# FIGURE8: A Novel System for Generating and Evaluating Figurative Language

**Sarah Harmon**
Computer Science Department
University of California, Santa Cruz
Santa Cruz, CA 95064 USA
smharmon@ucsc.edu

## Abstract

Similes are easily obtained from web-driven and case-based reasoning approaches. Still, generating thoughtful figurative descriptions with meaningful relation to narrative context and author style has not yet been fully explored. In this paper, the author prepares the foundation for a computational model which can achieve this level of aesthetic complexity. This paper also introduces and evaluates a possible architecture for generating and ranking figurative comparisons on par with humans: the FIGURE8 system.

## Introduction

Figurative language is embedded within and intimately connected to our cultures, behaviors, and models of the world. In fact, humans use figurative language so often that we seldom realize it (Lakoff and Johnson 1980); still, its utility for communication is clear. Using metaphors and similes, one can relate the unfamiliar, or the tenor, in terms of the familiar, or vehicle (Richards 1980). In Figure 1, for example, "moon" is the vehicle for "garden", the tenor. Attributes of the moon, such as its brilliance, are used to describe the beauty of the garden. Prior to the comparison, the garden's appearance is unknown (is it beautiful and luminous, or neglected and overgrown?). The simile helps to resolve this ambiguity and provide the reader with a clearer picture of the scene.

Comparison gives us the ability to delicately express irony and sarcasm ("clear as mud"), exaggeration ("that man was as tall as a giraffe"), and emotion ("my heart was a sinking ship"). With such tools, we can explain how we feel, what kinds of people we are, and what experiences we have had. Further, metaphors give color to dry speech and are understood faster than literal equivalents (Gibbs and Nagaoka 1985); this is likely due to their appeal to common previous experiences and memories.

For the purpose of this paper, we will consider two styles of figurative language: conventional (common analogies used in daily language, such as "I see what you mean") and creative (original comparisons that call attention to themselves as figures of speech, such as "Fear is a slinking cat I find / Beneath the lilacs of my mind" (Tunnell 1977)). Each type can provide value, although previous work on computational generation of figurative language has primarily focused on understanding and reconstructing conventional metaphors and similes.

Clichés (e.g., "fast as lightning") are arguably useful when fast, informal communication is required between a computer and a human, and such phrases can be learned via web query (Veale and Hao 2007a). Generating creative comparisons on par with human authors is a much more difficult challenge. A conventional metaphor is considered "good" if many others have used it before, but uniqueness and aesthetic qualities are critical in generating a strong creative metaphor. For instance, several aesthetic properties, such as syllable counts, phonetics, stressed syllable position, rhyme, and alliteration have been identified as "obvious" criteria for making creative poetic lines sound good, despite the fact that these "do not translate well into precise generative rules" (Gervás, Hervás, and Robinson 2007). While creative generators for figurative language exist, few address this concept of what makes for a high-quality metaphor or simile. I will describe a system, FIGURE8, which contains a novel underlying model for what defines creative and high quality figurative comparisons, and evaluates its own output based on these rules.

## Related Work

Modern research in creativity has generally defined a creative system as one that generates novel, context-appropriate output (Rothenberg and Hausman 1976; Sawyer 2012). Within the context of creative natural language generation, a third criterion has been noted: a creative system must generate context-appropriate knowledge outside of its pre-existing knowledge base (Pérez y Pérez and Sharples 2004).

Several computational systems exist which attempt to meet this benchmark. ASPERA, for instance, combines case-based reasoning with intelligent adaptation of examples from corpora (Gervás 2000). Psychological theories have further informed the art of generating figurative language, resulting in more advanced and thoughtful systems. Notably, Brown (Ortony 1993) and Glucksberg (Glucksberg 2001) have argued that categorization is inherent to metaphor. As a consequence, the concept of property-based concept mapping has inspired metaphor generation approaches, and has been cited as the best method for producing robust, scalable and useful metaphors (Hervás et al. 2007; Veale and Hao 2007a).

One must also consider how to develop an appropriate knowledge base without substantial manual authoring. Previous exemplary work in metaphor generation has emphasized the power of using the web to establish example cases of valid comparisons (Veale and Hao 2007a; 2007b). However, these systems merely generate large amounts of potentially creative descriptions, and cannot distinguish between original and poor quality comparisons (Veale and Hao 2007a). Further, they often ignore context, sentence construction, and aesthetics in the generation process, resulting in less evocative and meaningful language.

FIGURE8 is a system that uses a web-driven approach to form a preliminary knowledge base of nouns and their properties. The system is provided with a model of the current world and an entity in the world to be described. A suitable vehicle is selected from the knowledge base, and the comparison between the two nouns is clarified by obtaining an understanding via corpora search of what these nouns can do and how they can be described. Sentence completion occurs by intelligent adaptation of a case library of valid grammar constructions. Finally, the comparison is ranked by the system based on semantic, prosodic, and knowledge-based qualities. In this way, FIGURE8 simulates the human-authoring process of revision by generating many vehicle choices and linguistic variations for a single tenor, and choosing the best among them as its favorite. While FIGURE8 does not claim to have a comprehensive set of rules - for example, it does not consider phonetics in its evaluation of description quality - it provides a novel foundation for an intelligent figurative language generation and assessment system.

## Approach

Prior work has established that a strong creative metaphor is not only comprehensible (Tourangeau 1981), novel (Camac and Glucksberg 1984), and context-appropriate (Harwood and Verbrugge 1977; Tversky 1977; Gildea and Glucksberg 1983), but surprising (Tourangeau 1981). The following sections will illustrate how FIGURE8 considers these properties when generating metaphors and similes. A block diagram of the generation process is shown in Figure 3.

### Clarity

A strong metaphor must have an understandable, accurate link between tenor and vehicle. A vehicle is thus only considered acceptable if it has properties in common with the tenor. Further, associating the tenor with the capacities and known manifestations of the vehicle should enhance the clarity of the description. In the FIGURE8 system, these associations are found by mining existing literary corpora (Hart 2014) for instances of the vehicle and using NLTK's parts-of-speech tagging to identify associations (e.g., refer to Figure 2). This procedure enables the system to use words commonly associated with the vehicle to develop a fresh relation to the tenor. For example, if we were to compare a teacher to a horse, FIGURE8 may now be able to reason that the teacher would prance or trot into the room. In this way, a sentence can be generated by only implicitly referring to the vehicle ("The teacher pranced into the room" vs. "The teacher was a wild horse, prancing into the room"). Common verbs, such as forms of "to be", were culled from the generated list of association because - as all nouns have the capability to exist and be - such verbs do not lend clarity to the comparison.

Granted, the word chosen to relate to the tenor may not make sense (especially in the case of verbs), destroying the very clarity it was meant to enhance. FIGURE8 thus performs a web query using Python's urllib module to ensure that others have associated the chosen word with the tenor before. If a previous association has not been made, the metaphor is ranked lower in terms of estimated clarity. This evaluation measure ensures that nonsensical descriptions, such as "The turtle darkened like a blue ocean", are given a lower ranking overall.

### Novelty

Clichés are frowned upon by expert authors; as Salvador Dalí once said, "The first man to compare the cheeks of a young woman to a rose was obviously a poet; the first to repeat it was possibly an idiot" (1968). For computer-generated text, it is thus reasonable to expect that a quality metaphor is a fresh comparison. In the FIGURE8 system, each metaphor is checked against an existing knowledge base of comparisons (Friedman 1996), and all generations are ranked based on their similarity to conventional metaphors in this database.

### Aptness

Ideally, a strong metaphor will fit the context within which it lives. For usage in a narrative context, the FIGURE8 system can be passed a model of a simple world of objects and character models, and incorporate these appropriately into its eventual output along with a prepositional phrase generation module. Additionally, one may ask FIGURE8 to generate ironic comparisons, such as those generated by a sarcastic character when speaking. Irony is achieved by selecting for properties with the exact opposite meanings, in accordance with prior work (Veale and Hao 2007b). The FIGURE8 system also endeavors to match a given context during sentence completion, which will be described in a later section.

### Unpredictability

Metaphors are perceived as cleverer when the vehicle and tenor contain similarities, but the respective domains of these terms are distinct (Tourangeau 1981). A description is thus ranked as more surprising when the words are not very conceptually similar and contain fewer properties in common. With the assumption that they share at least one property in common, the chosen metaphor components are ranked by querying the UMBC Semantic Similarity System (Han et al. 2013). The degree to which the vehicle and tenor share major categories is also considered by using a function similar to WordNet's lexname query. This check is needed because if one or more major categories are shared, the metaphor is considerably less surprising. For

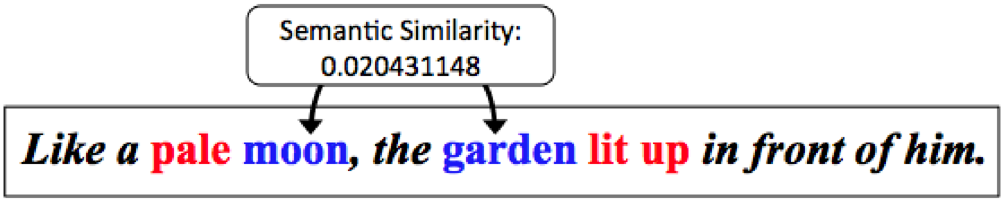


Figure 1: Example of a highly ranked output sentence by FIGURE8. Here, the tenor, vehicle, and associated phrases are *garden*, *moon*, *lit up*, and *pale*. The nouns *garden* and *moon* not only have low semantic similarity, but do not share a major category together. Likening a garden to a moon is also not a cliché comparison, lending to the description's potential novelty.

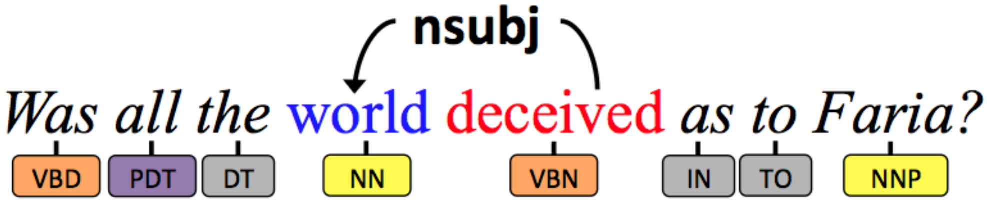


Figure 2: Example of how FIGURE8 discovers and associates a verb with a chosen vehicle, using text from *The Count of Monte Cristo*, and a part-of-speech parsing module similar to the Stanford Parser (Socher et al. 2013). Here, the *nsubj* label refers to a link between a verb ("deceived") and a noun phrase (in this case, the vehicle "world"). The remaining labels in the figure represent the part-of-speech tags.

instance, "the strawberry is a pomegranate" is considered a poor metaphor because strawberry and pomegranate are contained within a major category: fruits. Such a description may be produced by a web-based generator (for instance, the online MIT-licensed Metaphorgy system (Groff-Palermo and Lawson 2013) produces "My strawberry is a Phaeacian cherry"), but will be given a low ranking by FIGURE8.

### Prosody

The prosody of a metaphor can be defined as the rhythmic, tonal, and aesthetic qualities that distinguish one metaphor from another. Descriptions are ranked highly if their prosody is of consistent and high quality. For instance, consider the following similes:

(1) The serpent stretched into the horizon, like a deserted desert.
(2) The snake extended into the horizon, like an abandoned desert.

Although alliteration and assonance can be used beautifully in figurative language, the high similarity of consecutive words in (1) may be distracting. Example (2) depicts the same imagery, but uses words of greater distance in terms of consecutive string similarity.

At present, FIGURE8 conducts string similarity via Python's difflib to evaluate the prosody of its outputs. Using difflib's SequenceMatcher, one can determine a value indicating the degree of similarity between two input strings in a range from 0 (no similarity) to 1 (identical strings). FIGURE8 is thus able to quantify the string similarity for consecutive words, and ranks descriptions lower if there are many consecutive string similarity values above 0.7, which was deemed an appropriate threshold by the author. Consecutive words are also checked for alliteration and assonance, which are considered positive qualities by FIGURE8.

### Sentence Completion

Automated metaphor identification in text has been thoroughly explored (Neuman et al. 2013; Steen et al. 2010) and, as such, FIGURE8 has been provided with a case library of appropriate sentence constructions for metaphor and simile. By following the procedure of imaginative recall (Turner 1992), FIGURE8 first attempts to fit the provided context of the situation to an exact, pre-existing solution. If no solution exists, FIGURE8 searches its memory, solves the problem for a similar case, and adapts that solution to the provided context. As an illustration: if FIGURE8 notes that other authors have used the phrase "to the barn", it should recognize the barn as a noun denoting a man-made object via WordNet. Similarly, a "chair" is a man-made object, and thus, FIGURE8 may decide to replace "barn" with "chair" when told that a chair exists in the current narrative context. This adaptive process enables FIGURE8 to match its constructions to any provided context and complete statements creatively.

## Evaluation

Little research, if any, has worked towards developing a model of what makes a high quality computer-generated metaphor. Although there is no standard method to evaluate computationally-generated figurative descriptions, one reasonable way to judge would seem to be agreement with hu-

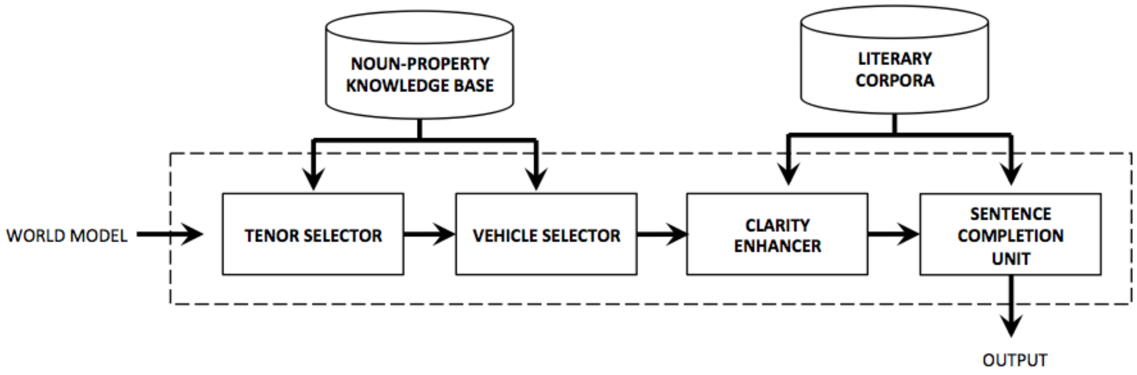


Figure 3: Block diagram of the FIGURE8 generation system. If no world model is given, a tenor is selected at random from the noun-property database. A vehicle is then selected with at least one property in common with the tenor. The Clarity Enhancer module requests verbs and adjectives associated with the vehicle from mined literary corpora. Finally, the sentence is completed by performing imaginative recall with known valid sentence constructions for metaphor identified from literary corpora.

| **Generated Description** | **FIGURE8 Clarity** | **Human Clarity** | **FIGURE8 Likability** | **Human Likability** |
|---|---|---|---|---|
| (A) *It was the pearl, fermenting like a wild apple.* | 3 | 2 | 3 | 2 |
| (B) *Like scenic music, the pearl danced in front of him.* | 1 | 1 | 1 | 1 |
| (C) *It was their pearl, sprawling like a wretched corpse.* | 4 | 4/5 | 4 | 3/4 |
| (D) *It was her pearl, crumpling like a drowned corpse.* | 5 | 4/5 | 5 | 5 |
| (E) *It was my pearl, bubbling like a treacherous swamp.* | 2 | 3 | 2 | 3/4 |

Table 1: Comparison 1 of FIGURE8 and human rankings for clarity and overall quality. In this set, FIGURE8 was asked to generate and rank figurative descriptions given "pearl" as the tenor. Human clarity and likability rankings were found to be highly correlated ($\rho = 0.684$). Spearman analyses also indicated positive correlations between human and FIGURE8 rankings (clarity: $\rho = 0.872$; quality: $\rho = 0.821$).

man ratings. This can be assessed by requesting humans to rank descriptions generated by the FIGURE8 algorithm, and determining if the majority are in agreement with the computer's (FIGURE8's) ranking. A pilot study indicated that providing each description with additional context would make the ranking process too time-consuming for participants. Thus, functions to enhance aptness were not included when generating outputs to be evaluated in the full-scale study.

## Method

One hundred participants (73 female, 27 male) were recruited via Amazon's Mechanical Turk. Each participant viewed a series of five sentences at a time, and were asked to rank the similes by how understandable they were (*clarity*), and by how much they, as individuals, enjoyed the comparison (*likability*). Each set of five sentences contained the same tenor, and were originally generated and ranked by FIGURE8. The sets were not hand-selected by the author. That is, the first eleven sets FIGURE8 generated and ranked were used in the study.

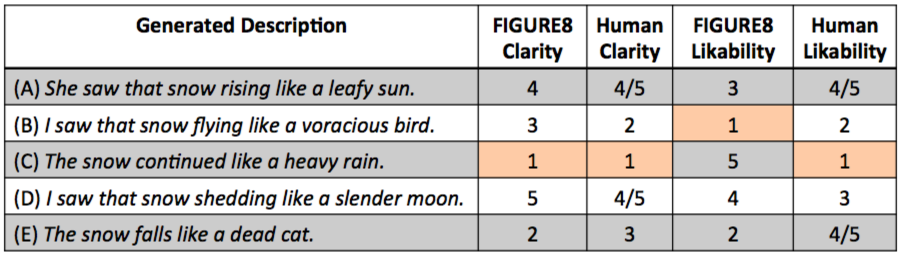
## Results

Human preferences were determined by following the majority criterion. As seen in Figures 4 and 5, human clarity ratings were often positively correlated with overall quality ratings, and this correlation was confirmed with Spearman analyses. Overall, FIGURE8's top result for clarity and overall quality generally agreed with the human rankings for each of the eleven sets. FIGURE8 exactly matched the first ranking 46% of the time for clarity and likability. Further, it matched either the first or second ranking 82% and 100% of the time for the clarity and likability categories, respectively. Examples of how FIGURE8 matched human ratings are shown in Tables 1, 2, and 3.

# Discussion and Future Work

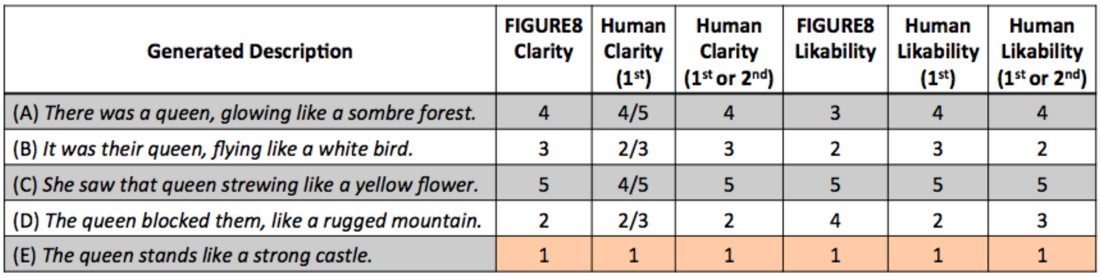
In this paper, the author has introduced the FIGURE8 system as a novel tool for generating and evaluating creative figurative descriptions. FIGURE8's assessments are grounded in psychological models of metaphor comprehension, and have thus far been found to adequately match human rankings when agreed upon.

Participants in the evaluation portion were not told that the descriptions were generated by a computer. Only two



| Generated Description | FIGURE8 Clarity | Human Clarity | FIGURE8 Likability | Human Likability |
|---|---|---|---|---|
| (A) *She saw that snow rising like a leafy sun.* | 4 | 4/5 | 3 | 4/5 |
| (B) *I saw that snow flying like a voracious bird.* | 3 | 2 | 1 | 2 |
| (C) *The snow continued like a heavy rain.* | 1 | 1 | 5 | 1 |
| (D) *I saw that snow shedding like a slender moon.* | 5 | 4/5 | 4 | 3 |
| (E) *The snow falls like a dead cat.* | 2 | 3 | 2 | 4/5 |

Table 2: Comparison 2 of FIGURE8 and human rankings for sentences of tenor "snow". Human clarity and likability rankings were found to be positively correlated ($\rho$ = 0.763). Spearman correlation analysis suggested that FIGURE8 clarity rankings were positively associated with human clarity rankings ($\rho$ = 0.872), but no significant association was found between likability rankings in this case ($\rho$ = -0.359).



| Generated Description | FIGURE8 Clarity | Human Clarity (1st) | Human Clarity (1st or 2nd) | FIGURE8 Likability | Human Likability (1st) | Human Likability (1st or 2nd) |
|---|---|---|---|---|---|---|
| (A) *There was a queen, glowing like a sombre forest.* | 4 | 4/5 | 4 | 3 | 4 | 4 |
| (B) *It was their queen, flying like a white bird.* | 3 | 2/3 | 3 | 2 | 3 | 2 |
| (C) *She saw that queen strewing like a yellow flower.* | 5 | 4/5 | 5 | 5 | 5 | 5 |
| (D) *The queen blocked them, like a rugged mountain.* | 2 | 2/3 | 2 | 4 | 2 | 3 |
| (E) *The queen stands like a strong castle.* | 1 | 1 | 1 | 1 | 1 | 1 |

Table 3: A third comparison of FIGURE8 and human rankings for sentences of tenor "queen". In addition to showing first choice rankings, this table displays human rankings when considering first and second choices. That is, "the queen stands like a strong castle" was ranked as either first or second for the majority of respondents. In both cases, human clarity and likability rankings were found to be positively correlated ($\rho > 0.9$). Spearman analyses also suggested for both cases that FIGURE8 and human rankings for clarity and likability were positively correlated with high significance ($\rho > 0.7$).

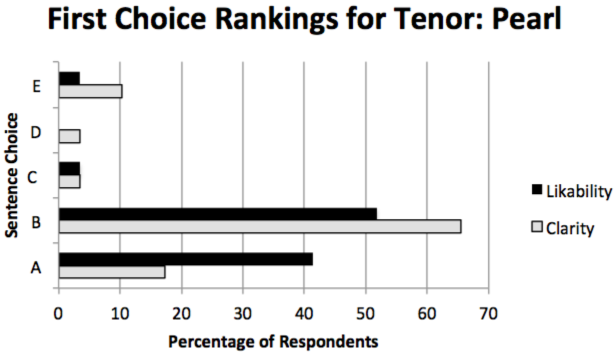


Figure 4: First choice rankings for the generated set of sentences using *pearl* as the tenor. Although some disparities existed, the majority of respondents generally agreed upon which sentence was the most understandable.

comments were made about checking sentences for validity prior to including them in the study, and one regarding how painful it was to rank "bad poetry". Most participants, however, enjoyed the task and provided positive feedback about their experience ("cool hit","super fun","I love this"). It is conceivable that task enjoyment affected user responses, but controlling for explicit indication of task enjoyment yielded no significant difference in the results. Controlling for gender also did not reveal significantly different outcomes.

Interestingly, for roughly half (50-60% per set) of the participants, how much they liked the figurative description was directly correlated with how well they understood it. The most highly ranked phrases for clarity were also often ranked first for likability, and the Spearman coefficient was used to confirm these positive associations. This was a surprising finding, because more variation and subjectivity was expected for these ratings. Discrepancies between human and FIGURE8 likability rankings, such as in Table 2, could potentially be explained by a human tendency to prefer metaphors containing words of positive sentiment value. However, more analysis is required to confirm this idea, and further study is needed to evaluate how qualities of language are weighted across general and expert populations. Judging from participant comments, it is also possible that some people may like metaphors primarily based on qualities other than clarity (such as prosody, sentiment, or whimsy). If these groups could be automatically identified, perhaps future computer-produced descriptions could adapt to generate more personalized descriptions for the optimum enjoyment of the reader.

While FIGURE8 is able to rank its figurative descriptions over various measures of quality, how well its output com-

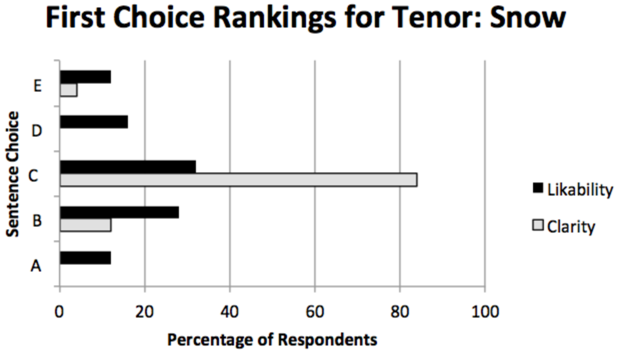


Figure 5: Clarity rankings for the generated set of sentences using *snow* as the tenor. Participants rated what FIGURE8 considered the most unsurprising metaphor as the most clear, but there was no highly significant consensus regarding the most likable description.

pares with human-authored descriptions was not assessed. The fact that most participants in the evaluation did not question the source of the texts is a promising sign that the system presented here generates human-like output. Regardless, its present constructions can be automatically assigned rankings on par with human evaluations. It is assumed that as the quality of FIGURE8's generations increases, it will be able to extract the best output from the results of its "brainstorming". Future research should build upon this foundation and work towards evaluating computer-generated descriptions in terms of aptness, prosody, and unpredictability. When machines are fully able to grasp the subtleties and aesthetics of figurative language, we as humans will be able to relate to them as never before.

## References

Camac, M. K., and Glucksberg, S. 1984. Metaphors do not use associations between concepts, they are used to create them. *Journal of Psycholinguistic Research* 13(6):443–455.

Friedman, S. M. 1996. Cliche finder. Retrieved 2 Mar 2015 from http://www.westegg.com/cliche/.

Gervás, P.; Hervás, R.; and Robinson, J. R. 2007. Difficulties and challenges in automatic poem generation: Five years of research at UCM. *e-poetry 2007*.

Gervás, P. 2000. An expert system for the composition of formal Spanish poetry. *Journal of Knowledge-based Systems* 14:200–201.

Gibbs, R. W., and Nagaoka, A. 1985. Getting the hang of American slang: Studies on understanding and remembering slang metaphors. *Language and Speech* 28(2):177–194.

Gildea, P., and Glucksberg, S. 1983. On understanding metaphor: The role of context. *Journal of Verbal Learning and Verbal Behavior* 22:577–590.

Glucksberg, S., ed. 2001. *Understanding figurative language: From metaphors to idioms*. Oxford: Oxford: University Press.

Groff-Palermo, S., and Lawson, J. 2013. Metaphorgy: Metaphor generator. Retrieved 21 Dec 2014 from http://www.metaphor.gy/.

Han, L.; Kashyap, A. L.; Finin, T.; Mayfield, J.; and Weese, J. 2013. UMBC_EBIQUITY-CORE: Semantic textual similarity systems. In *Proc. 2nd Joint Conf. on Lexical and Computational Semantics, Association for Computational Linguistics*.

Hart, M. 2014. Free ebooks - Project Gutenberg. Gutenberg.org. http://www.gutenberg.org/.

Harwood, D. L., and Verbrugge, R. R. 1977. Metaphor and the asymmetry of similarity. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Hervás, R.; Costa, R. P.; Costa, H.; Gervás, P.; and Pereira, F. C. 2007. Enrichment of automatically generated texts using metaphor. *MICAI 2007, LNAI 4827* 944–954.

Lakoff, G., and Johnson, M., eds. 1980. *Metaphors We Live By*. Chicago, IL: University Of Chicago Press.

Neuman, Y.; Assaf, D.; Cohen, Y.; Last, M.; Argamon, S.; Howard, N.; and Frieder, O. 2013. Metaphor identification in large texts corpora. *PLoS ONE* 8(4): e62343:doi:10.1371/journal.pone.0062343.

Ortony, A., ed. 1993. *Metaphor and Thought*. Cambridge University Press.

Pérez y Pérez, R., and Sharples, M. 2004. Three computer-based models of storytelling: BRUTUS, MINSTREL and MEXICA. *Knowledge-Based Systems* 17(1):15–29.

Richards, I. A., ed. 1980. *The Philosophy of Rhetoric*. Oxford: Oxford University Press.

Rothenberg, A., and Hausman, C. R., eds. 1976. *The Creative Question*. Durham NC: Duke University Press.

Sawyer, R. K. 2012. *Explaining Creativity: The Science of Human Innovation*. Oxford University Press.

Socher, R.; Bauer, J.; Manning, C. D.; and Ng., A. Y. 2013. Parsing with compositional vector grammars. In *Proceedings of ACL 2013*.

Steen, G. J.; Dorst, A. G.; Herrmann, J. B.; Kaal, A.; Krennmayr, T.; and Pasma, T. 2010. *A method for linguistic metaphor identification*. From MIP to MIPVU. Amsterdam: John Benjamins.

Tourangeau, R. 1981. Aptness in metaphor. *Cognitive Psychology* 13(1):27–55.

Tunnell, S. 1977. *The Quotable Women, 1800-1975*. Corwin Books.

Turner, S. 1992. Minstrel: a computer model of creativity and storytelling. Technical Report CSD-920057, Ph.D. Thesis, Computer Science Department, University of California, Los Angeles, CA.

Tversky, A. 1977. Features of similarity. *Psychological Review* 84:327–352.

Veale, T., and Hao, Y. 2007a. Comprehending and generating apt metaphors: A web-driven, case-based approach to figurative language. In *Proceedings of the Twenty-*

*Second AAAI Conference on Artificial Intelligence (AAAI-07)*, 1471–1476. Vancouver, British Columbia: AAAI Press.

Veale, T., and Hao, Y. 2007b. Learning to understand figurative language: From similes to metaphors to irony. In *Proceedings of Cog Sci*, 683–688.