait of his aging, sin-wracked face in the attic. This inner blend puts a new face on Wilde's tale, to create a new chunk from familiar elements that can now be reused in other blends, as though it had always existed in our cultural lexicons.

The *Mendes-as-Clooney-as-Dorian* and *Kristofferson-as-Pitt* blends show just how complex a blend can be, while nonetheless remaining intelligible to a reader: when we interpret these constructs, we are not aware of any special challenge being posed, or of any special machinery being engaged. Nonetheless, this kind of blend poses significant problems for our computers and their current cognitive/linguistic modeling abilities. So in this chapter we present a computational middle-ground, called a conceptual mash-up, that captures some of the power and utility of a conceptual blend, but in a form that is practical and robust to implement on a computer. From this starting point we can begin to make progress toward the larger goal of creative computational systems that - to use Hawke's word - can formulate truly badass blends of their own.

Creative language is a knowledge-hungry phenomenon. We need knowledge to create or comprehend an analogy, metaphor or blend, while these constructs allow us to stretch our knowledge into new forms and niches. But computers cannot be creative with language unless they first have something that is worth saying creatively, for what use is a poetic voice if one has no opinions or beliefs of one's own that need to be expressed? This current work describes a re-usable resource - a combination of knowledge and of tools for using that knowledge - that can allow other computational systems to form their own novel hypotheses from mash-ups of common stereotypical beliefs. These hypotheses can be validated in a variety of ways, such as via a Web search, and then expressed in a concise and perhaps creative linguistic form, such as in poem, metaphor or riddle. The resource, which is available as a public web service called *Metaphor-Eyes*, produces conceptual mash-ups for its input concepts, and returns the resulting knowledge structures in an XML format that can then be used by other computational systems in a modular, distributed fashion. The *Metaphor-Eyes* service is based on an approach to creative introspection first presented in Veale and Li (2011), in which stereotypical beliefs about everyday concepts are acquired from the Web, and then blended on demand to create hypothe-

ses about topics that the computer may know little or nothing about. We present the main aspects of *Metaphor-Eyes* in the following sections, and show how its capacity for conceptual mash-ups can be exploited by other systems via the Web.

1.1 Structure of this Chapter

Our journey begins in the next section, with a brief overview of relevant computational work in the areas of metaphor and blending. It is our goal to avoid hand-crafted representations, so in the section after that we describe how the system can acquire its own common-sense knowledge from the web, by eavesdropping on the revealing questions that users pose everyday to a search engine such as Google. This knowledge provides the basis for conceptual mash-ups, which are constructed by re-purposing web questions to form new introspective hypotheses about a topic. We also introduce the notion of a *multi-source mash-up*, which allows us to side-step the vexing problem of context and user-intent in the construction of conceptual blends. Finally, an empirical evaluation of these ideas is presented, and the chapter concludes with thoughts on future directions.

2 Related Work and Ideas

We use metaphors and blends not just as rhetorical flourishes, but as a basis for extending our inferential powers into new domains (Barnden, 2006). Indeed, work on analogical metaphors shows how metaphor and analogy use knowledge to create knowledge (Veale et al., 2016). *Structure-Mapping Theory*, or SMT (Gentner, 1983; Gentner et al., 1989), argues that analogies allow us to impose structure on a poorly-understood domain, by mapping knowledge from one that is better understood. SME, the *Structure-Mapping Engine* (Falkenhainer et al., 1989) implements these ideas by identifying sub-graph isomorphisms between two mental representations (see also Veale and Keane (1997) for a discussion of why this task is NP-Hard). SME then projects connected sub-structures from the source to the target domain. SMT prizes analogies that are systematic in their preservation of causal structure (see also Winston (1980) and Carbonell (1982)), yet a key issue in any structural approach is how a computer can acquire structured representations for itself.

Veale and O'Donoghue (2000) proposed an SMT-based model of conceptual blending that was perhaps the first computational model of the phenomenon. The model, called *Sapper*, addresses many of the problems faced by SME - such as deciding for itself which knowledge is relevant to a blend - but succumbs to others, such as the need for a hand-crafted knowledge base. Pereira (2007) and Martins, Pereira, and Cardoso (2017) present an alternative computational model that combines SMT with other computational techniques, such as using genetic algorithms to search the space of possible blends. Pereira's model was applied both to lin-

guistic problems (such as the interpretation of novel noun-noun compounds) and to visual problems, such as the generation of novel monsters/creatures for video games. Nonetheless, Pereira's approach was just as reliant on hand-crafted knowledge. To explore the computational uses of blending without such a reliance on specially-crafted knowledge, Veale (2006) showed how blending theory can be used to understand novel portmanteau words - or "formal" blends - such as "Feminazi" (*Femi*nist + *Nazi*). This approach, called *Zeitgeist*, automatically harvested and interpreted portmanteau blends from Wikipedia, using only the topology of Wikipedia itself and the contents of Wordnet (Fellbaum, 1998) as resources.

The availability of large corpora and the web suggests a means of relieving the knowledge bottleneck that afflicts computational models of metaphor, analogy and blending. Turney and Littman (2005) show how a statistical model of relational similarity can be constructed from web texts for handling proportional analogies of the kind used in SAT and GRE tests. No hand-coded or explicit knowledge is employed, yet Turney and Littman's system achieves an average human grade on a set of 376 SAT analogies (such as *mercenary:soldier::?:?* where the best answer among four alternatives is *hack:reporter*). Almuhareb and Poesio (2004) describe how attributes and values can be harvested for word-concepts from the web, showing how these properties allow lexical concepts to be clustered into category structures that replicate the semantic divisions made by a curated resource such as WordNet (Fellbaum, 1998). Veale and Hao (2007a, 2007b, 2008) describe how stereotypical knowledge can be acquired from the web by harvesting similes of the form "*as P as C*" (as in "as *smooth* as *silk*"), and go on to show, in Veale (2012), how a body of 4000 stereotypes is used in a web-based model of metaphor generation and comprehension.

Shutova, Sun, and Korhonen (2010) combined elements of several of these approaches. They annotated verbal metaphors in corpora (such as "to stir excitement", where the verb "stir" is used metaphorically) with the corresponding conceptual metaphors identified by Lakoff and Johnson (1980) and listed in Lakoff, Espenson, and Schwartz (1991). Statistical clustering techniques were then used to generalize from the annotated exemplars, allowing their system to recognize other metaphors in the same vein (e.g. "he swallowed his anger"). These clusters can also be analyzed to suggest literal paraphrases for a given metaphor (such as "to provoke excitement" or "suppress anger"). This approach is noteworthy for the way it operates with Lakoff and Johnson's inventory of conceptual metaphors without actually using any explicit knowledge of its own.

The questions people ask, and the web queries they pose, are an implicit source of common-sense knowledge. The challenge we face as computationalists lies in turning this *implicit* world knowledge into *explicit* representations. For instance, Pasca and Durme (2007) show how knowledge of classes and their attributes can be extracted from the queries that are processed and logged by web search engines. We intend to show in this chapter how a common-sense representation that is derived from web questions can be used in a model of conceptual blending. We focus on well-formed questions, found either in the query logs of a search engine or harvested from documents on the web. These questions can be viewed as atomic properties of their topics, but they can also be parsed to yield logical forms for reasoning. We

show here how we might, by representing topics via the questions we ask about them, also grow our knowledge-base via blending, by posing these questions introspectively of other topics as well.

3 "Milking" Knowledge from the Web

Amid the ferment and noise of the world-wide-web sit nuggets of stereotypical world knowledge, in forms that can be automatically harvested. To acquire a property P for a topic T, one can look for explicit declarations of T's *P-ness*, but such declarations are rare, as speakers are loathe to explicitly articulate truths that are tacitly assumed by others. Hearst (1992) observed that the best way to capture tacit truths in large corpora (or on the web) is to look for stable linguistic constructions that presuppose the desired knowledge. So rather than look for "*all Xs are Ys*", which is a laudably direct but exceedingly rare pattern in everyday usage, more frequent *Hearst*-patterns such as "*Xs and other Ys*" presuppose exactly the same hypernymic relations. By mining presuppositions rather than declarations, a harvester can cut through the layers of noise and misdirection that are endemic to the web.

If W is a count noun denoting a topic T_W, then the query "why do W_{plural} *" allows us to retrieve questions posed about T_W on the web, in this case via the Google API. (If W is a mass noun or a proper-name, we can instead use the query "why does W * ?") These two formulations show the benefits of using questions as extraction patterns: a query is framed by an opening WH-question word and a closing question mark, ensuring that a complete statement is retrieved (Google snippets often contain sentence fragments); and number agreement between "do"/"does" and W suggests that the question is syntactically well-formed (good grammar helps discriminate well-formed musings from random noise). Queries with the subject T_W are dispatched whenever the system wishes to learn about a topic T. We ask the Google API to return 200 snippets per query, which are then parsed to extract well-formed questions and their logical forms. Questions that cannot be parsed in this way are rejected as being too complex for later re-use in conceptual blending.

For instance, the topic *Pirate* yields the query "why do pirates *" which can be used to retrieve snippets about pirates. The retrieval set includes these questions:

Why do pirates wear eye patches?
Why do pirates hijack vessels?
Why do pirates have wooden legs?

Parsing the second question above, we obtain its logical form:

$$\forall x\ pirate(x) \rightarrow \exists y\ vessel(y) \wedge hijack(x,y)$$

A computational system needs a critical mass of such commonsense knowledge before it can be usefully applied to problems such as conceptual blending. Ideally, we could extract a large body of everyday musings from the query logs of a search

engine like Google, since many users persist in using full NL questions as web queries. Yet such logs are jealously guarded, not least on concerns about privacy. Nonetheless, engines like Google do expose the most common queries in the form of text completions: as one types a query into the search box, Google anticipates the user's query by matching it against past queries, and offers a variety of popular completions. These completions are a rich source of knowledge for a machine.

In an approach we call Google "milking", we coax completions from the Google search box for a long list of strings with the prefix "why do", such as "why do *a*" (which prompts "why do *a*nimals hibernate?"), and "why do *aa*' (which prompts "why do *aa* batteries leak?"). We use a manual trie-driven approach, using the input "why do *X*" to determine if any completions are available for a topic prefixed with *X*, before then drilling deeper with "why do *Xa*" ... "why do *Xz*". Though laborious, this process taps into a veritable mother lode of nuggets of conventional wisdom. Two weeks of milking yields approximately 25,000 of the most common questions on the web, for over 2000 topics, providing critical mass for the processes to come.

4 Conceptual "Mash-Ups"

Google milking yields these frequent "Why do ..." questions about poets:

Why do poets repeat words?
Why do poets use metaphors?
Why do poets use alliteration?
Why do poets use rhyme?
Why do poets use repetition?
Why do poets write poetry?
Why do poets write about love?

Querying the web directly, the system finds other common presuppositions about poets, such as "*why do poets die poor?*" and "*why do poets die young?*", precisely the kind of knowledge that shapes our stereotypical view of poets yet which one is unlikely to see reflected in a dictionary's entries. Suppose a user asks the system to explore the ramifications of the blend *Philosophers are Poets*: this prompts the system to introspectively ask "*how are philosophers like poets?*". This question spawns others, which are produced by replacing the subject of the poet-specific questions above, yielding new introspective musings such as "*do philosophers write poetry?*", "*do philosophers use metaphors?*", and "*do philosophers write about love?*"

Each repurposed question can be answered by again appealing to the web: the system simply looks for evidence that the hypothesis in question (such as "philosophers use metaphors") is attested by literal usage in one or more web texts. The Google API finds supporting matches for the following hypotheses: "*philosophers die poor*" (3 hits), "*philosophers die young*" (6 hits), "*philosophers use metaphors*" (156 hits), and "*philosophers write about love*" (just 2 hits). The goal is not to show that these behaviors are as salient for philosophers as for poets, merely that they are attested to be meaningful for philosophers too. We refer to the construct *Philoso-*

phers are Poets as a *conceptual mash-up*, since knowledge about a source concept, *Poet*, has been mashed-up with that of a target idea, *Philosopher,* to yield a new knowledge network for the latter. Conceptual mash-ups are a specific kind of conceptual blend, one that is easily constructed via simple computational processes.

To generate a mash-up, the system starts from a given target idea T and searches for the source concepts $S_1 \ldots S_n$ that might plausibly yield a meaningful blend. A locality assumption limits the scale of the search space for $S_1 \ldots S_n$, by assuming that T must exhibit a pragmatic similarity to any source concept S_i. Budanitsky and Hirst (Budanitsky & Hirst, 2006) describe a raft of term-similarity measures based on WordNet (Fellbaum, 1998), but what is needed for blending is a generative measure: one that can quantify the similarity of T to S as well as suggest a range of likely S_i's for any given topic T. We construct such a measure via corpus analysis, since a measure trained on corpora can easily be made corpus-specific and thus domain- or context-specific. The Google ngrams (Brants & Franz, 2006) provide a large collection of word sequences from web texts. Looking to the 3-grams, we extract coordinations of generic nouns of the form "Xs and Ys". For each coordination, such as "tables and chairs" or "artists and scientists", X is considered a pragmatic (rather than semantic) neighbor of Y, and vice versa. When identifying blend sources for a topic T, we consider the neighbors of T as candidate sources for a blend. Furthermore, if we consider the neighbors of T to be features of T, then a vector space representation for topics can be constructed, such that the vector for a topic T contains all of the neighbors of T that are identified in the Google 3-grams. This vector representation allows us to calculate the similarity of a topic T to a source S, and rank the neighbors $S_1 \ldots S_n$ of T by their similarity to T (Veale & Li, 2013).

Intuitively, writers use the pattern "Xs and Ys" to denote an ad-hoc category, so that the topics linked by this pattern are not just similar but truly comparable, or even interchangeable. Potential sources for T are ranked by their perceived similarity to T, as described above. So if generating mash-ups for *Philosopher*, the top-ranked sources found in the Google 3-grams are: *Scholar, Epistemologist, Ethicist, Moralist, Naturalist, Scientist, Doctor, Pundit, Savant, Explorer, Intellectual* and *Lover*.

4.1 Multi-Source Mash-Ups

The problem of finding good sources for a topic T is highly under-constrained, and depends on the contextual goals of the speaker. However, when blending is used for knowledge acquisition, multi-source mash-ups allow us to blend a range of sources into a rich, context-free structure. If $S_1 \ldots S_n$ are the n closest neighbors of T as ranked by similarity to T, then a mash-up can be constructed to describe the semantic potential of T by collating all of the questions from which the system derives its knowledge of $S_1 \ldots S_n$, and by repurposing each question for T. A complete mash-up collates questions from all the neighbors of a topic, while a 10-neighbor mash-up for *Philosopher*, say, would collate all the questions associated with the top 10 neighbors *Scholar* ... *Explorer* and insert "philosopher" as the subject of each. In

this way a conceptual picture of *Philosopher* could be created, by drawing on beliefs such as that naturalists tend to be pessimistic and humanists care about morality.

A 20-neighbor mash-up for *Philosopher* would also integrate the system's knowledge of *Politician* into this picture, to suggest the possibilities that e.g. *philosophers lie, philosophers cheat, philosophers equivocate* and even that *philosophers have affairs* and *philosophers kiss babies*. Each of these hypotheses can be put to the test in the form of a specific web query; thus, the hypotheses "philosophers lie" (586 Google hits), "philosophers cheat" (50 hits) and "philosophers equivocate" (11 hits) are all validated with Google queries, whereas "philosophers kiss babies" (0 hits) and "philosophers have affairs" (0 hits) are not. As one might expect, the most domain-general hypotheses show the greatest promise of taking root in a target domain. For example, "why do artists use Macs?" is more likely to be successfully transplanted into another domain than "why do artists use perspective drawing?"

The generality of a question is related to the number of times it appears in our knowledge-base with different subjects. Thus, "why do $< Xs >$ wear black" appears 21 times, while "why do $< Xs >$ wear black hats" and "why do $< Xs >$ wear white coats" each just appear twice. When a mash-up for a topic T is presented to the user, each imported question Q is ranked according to two criteria: Q_{count}, the number of neighbors of T that suggested Q as a hypothesis for T; and Q_{sim}, the similarity of T to its most similar neighbor that suggested Q (as calculated using a WordNet-based metric; e.g., see (Budanitsky & Hirst, 2006) for a survey; we use the metric in Veale and Li (2013) here). Both combine to give the single salience measure $Q_{salience}$, which is defined as follows:

$$Q_{salience} = Q_{sim} \times Q_{count} / (Q_{count} + 1)$$

Note that Q_{count} is always greater than 0, since each question Q must be suggested by at least one neighbor of T. Note also that salience is a measure of expectedness (e.g. see Grace and Maher (2017)) and thus of plausibility too, so when Q_{count} is large then so is $Q_{salience}$. It is time-consuming to dispatch every question in a mash-up to the web, as a mash-up of m questions requires m web queries. It is more practical to choose a cut-off n and simply test the top n questions, as ranked by $Q_{salience}$. In the next section we evaluate the ranking of questions in a mash-up, and estimate the likelihood of successful knowledge transfer from one topic to another.

5 Empirical Evaluation

Our corpus-attested, neighborhood-based approach to similarity does not use WordNet (Fellbaum, 1998), but is capable of replicating the same semantic divisions made by WordNet. In earlier work, Almuhareb and Poesio (2004) extracted features for concepts from text-patterns instantiated on the web. Those authors tested the efficacy of the extracted features by using them to cluster 214 words taken from 13 semantic categories in WordNet (henceforth, we denote this experimental setup as

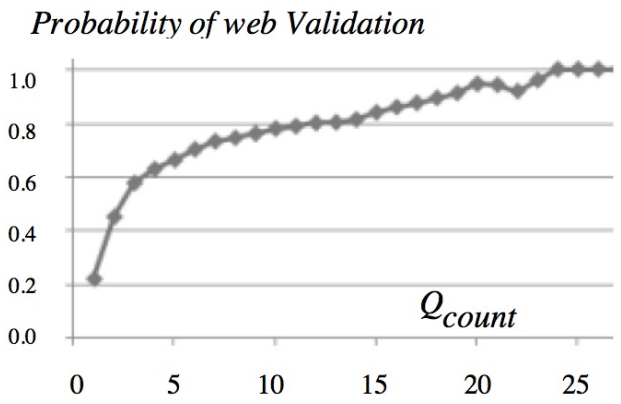

Fig. 2: Likelihood of a hypothesis in a mash-up being validated via web search (y-axis) for hypotheses that are suggested by Q_{count} neighbors (x-axis).

AP214), and reported a cluster purity of 0.85 in replicating the category structures of WordNet. But if the neighbors of a term are instead used as features for that term, and if a term is also considered to be its own neighbor, then an even higher purity/accuracy of 0.934 is achieved on *AP214*. Using neighbors as features in this way requires a vector space of just 8,300 features for *AP214*, whereas Almuhareb and Poesio's original approach to AP214 used approximately 60,000 features.

The locality assumption underlying this notion of a pragmatic neighborhood constrains the number of sources that can contribute to a multi-source mash-up. Knowledge of a source S can be transferred to topic T only if S and T are neighbors, as identified via corpus analysis. Yet, the Google 3-grams suggest a wealth of neighboring terms, so locality does not unduly hinder the transfer of knowledge. Consider a test-set of 10 common terms, *Artist, Scientist, Terrorist, Computer, Gene, Virus, Spider, Vampire, Athlete* and *Camera*, where knowledge harvested for each of these terms is transferred via mash-ups to all of their neighbors. For instance, "why do artists use Macs?" suggests "musicians use Macs" as a hypothesis because artists and musicians are close neighbors, semantically (in WordNet) and pragmatically (in the Google n-grams); this hypothesis is in turn validated by 5,700 web hits for "musicians use Macs". In total, 410,000 hypotheses are generated from these 10 test terms, and when posed as web queries to validate their content, approximately 90,000 (21%) hypotheses are validated by at least one attested use on the web.

Just as knowledge tends to cluster into pragmatic neighborhoods, hypotheses likewise tend to be validated in clusters. As shown in Fig. 2, the probability that a hypothesis is valid for a topic T grows with the number of neighbors of T for which it is known to be valid (that is, Q_{count}). Unsurprisingly, the closest neighbors with the highest similarity to the topic exert the most influence. Fig. 3 shows that the probability of a hypothesis for a topic being validated by attested web usage grows with the number of the topic's neighbors that suggest it and its similarity to the

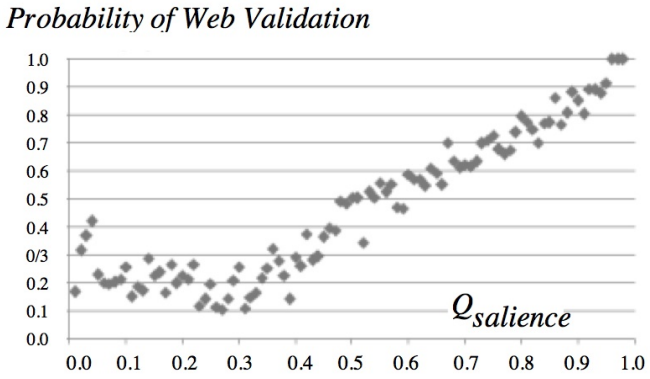

Fig. 3: Likelihood of a hypothesis in a mash-up being validated via web search (y-axis) for hypotheses with a particular $Q_{salience}$ measure (x-axis).

closest of these neighbors (that is, $Q_{salience}$). In absolute terms, hypotheses perceived to have high salience (e.g. $> .6$) are much less frequent than those with lower ratings. So a more revealing test is the ability of the system to rank the hypotheses in a mash-up so that the top-ranked hypotheses have the greatest likelihood of being validated on the web. That is, to avoid information overload, the system should be able to distinguish the most plausible hypotheses from the least plausible, just as search engines like Google are judged on their ability to push the most relevant hits to the top of their rankings.

Fig. 4 shows the average rate of web validation for the top-ranked hypotheses (ranked by salience) of complete mash-ups generated for each of our 10 test terms from all of their neighbors. Since these are common terms, they have many neighbors that suggest many hypotheses. On average, 85% of the top 20 hypotheses in each mash-up are validated by web search as plausible, while just 1 in 4 of the top 60 hypotheses in a mash-up are not validated by attested usage in web documents. Figs. 2 – 4 show that the system is capable of acquring knowledge from the web that can be successfully transferred to neighboring terms via metaphors and mash-ups, and then meaningfully ranked by salience. But just how useful is this knowledge? To determine if it is the kind of knowledge that is useful for categorization, and thus the kind that captures the essence of a concept, we use it to replicate the AP214 test of Almuhareb and Poesio (2004). Recall that AP214 tests the ability of a feature-set to support the category distinctions imposed by WordNet, so that the 214 words can be clustered back into the 13 WordNet categories from whence they came.

So for each of these 214 words, we harvest questions from the web, and treat each question body as an atomic feature of its subject; thus, for example, we treat "kisses babies" as a feature of *Politician*. Clustering over these features alone offers poor accuracy when reconstructing WordNet categories, yielding a cluster purity of just over 0.5. One AP214 category in particular, comprising time units such as *week* and *year*, offers no traction to the question-based approach, and accuracy / purity

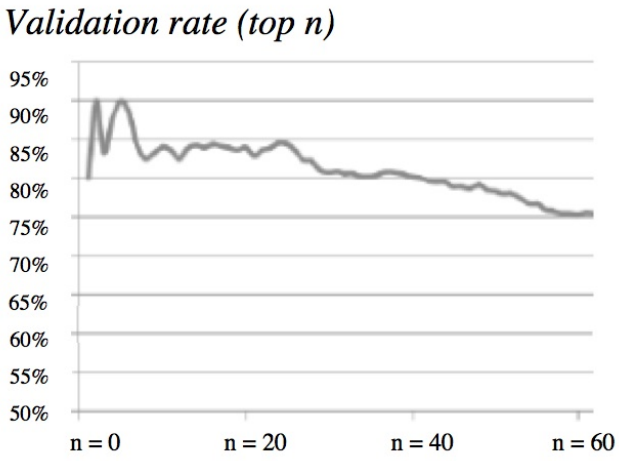

Fig. 4: Average percentage of the top-n hypotheses in a mash-up (as ranked by $Q_{salience}$) that are validated by web search.

increases to 0.6 when this category is excluded. People, it seems, rarely question the conceptual status of an abstract temporal unit on the web. Yet as knowledge is gradually transferred to the terms in AP214 from their corpus-attested neighbors, so that each term is represented as a conceptual mash-up of its *n* nearest neighbors, categorization markedly improves. Fig. 5 demonstrates the increasing accuracy of the system on AP214 (excluding the vexing *time* category) when using mash-ups of increasing numbers of neighbors. Blends really do bolster our knowledge of a topic with insights that are relevant to categorization.

6 Conclusions

We have explored how the most common questions on the web can provide the world knowledge needed to drive a robust, if limited, form of blending called a conceptual mash-up. The ensuing powers of self-questioning introspection, though basic, can be used to speculate upon the conceptual make-up of any given topic, not only in individual metaphors but in rich, informative mash-ups of multiple concepts. The world-wide-web is central to this approach: not only are questions harvested from the web (e.g., via Google "milking"), but newly-formed hypotheses are validated by means of simple web queries. The approach is practical, robust and quantifiable, and uses an explicit knowledge representation that can be acquired on demand for a given topic. Most importantly, the approach makes a virtue of blending, and argues that we and our machines should view blending not just as a problem to be solved, but as a tool of creative computational engineering.

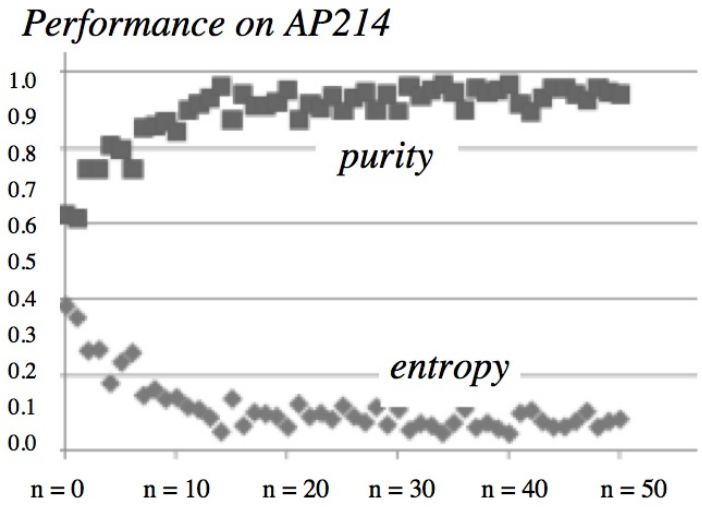

Fig. 5: Clustering performance on AP214 improves (that is, purity is higher and entropy is lower) as knowledge is transferred from the *n* closest neighbors of a term.

The ideas described here have been computationally realized in a web application and web service called *Metaphor-Eyes*. Fig. 6 provides a snapshot of the service in action. The user enters a query - in this case the provocative assertion "Google is a cult" – and the service provides an interpretation based on a mash-up of its knowledge of the source concept (*cults*) and of the target concept (*Google*). Two kinds of knowledge are used to provide the interpretation of Fig. 6. The first is common-sense knowledge of cults, of the kind that we expect most adults to possess. This knowledge includes widely-held stereotypical beliefs such as that cults are lead by gurus, that they worship gods and enforce beliefs, and that they recruit new members, especially celebrities, which often act as apologists for the cult. The system possesses no stereotypical beliefs about Google, but using the Google 2-grams (somewhat ironically, in this case), it can find linguistic evidence for the notions of a "Google guru", a "Google god" and a "Google apologist". The corresponding stereotypical beliefs about cults are then projected into the new blend space of *Google-as-a-cult*.

Metaphor-Eyes derives a certain robustness from its somewhat superficial treatment of blends as mash-ups. In essence, the system manipulates conceptual-level objects (ideas and other blends) by using language-level objects (strings, phrases, collocations) as proxies: a combination at the concept-level is deemed to make sense if a corresponding combination at the language-level can be found in a web corpus (or in the Google n-grams). As such, any creativity exhibited by the system is often facile or glib. Because the system looks for conceptual novelty in the veneer of surface language, it follows in the path of humour systems that attempt to generate interesting semantic phenomena by operating at the level of words and their con-

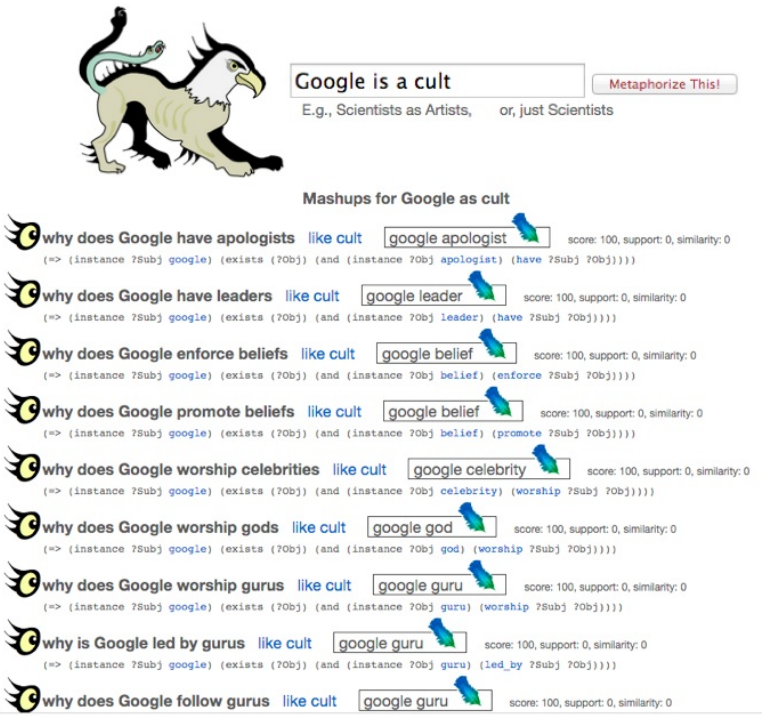

Fig. 6: A screen-shot from the computational system *Metaphor-Eyes*, which implements the model described in this chapter. *Metaphor-Eyes* shows how we can use conceptual mash-ups to explore counterfactual amd/or hybrid ideas and thus stimulate human creativity. (Note: Because the system has no prior ontological knowledge about Google, each entry above shows a default score of 100 and a support/similarity measure of 0). Visit *http://Afflatus.UCD.ie* to interact with the *Metaphor-Eyes* system for yourself, or to find out more about the system's XML functionality.

ventional significations (see Gatti, Ozbal, Guerini, Stock, and Strapparava (2017) for other work in this vein).

We have thus delivered on just one half of the promise of our title. While conceptual mash-ups are something a computer can handle with relative ease, "bad-ass" blends of the kind discussed in the introduction still lie far beyond our computational reach. Nonetheless, we believe the former provides a solid foundation for development of the tools and techniques that are needed to achieve the latter. Several areas of future research suggest themselves in this regard, and one that appears most promising at present is the use of mash-ups in the generation of poetry (see Veale (2013) for work in this direction). The tight integration of surface-form and meaning that is expected in poetry means this is a domain in which a computer can serendipitously allow itself to be guided by the possibilities of word combination

while simultaneously exploring the corresponding idea combinations at a deeper level (see (Gervás, 2017) for an exploration of key issues in computational poetry generation). Indeed, the superficiality of mash-ups makes them ideally suited to the surface-driven exploration of deeper levels of meaning.

Metaphor-Eyes should thus be seen as a community resource thru which the basic powers of creative introspection (as first described in (Veale & Li, 2011)) can be made available to a wide variety of third-party computational systems. In this regard, *Metaphor-Eyes* is a single instance of what will hopefully become an established trend in the maturing field of computational creativity: the commonplace sharing of resources and tools, perhaps as a distributed network of creative web services (Veale, 2014), that will promote a wider cross-fertilization of ideas in our field. The integration of diverse services and components will in turn facilitate the construction of systems with an array of creative qualities. Only by pooling resources in this way can we hope to go beyond one-note systems and produce the impressive multi-note "badass blends" of the title.

Acknowledgements This research was in part supported by the EC project WHIM, *The What-If Machine*, and by the EC project PROSECCO, a coordination action funded to *PROmote the Scientific Exploration of Computational Creativity.*

References

Almuhareb, A., & Poesio, M. (2004, June). Attribute-based and value-based clustering: An evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing (emnlp)* (pp. 158–165). Association for Computational Linguistics.

Barnden, J. (2006). Artificial intelligence, figurative language and cognitive linguistics. In G. Kristiansen, M. Achard, R. Dirven, & F. J. R. de Mendoza Ibanez (Eds.), *Cognitive linguistics: Current application and future perspectives* (p. 431-459). Berlin: Mouton de Gruyter.

Brants, T., & Franz, A. (2006). *Web 1t 5-gram version 1*. Linguistic Data Consortium.

Budanitsky, A., & Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, *32*(1), 13–47.

Carbonell, J. (1982). Metaphor: An inescapable phenomenon in natural language comprehension. In W. Lehnert & M. Ringle (Eds.), *Strategies for natural language processing* (pp. 415–434). Lawrence Erlbaum.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). Structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, *41*, 1-63.

Fauconnier, G. (1994). *Mental spaces: aspects of meaning construction in natural language*. Cambridge University Press.

Fauconnier, G. (1997). *Mappings in thought and language*. Cambridge University Press.

Fauconnier, G., & Turner, M. (1994). *Conceptual projection and middle spaces (technical report 9401)* (Tech. Rep.). University of California at San Diego, Department of Computer Science.

Fauconnier, G., & Turner, M. (2002). *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. New York: Basic Books.

Fellbaum, C. (Ed.). (1998). *Wordnet: An electronic lexical database (isbn: 0-262-06197-x)* (First ed.). MIT Press.

Gatti, L., Ozbal, G., Guerini, M., Stock, O., & Strapparava, C. (2017). Computer-supported human creativity and human-supported computer creativity. In T. Veale & F. A. Cardoso (Eds.), *Computational creativity: The philosophy and engineering of autonomously creative systems.* Springer International Publishing.

Gentner, D. (1983). Structure-mapping: A theoretical framework. *Cognitive Science*, *7*(2), 155170.

Gentner, D., Falkenhainer, B., & Skorstad, J. (1989). Metaphor: The good, the bad and the ugly. In Y. Wilks (Ed.), *Theoretical issues in nlp.* Hillsdale, NJ.: Lawrence Erlbaum Associates.

Gervás, P. (2017). Exploring quantitative evaluations of the creativity of automatic poets. In T. Veale & F. A. Cardoso (Eds.), *Computational creativity: The philosophy and engineering of autonomously creative systems.* Springer International Publishing.

Goel, A. (2017). Revisiting design, analogy, and creativity. In T. Veale & F. A. Cardoso (Eds.), *Computational creativity: The philosophy and engineering of autonomously creative systems.* Springer International Publishing.

Grace, K., & Maher, M. L. (2017). Placing expectation at the centre of computational creativity evaluation. In T. Veale & F. A. Cardoso (Eds.), *Computational creativity: The philosophy and engineering of autonomously creative systems.* Springer International Publishing.

Hao, Y., & Veale, T. (2010). An ironic fist in a velvet glove: Creative misrepresentation in the construction of ironic similes. *Minds and Machines*, *20*(4), 483-488.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on computational linguistics - volume 2* (pp. 539–545). Stroudsburg, PA, USA: Association for Computational Linguistics.

Lakoff, G., Espenson, J., & Schwartz, A. (1991). *The master metaphor list.* (Tech. Rep.). University of California at Berkeley.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

Lavrač, N., Juršič, M., Sluban, B., Perovšek, M., Urbančič, T., & Cestnik, B. (2017). Bisociative knowledge discovery for cross-domain literature mining. In T. Veale & F. A. Cardoso (Eds.), *Computational creativity: The philosophy and engineering of autonomously creative systems.* Springer International Publishing.

Martins, P., Pereira, F. C., & Cardoso, F. A. (2017). The nuts and bolts of con-

ceptual blending: Multi-domain concept creation with divago. In T. Veale & F. A. Cardoso (Eds.), *Computational creativity: The philosophy and engineering of autonomously creative systems.* Springer International Publishing.

Pasca, M., & Durme, B. V. (2007). What you seek is what you get: Extraction of class attributes from query logs. In *Proceedings of the 20th international joint conference on artificial intelligence* (pp. 2832–2837).

Pereira, F. C. (2007). *Creativity and artificial intelligence: a conceptual blending approach*. Berlin: Walter de Gruyter.

Shutova, E., Sun, L., & Korhonen, A. (2010). Metaphor identification using verb and noun clustering. In *Proceedings of coling 2010* (pp. 1002–1010). Beijing, China.

Turney, P. D., & Littman, M. L. (2005). Corpus-based learning of analogies and semantic relations. *Machine Learning*, *60*(1-3), 251-278.

Veale, T. (2006). Tracking the lexical zeitgeist with wikipedia and wordnet. In *Proceedings of ecai2006, the 17th european conference on artificial intelligence.* Trento, Italy: John Wiley.

Veale, T. (2012). *Exploding the creativity myth: The computational foundations of linguistic creativity*. London, UK: Bloomsbury.

Veale, T. (2013). Less rhyme, more reason: Knowledge-based poetry generation with feeling, insight and wit. In *Proceedings of iccc-2013, the 4th international conference on computational creativity.* Sydney, Australia.

Veale, T. (2014). A service-oriented architecture for metaphor processing. In *Proceedings of the second workshop on metaphor in nlp* (pp. 52–60). Baltimore, Maryland.

Veale, T., & Hao, Y. (2007a). Comprehending and generating apt metaphors: A web-driven, case-based approach to figurative language. In *Proceedings of aaai'2007, the 22nd conference of the association for the advancement of artificial intelligence.*

Veale, T., & Hao, Y. (2007b). Making lexical ontologies functional and context-sensitive. In *Proceedings of acl2007, the 46th annual meeting of the association for computational linguistics: Human language technologie.*

Veale, T., & Hao, Y. (2008). A fluid knowledge representation for understanding and generating creative metaphors. In *Proceedings of coling 2008* (pp. 945–952). Manchester, UK.

Veale, T., & Keane, M. (1997). The competence of sub-optimal structure mapping on hard analogies. In *Proceedings of ijcai97, the 15th international joint conference on artificial intelligence.* San Mateo, California: Morgan Kaufmann.

Veale, T., & Li, G. (2011). Creative introspection and knowledge acquisition: Learning about the world thru introspective questions and exploratory metaphors. In *Proceedings of aaai2011, the 25th conference of the association for the advancement of artificial intelligence.* AAAI press.

Veale, T., & Li, G. (2013). Creating similarity: Lateral thinking for vertical similarity judgments. In *Proceedings of acl 2013, the 51st annual meeting of the association for computational linguistics.*

Veale, T., & O'Donoghue, D. (2000). Computation and blending. *Cognitive Lin-*

guistics, *11*(3-4), 253-281.

Veale, T., Shutova, E., & Klebanov, B. B. (2016). *Metaphor: A computational perspective*. USA: Morgan Claypool: Synthesis Lectures on Human Language Technologies.

Winston, P. (1980). Learning and reasoning by analogy. *Communications of the ACM*, *23*(12), 689-703.