Metaphoria: An Algorithmic Companion for Metaphor Creation

Katy Ilonka Gero
Columbia University
katy@cs.columbia.edu

Lydia B. Chilton
Columbia University
chilton@cs.columbia.edu

ABSTRACT

Creative writing, from poetry to journalism, is at the crux of human ingenuity and social interaction. Existing creative writing support tools produce entire passages or fully formed sentences, but these approaches fail to adapt to the writer's own ideas and intentions. Instead we posit to build tools that generate ideas coherent with the writer's context and encourage writers to produce divergent outcomes. To explore this, we focus on supporting metaphor creation. We present Metaphoria, an interactive system that generates metaphorical connections based on an input word from the writer. Our studies show that Metaphoria provides more coherent suggestions than existing systems, and supports the expression of writers' unique intentions. We discuss the complex issue of ownership in human-machine collaboration and how to build adaptive creativity support tools in other domains.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; *Natural language interfaces*; • **Applied computing** → *Arts and humanities*;

KEYWORDS

human-computer collaboration; co-creativity; generative art; writing support; natural language processing

ACM Reference Format:
Katy Ilonka Gero and Lydia B. Chilton. 2019. Metaphoria: An Algorithmic Companion for Metaphor Creation. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3290605.3300526

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300526

Figure 1: A poet using Metaphoria to find metaphorical connections between *america* and *wood*.

1 INTRODUCTION

Creative writing, from poetry to journalism, is at the crux of human ingenuity and social interaction. It conveys not only information but also experience, emotion, and beauty. While computation has opened a floodgate of creative tools for music and the visual arts, little of that fervor has transferred to text. Word processors that detect grammatical errors are useful, but do not support the creative elements of writing.

Past work in computational support for creative writing has focused on suggesting next sentences while writing stories [3, 26, 36] or fully generating a creative output based on a topic [11]. These ideas have potential, but current systems fail to provide strong coherence with the intention of the writer—either the text that they have already written or their intention for the entire output. Since these tools are not user-centric, they are most useful during ideation when there are fewer constraints. In this case, a system's failure to provide coherence can be seen as a feature: a random suggestion can help a writer move in an unexpected direction.

We can improve tools for creative writing by designing them from a user-centric perspective. To do so, we propose focusing on the building blocks of creative writing, in which writers have more specific goals. Instead of providing complete sentences generally applicable to wherever the writer is, we can improve the relevance of our support by constraining the idea space to a specific writing goal, and allow it to be used at more points in the writing process. We focus on metaphor, which famously conveys complex or abstract ideas succinctly and is used in everything from poetry to journalism to science education [19, 28, 29].

Creating unconventional and expressive metaphors is challenging [10], requiring divergent and lateral cognitive processes [13]. We present Metaphoria: an interactive system that generates potential metaphorical connections for any input word. Metaphoria uses an open source knowledge graph and a modified Word Mover's Distance algorithm to find a large, ranked list of suggested metaphorical connections. These suggestions are embedded in an interactive interface that allows writers to generate ideas for any input. Figure 1 shows the system while used by a professional poet.

We ran three studies to evaluate Metaphoria. First, we compare our method for generating suggestions to state-of-the-art systems and show it performs better across three metrics for metaphor quality. Second, we have novices write extended metaphors with and without Metaphoria and show that Metaphoria generates meaningful and inspirational suggestions given a specific writing task. Third, we have professional poets write poems with Metaphoria and show the range of expression using the system. In the Discussion, we report on issues of ownership that arise when a computational system produces "human-like" output, and suggest future work to mitigate these concerns.

We make the following contributions:

- A computational method for producing metaphorical connections better than state-of-the-art algorithms.
- Metaphoria, an interactive system for collaboratively writing metaphors with a computer.
- User studies with novice and expert writers, showing that Metaphoria gives people useful and inspirational suggestions and increases the diversity of responses.
- Design implications for ownership in co-creative systems more generally.

2 RELATED WORK

Writing support

Writing support has a long history; editing has existed perhaps as long as writing and the introduction of dictionaries and thesauri gave writers external tools they could use on their own. Experimental writing movements, such as the Dadaists with their cut-up technique and the Oulipo with their constrained methods, employed algorithmic ideas to trigger inspiration, pre-dating the advent of computers.

One of the early successes of computation was the development of spell-check [33], and grammar-checking remains an active area of research today [20]. Recent computational work has leveraged cognitive apprenticeship models to improve writing with highly specific goals, such as an email to request help [15], an essay for a standardized test [2], or a piece of journalism [25]. Work on collaborative writing [1, 17, 39] has shown that writing can be broken into micro-tasks in which individuals can contribute usefully without access to the full writing document.

This success suggests applying user-centric ideas to creative writing. Support for creative writing has focused on generating next sentences for a story [3, 36, 38] or generating entire poems given a topic [11, 31]. While this paradigm has potential to trigger inspiration similar to the earlier, experimental movements, we focus on providing more coherent suggestions by responding to the need for rhetorical devices. We provide support for metaphor creation, a common but challenging rhetorical device [10]. This narrowing of the goal, similar to previous HCI work on writing, allows us to achieve the coherence necessary to move beyond random association and support the creation of meaning.

Creativity support and co-creativity

Creativity support tools have flourished for music and the visual arts, from the widespread adoption of software for generation and editing to the development of medium-specific programming languages [22, 34, 45]. These tools are beginning to tackle how to be compatible with existing manual practices [16], as well as how to be more compatible with current artificial intelligence frameworks [6, 30].

The way in which creativity support tools integrate with an artist's practice is at the heart of these issues. When a support tool provides more complete or conceptual contributions, or provides contributions without a request from the artist (as in mixed-initiative user interfaces [14]), the term co-creativity is often used. Critically, Davis defines human-computer co-creativity as when the "program is adapting to the input of the user" [5]. This distinguishes co-creative systems from more procedural contributions, in which an artist either has a high level of control over the outputs, as in a synthesizer, or little to no control over the outputs, as in a computer-generated poem based on a topic [11].

It is essential to think about tools as supporting artists in their desired practice, rather than replacing aspects deemed computationally tractable. Support for creative writing should align with the 'wide walls' design principle of creativity support tools, in which tools aim to "support and suggest a wide range of explorations" [35]. Unlike more specified writing tasks (such as writing an email to request help), creative writers do not want tools that will make their writing sound the same as others [38]. Thus, in co-creative domains, systems should be conducive to divergent outcomes.

Metaphor generation algorithms

Metaphor generation is a version of conceptual blending [7] that has been correlated with general fluid intelligence [37] and is considered an important challenge in artificial intelligence [44].

Current metaphor generation systems find properties that can be attributed to the two concepts in the metaphor. Two prominent algorithms are Thesaurus Rex [40, 42] and Intersecting Word Vectors [8]. Thesaurus Rex [40, 42] is a web service that provides shared attributes and categories for input concepts. For example, inputting *coffee & cola* produces results such as *acidic food* and *nonalcoholic beverage*. Thesaurus Rex is explicitly designed to support metaphor generation [41, 43]. Intersecting Word Vectors [8] is a metaphor generation algorithm in which connector words are found using word embeddings. Connector words are those found in the intersection of the 1000 words closest to each of the concept words. For example, connector words for *storm & surrender* include *barrage* and *onslaught*. These systems are strong baselines for metaphor generation from the artificial intelligence and natural language processing communities.

Theories of metaphor often conform to structural alignment theory [9] in which analogies are discovered by finding isomorphic sections of knowledge graphs, where each edge is a structural relation between concepts. Work on using analogies for product design [12] has focused on the difference between structural and functional aspects of products for ideation. We draw on these ideas of structural and functional connections as a search function for concept attributes.

3 DESIGN OF METAPHORIA

Design Goals

Based on our literature review, **coherence to context** is the biggest barrier to use for creative writing support tools [3, 26, 36]. Secondarily, writers do not want tools that make their writing sound the same as others [38]. Thus, suggestions that result in **divergent outcomes** for writers is crucial. These goals map to previous methodology in HCI for the evaluation of generative drawing tools; Jacobs et. al. [16] evaluate their drawing tool on compatibility (coherence to context) and expressiveness (ability to express a divergent set of ideas) .

A system that is **coherent to context** provides suggestions that are relevant to the task at hand. If writers come to the system with an idea or intention, the system should generate metaphorical phrases coherent with this context, and should be flexible enough to be coherent for a wide range

high	envy is used for getting attention like a bell
	envy is for alerting you to something like a bell
	...
	envy is used to toll like bell
low	envy is for playing music like a bell

Table 1: Examples of connections with high and low relevance for the seed *envy is a bell.*

of writer ideas and intentions. A system that encourages **divergent outcomes** provides many compelling options and increases the variation in writers' work rather than propel all writers toward similar metaphors.

To address **coherence to context**, we focus on generating metaphorical connections for a given "seed metaphor". Seed metaphors are of the form *[source] is [vehicle]*, e.g. *envy is a bell*, where *envy* is the source and *bell* is the vehicle. By focusing on connections between the words, such as 'envy can sound the alarm like a bell', rather than the selection of the seed words, we leave open the possibility that the writer inputs one or both words of the seed metaphor.

To address **divergent outcomes**, we generate and present multiple, distinct suggestions for each seed metaphor. This approach allows writers to select a suggestion salient for them in particular.

Generating coherent connections

Starting with a seed metaphor, our approach is to first generate many features of the vehicle (*bell*), and then rank these features by how related they are to the source (*envy*). This aligns with traditional metaphor usage, in which features of the vehicle are used to explain the source.

To find features of the vehicle we use ConceptNet [24], an open-source knowledge graph, as a source of structural and functional properties of words. Structural properties are elements that define or compose an object. For example, a *bell* has a *clapper* and a *mouth*. In ConceptNet, we select for structural features by querying the "HasA" relations of the vehicle. Functional properties focus on an object's actions and purpose. For example, a *bell* can *make noise* and be used for *alerting*. In ConceptNet, we select for functional features by querying the "UsedFor" and "CapableOf" relations. Together, structural and functional properties provide a large set of potential connections from the vehicle to the source.

Not all features of the vehicle (*bell*) will metaphorically map to the source (*envy*). To find the most relevant ones, we rank how related the vehicle features (e.g. *used for getting attention*) are to the source (*envy*). To rank suggestions we use GloVe word embeddings [32] trained on Wikipedia 2014 + Gigaword 5. Word embeddings are a common way to measure the semantic similarity between words [27]. Here, we use them to measure the semantic similarity between the vehicle property and source word. Examples of vehicle properties with high and low relevance are found in Table 1.

To find the semantic distance between vehicle features and the source word, we use a modified Word Mover's Distance (WMD) [18]. WMD is an algorithm for finding the smallest distance between two documents, i.e. sets of words, in a word embedding space. It formulates distance between documents as a transportation problem: we denote $c(i, j)$ as the distance between words x_i and x_j, where $c(i, j)$ is the cosine distance between the two word vectors. Given two documents D_1 and D_2, we allow each word i in D_1 to be transformed into any word in D_2 in total or in parts. We denote T_{ij} as how much of word i in D_1 is transformed to word j in D_2; therefore $\sum_{i,j} T_{ij} = 1$.

We can define the distance between two documents as the minimum cumulative cost of moving all words in D_1 to all words in D_2. This is equivalent to solving the linear program

$$\min \sum_{i,j} T_{ij} * c(i, j) \tag{1}$$

for which specialized solvers have been developed. For example, this would find the shortest distance from *making noise* to *envy*.[1] From this ranking of connections, we can select the top n as the most coherent.

Selecting multiple distinct connections

In order to promote diverse outcomes, our systems presents writers with 10 coherent suggestions that are semantically distinct. For instance *get attention* and *getting people's attention* may both be coherent, yet they give effectively the same idea to the writer. For this reason, as we build our list of suggestions to show the writer, we throw out any feature that is too close to any of the features already ranked. This closeness is again calculated with the Word Mover's Distance, this time between two features. Through observation, we find a distance of less than 4 indicates two features are not semantically distinct.

Additional coherence with valence ranking

The word embedding space is not sensitive to antonyms and thus some highly ranked features have a mismatched sentiment with the source concept. Pilot testing showed that people found mismatched sentiments to be jarring and caused them to lose faith in the system. However, people who are first shown more intuitive features were more likely to appreciate the antonym features. Thus, we first select the suggestions as shown above, and then re-rank them by how similar the valence of each one is to the source concept.

[1] In this usage, D_2 is always a single word, the source concept, although our implementation allows for natural expansion into multi-word sources.

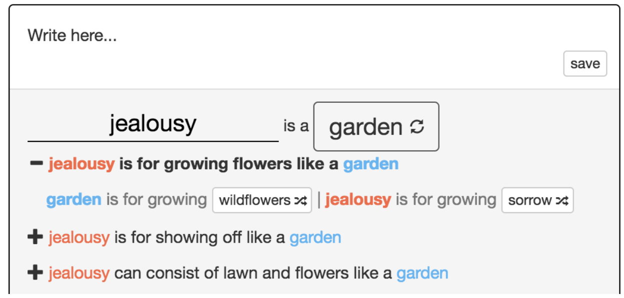

Figure 2: Screenshot of Metaphoria with suggestion for *jealousy is a garden* expanded.

Valence is the positive or negative connotation of a word and we assign valence scores to all words based on Warriner, et. al's database [46]. We denote the valence of the source as V_{source} and the valence of word i in the feature V_i for words $1, ..., n$. Then we define the valence distance as

$$V_{dist} = |V_{source} - \text{avg}(V_1, ..., V_n)| \tag{2}$$

We can then reorder the suggestions from the smallest valence distance to the largest.

Finally, we rephrase all connections into a suggestion for the writer; given the source *envy*, the vehicle *bell* and the connecting feature *making noise*, the suggestion is presented as 'envy is used for making noise like a bell'.

Additional distinctness with suggestion expansion

Great metaphors are specific; we want to support writing specific metaphors by expanding them to include more details of how the source and vehicle are connected. If envy *makes noise* like a bell, we can expand on the details of the noise a bell makes (e.g. *vibrato, reverberation, high/low pitch*) and how these details relate to envy. For example, the noise of a bell has *reverberation*; and envy has *lasting bitterness*. Metaphoria provides multiple detailed metaphoric expansions for each suggestion to give writers more diverse options.

To generate the expanded metaphors, we first split each suggestion into two parallel sentences: one about the vehicle (*bells make noise*) and one about the source (*envy makes noise*). We want to find several alternative words to replace *noise* in each sentence. To generate these words, we again rely on word embeddings. This time, however, we want to discover words that will syntactically match the sentence–for this reason, we use word embeddings trained using a dependency parse as the context [21]. This results in similar words also having a similar part of speech. We use the word embeddings to create list of 60 words similar to the content word (*noise*) and 60 words similar to source (*envy*). Then, we order these words by similarity to the vehicle (*bell*) and original word (*noise*), respectively, and return the 10 most related words in each case. Figure 2 shows the interface where a writer selects the suggestion "jealousy is for growing flowers like a garden" and can click through suggested expansions such as "jealousy is for growing sorrow."

Interactivity

The above methods are embedded in a Flask-based web application, as shown in Figure 1. Writers can input their own source and click through a set of common vehicles. Each combination will generate a list of up to 10 suggestions, and each suggestion can be expanded.

The design of Metaphoria has our goals of **coherence to context** and **divergent outcomes** in mind. By allowing writers to input a source and change the vehicle, we adapt to the intention of the writer, allowing greater coherence. Showing writers 10 semantically relevant suggestions, and enabling writers to 'shift' the suggestions with the detail words, enables a diversity of ideas and, hopefully, responses.

4 STUDY 1: SUGGESTION QUALITY

This study evaluates the quality of the suggestions Metaphoria generates. To achieve **coherence to context**, suggestions should make sense given their seed metaphor and enact principles of high quality writing.

Methodology

To evaluate the suggestions, we compare them to two other state-of-the-art metaphor generation algorithms: Thesaurus Rex [42] and Intersecting Word Vectors [8]. These algorithms are described fully in the Related Works section. As our system produces a ranked set of suggestions, we also compare both the highest ranked suggestions with the lowest to evaluate the effectiveness of the ranking algorithm.

Thesaurus Rex produces shared attributes; for example *envy & bell* produces attributes such as *loud*. Intersecting similarly produces connector words; for *envy & bell* it produces words such as *behold*. In both cases we formulate these into sentences comparable with Metaphoria suggestions. Table 2 shows examples of this.

Metaphoria
envy is used for getting attention like a bell
envy is for alerting you to something like a bell
Thesaurus Rex
envy is loud like a bell
envy is audible like a bell
Intersecting
envy is shiny like a bell
envy can behold like a bell

Table 2: Examples of metaphors from Metaphoria and two comparable, state-of-the-art metaphor generation algorithms for the seed *envy is a bell.*

For each system we select the top three ranked suggestions. Ranking for Metaphoria is done using the WMD distance to the source concept (as explained in the Design section); both Thesaurus Rex and Intersecting generate ranked lists.

To compare the systems, we define three metrics for evaluating metaphor strength. The first is **aptness**, in which a metaphor accurately describes a connection between the concepts; this is the ground level of metaphors. The second is **specificity**, in which a metaphor describes a connection unlikely to be transferable other concepts. The third is **imageability**, in which a metaphor describes a connection the reader can visualize.

We expect that Intersecting will not be particularly apt as it relies solely on the embedding space to provide meaning and embedding spaces notoriously lack consistent discrete semantics [23]. Thesaurus Rex uses textual evidence, so we expect its connections to be apt, but because of this we also expect it to be less imageable and specific as it may only find higher level, and thus vaguer, attributes.

We have three hypotheses:

- H1: Metaphoria suggestions are **more apt** than Intersecting and **at least as apt** as Thesaurus Rex.
- H2: Metaphoria suggestions are **more specific** than Thesaurus Rex and Intersecting.
- H3: Metaphoria suggestions are **more imageable** than Thesaurus Rex and Intersecting.

Additionally, we want to know if top-ranked Metaphoria suggestions are more apt than bottom-ranked ones. For this, we compare the top three and bottom three ranked suggestions. Our hypothesis is:

- H4: Top-ranked Metaphoria suggestions are **more apt** than bottom ranked ones.

We have two professional writers with an MFA in Creative Writing act as annotators. We consider 12 different seed metaphors, e.g. *hope is a stream*, and for each generate the top 3 metaphor suggestions from each system. Additional we generate the bottom 3 metaphor suggestions for Metaphoria. This results in 144 suggestions total.

The annotators consider each metaphor suggestion and mark whether it is apt, specific, and imageable. They are told that all suggestions are generated by computers, but they are not told anything about how or the fact that they come from different systems. They are shown the suggestions for each seed metaphor in random order.

In addition to definitions of the metrics, annotators were also provided with examples of positive and negative cases for each category, as shown found in Table 3.

	Apt: makes sense
strong example	Love can *come on unexpectedly.*
weak example	Love is a *weather event.*
	Specific: uniquely belonging
strong example	Love can *last through the night.*
weak example	Love is *dark.*
	Imageable: evokes visual
strong example	Love can *rain down on our heads.*
weak example	Love can *scare people.*

Table 3: Examples of strong and weak sentences for each of the metaphor evaluation metrics. All sentences are based on the seed metaphor *love is a storm.*

	Apt	Specific	Imageable
Metaphoria (M)	97%	**82%**	**100%**
Thesaurus Rex (TR)	**100%**	47%	**100%**
Intersecting (I)	49%	43%	53%

Table 4: While both Metaphoria and Thesaurus Rex generate apt and imageable metaphors, only Metaphoria consistently produces specific metaphors.

Hypothesis	diff	t-value	p-value
H1a M *more apt* than I	0.48	5.83	2.8e-08
H1b TR *more apt* than I	0.51	6.16	4.8e-09
H2a M *more specific* than TR	0.34	3.36	2.7e-03
H2b M *more specific* than I	0.38	3.55	6.7e-04
H3a M *more imageable* than TR	0.00	n/a	n/a
H3b M *more imageable* than I	0.47	5.59	1.4e-09

Table 5: T-tests confirm that Metaphoria is as good or better across all metrics than state-of-the-art metaphor generation algorithms. P-values are Bonferonni corrected.

	Apt	Specific	Imageable
Top-ranked	**97%**	82%	**100%**
Bottom-ranked	78%	**85%**	89%

Table 6: Top-ranked metaphors perform significantly better than bottom-ranked metaphors on aptness and imageability; there is no significant difference for specificity.

As in any evaluation of linguistic artifacts, it is not clear that there are precise or correct rankings for all of these attributes. Instead, there are general trends that most native English speakers would agree with. We first have the annotators evaluate suggestions for 2 seed metaphors together and discuss their evaluation in order to establish common understandings of the metrics. They then individually evaluate the suggestions for the 12 seed metaphors.

Results

We report the percent agreement between the two annotators for apt, specific, and imageable (and the Cohen's Kappa correlation coefficients) to be 85% (0.63), 83% (0.67) and 88% (0.64), respectively. Given the natural ambiguity of metaphors and creative writing, this is a high level of agreement.

The following results are determined by combining the evaluations of the two annotators; the higher evaluation is used in cases of disagreement. Table 4 shows the percent of times a given systems' suggestions was marked as apt, specific, or imageable. While Metaphoria and Thesaurus Rex metaphors are both consistently apt and imageable, Metaphoria outperforms all systems on specificity.

To test H1-3, we perform paired t-tests (Bonferonni corrected) on the relevant pairs and disprove the null hypothesis for H1 and H2. However, it is clear that H3 does not hold as both Metaphoria and Thesaurus Rex were 100% imageable. The results of the statistical tests can be found in Table 5.

Surprisingly, Thesaurus Rex metaphors were as imageable as Metaphoria ones. In general the annotators found adjectives like *hard* more imageable than we expected. However, Metaphoria still outperforms other systems on specificity.

We also consider the difference between the top and bottom ranked Metaphoria suggestions; Table 1 shows examples. Table 6 shows the percent of times a given systems' suggestions was marked as apt, specific, or imageable. Top ranked suggestions are more apt than bottom ranked ones (t = 2.49, p-value = 0.01) which confirms H4. There is no significant difference for specificity (t = -0.30, p-value = 0.76). However, top ranked suggestions are slightly more imageable than bottom ranked suggestions (t = 2.09, p-value = 0.04). It could be that aptness makes it easier visualize the suggestion.

This shows that Metaphoria creates high quality metaphors and can provide strong suggestions to writers.

5 STUDY 2: NOVICE USERS

This study evaluates the quality of the suggestions Metaphoria generates in the context of a specific writing task: writing extended metaphors. This allows us to test **coherence to context**, as well as if Metaphoria supports **divergent outcomes** when writers are given the same list of suggestions.

Methodology

We recruited 16 undergraduates: 8 female, 8 male, with an average age of 19.5 ($\sigma^2 = 1.2$). Each participant did a writing task and a semi-structured interview.

Each participant was asked write a metaphor that expresses a connection between an abstract concept and concrete object presented to them. They are given the following example for the seed *love is a stream*:

> Love is something that just drags me along. Like a stream it just takes me in whatever direction it is going.

We present each participants with six seed metaphors. The metaphors are generated by combining a random word from a set of poetic themes (e.g. *love*) with a random word from a set of concrete nouns (e.g. *stream*) [8]. Participants are asked to write about these seed metaphors one at a time–3 with Metaphoria and 3 without. All participants were given the same seed metaphors in the following order:

- *gratitude is a stream*
- *peace is a king*
- *jealousy is sand*
- *consciousness is a shadow*
- *loss is a wing*
- *friendship is snow*

To counterbalance the experiment, half the participants could use Metaphoria with the first three metaphors, and half use it with the last three metaphors. Figure 3 shows how the interface is presented in each case.

After the participant completes the task, the first author conducts a semi-structured interview in which all participants are asked the same set of core questions, with follow-up questions asked as specific issues come up. During the interview, the participant or interviewer could use the interface to go back and look at what the participant wrote, or interact with the suggestions again.

In this study we are testing Metaphoria for coherence to context. If the suggestions are not coherent, participants will not be able to use them to write coherent sentences, which is their goal. Thus, usage is a strong signal for coherence. We also test for divergent outcomes by looking at the variety of responses. If Metaphoria does not support divergent outcomes, metaphors written across participants will be more similar when using Metaphoria than not.

Results

Coherence to context. 12 of 16 participants used the suggestions to the complete the task. Although all participants were given the same suggestions in the same order, they used a variety of different suggestions. For instance, given the seed metaphor *peace is a king*, P10 used the suggestion 'peace is for

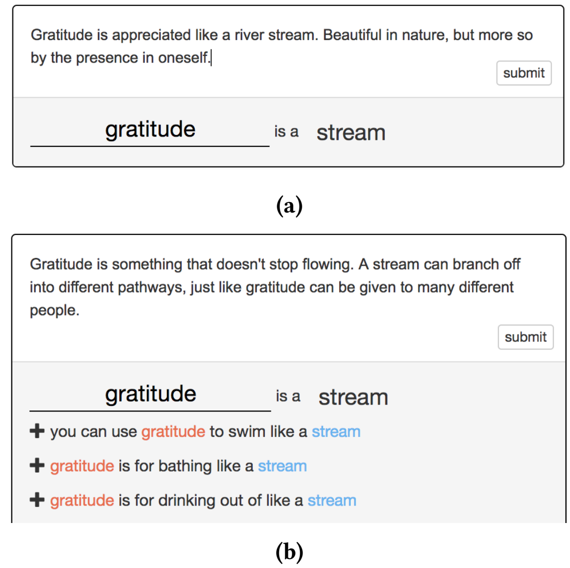

Figure 3: Interface for constrained writing task, in which participants wrote extended metaphors without suggestions (a) and with suggestions (b). Figure includes responses from P12 (a) and P10 (b).

leading the people like a king' while P6 used the suggestion 'peace is for rallying the troops like a king'.

Some participants were inspired by multiple suggestions, like P1 who used two suggestions, 'friendship is for beautiful vistas like snow' and 'friendship often arrives with a storm like snow', to write the following metaphor:

> Friendship often breaks out from kindness. It is a snow that often falls around christmas.

Many participants were impressed by the quality of the suggestions, like P8 who said:

> "I like 'you can use gratitude to wash something like a stream'. That's something I wish I had come up with. That's creative."

Several of these participants acknowledged that the quality of the suggestions varied. P3 said that although some of the metaphors didn't make immediate sense, they thought that the metaphors could make immediate sense to someone else.

All participants were asked to choose one suggestion that was bad in some way and discuss why. Most participants spent some time rereading suggestions to select one. During this process, several participants discovered that a suggestion they previously thought did not make sense they could in fact interpret. P4 said:

> "With this one I was sort of a little confused, 'peace is for moving forward and backwards in checkers like a king', I guess it makes sense

> now that I say it out loud. It's saying that peace doesn't have any limits on it."

Of the 4 participants who did not use the suggestions, 3 said this was because the suggestions did not make sense. They often said the suggestions were too literal or simply nonsensical. However, P12 said the suggestions did make sense, but she did not want to use them because she wanted to demonstrate that she could write creative metaphors on her own. We come back to this in the Discussion section.

Divergent outcomes. The suggestions may be coherent, but if participants end up writing very similar responses then Metaphoria is not supporting divergent outcomes for writers. We report both quantitative and qualitative results.

To quantitatively measure this, we measure the variation of responses across all participants when they did or did not use Metaphoria. Here we define variation as the distribution of distances between all responses–high variation means all responses were very different from all other responses. We measure distance as the Word Mover's Distance between two responses.

The responses without Metaphoria act as a baseline for the variance we expect to see in the responses. If participants were staying close the suggestions, as opposed to expanding or shifting the ideas, we would expect there to be less variation with Metaphoria. Less variation means similar ideas, words, and phrasing. As a reminder, all participants received the same suggestions when they had access to Metaphoria.

Our hypothesis is as follows:

- H5: The variation in responses with Metaphoria is as least as large as the variation in responses without.

We compare the variation per seed metaphor with and without Metaphoria. There is no significant difference in the variation of the responses for 4 of the 6 seed metaphors. For *consciousness is a shadow* there is significantly greater variation with Metaphoria; for *jealousy is sand* there is significantly greater variation without.

Table 7 shows examples from participants who said they were inspired by the same suggestion, demonstrating the wide range of directions participants took the idea, as well as examples of the more convergent responses.

Qualitatively participants did not feel like the suggestions boxed them in but rather inspired them to come up with new ideas. P4 expressed well how he would be inspired by a suggestion:

> "I saw 'gratitude is for bathing like a stream' and that made me think, well, how big is a stream? It started making me think about its size."

To demonstrate how far he took this idea, here is his final response to *gratitude is a stream*:

> Gratitude can be difficult to feel, or to notice, much like a stream that runs down the gutter of the road in a rainstorm. And like all streams, it can easily run dry–and you might not realize it's gone until it's too late.

	'gratitude is for bathing like a stream'
P6	Like a stream, you can bathe in gratitude and as the stream cleans your body, gratitude cleans your soul.
P13	A stream, to me, is rapid and powerful and has the ability to sweep you away. Gratitude offered by a friend or even a stranger is a stream in this way; it has the unexpected power to swell your heart with positive emotions and completely sweep you away.
	'jealousy can irritate skin like sand'
P16	Jealousy is a sand. It finds a way to irritate and conflict trouble of mind upon those whom it possesses.
P2	Jealousy can itch and irritate your mental behavior similar to the sand that clings on to your clothes and feet.

Table 7: Metaphoria mostly resulted in distinct responses, even when writers used the same suggestion, as in the 'gratitude' examples. But sometimes suggestions resulted in very similar responses, as in the 'jealousy' example.

We were worried that certain suggestions would be far more coherent than others, or that there would be a strong ordering effect, and therefore participants would always choose the same suggestions and write similar responses. However, as seen in the above analysis, this was not the case. Even when participants chose the same response, they would write radically different things.

6 STUDY 3: EXPERT WRITERS

This study evaluates if Metaphoria can adapt to a writer's own goals, and tests the system on inputs we did not expect. Our previous studies show Metaphoria is **coherent to context** and produces **divergent outcomes**; now we tackle whether these properties hold in real tasks which span a wide range of writer intentions.

Methodology

We gave three professional poets a 15 minute tutorial of Metaphoria and then asked them to write a poem on a subject of their own choosing using Metaphoria in some way. The poets wrote for around 30 minutes each. We then conducted a semi-structured interview, and utilized having Metaphoria available to discuss their process and response.

In this study, we gave participants access to the full interactivity of Metaphoria: they could enter in their own source concept, as well as a generate new vehicles, which are drawn randomly from a list of common poetic vehicles.

The poets were recruited through a mailing list for current and past MFA in Creative Writing students at a local university. All had a regular writing practice, were published poets, and one also held a teaching position in which they taught poetry writing workshops to undergraduates.

Results

Coherence to context. All poets used several of the suggestions in their poem. Part of each poem is reproduced in Table 8, where words they input into Metaphoria are highlighted in pink and phrases from the suggestions they used are highlighted in green.

The context each poet brought to Metaphoria was very different. PO1 initially entered the word *island*; the first line of their poem was inspired by the suggestion 'island can fill a glass like wine', though they first spent several minutes with other suggestions like 'island can travel over water like a ship' and 'island can age over time like wine'. PO2 was initially inspired by suggestions for the seed metaphor *work is a garden*, where *work* was input during the tutorial; several words in the first stanza came from the suggestions for this seed. Later they input the words *swaying* and *she*.

PO3 brought a very different type of context. They input many more words than the other two poets, more interested in finding interesting suggestions than crafting a poem with a particular direction; almost every line derives from some part of Metaphoria. They first input *sales*, then *marketing*, before exploring the word *metaphor*. Their first line is inspired by the suggestion 'metaphor is for restoring quiet like a bell'. Later they input words like *time*, *guns*, *history*, *elections*, *laughter*, and *stone*, to mention only a small number.

All poets found suggestions that resonated with them, though they were discriminate and often searched through several seeds before finding something they used. However, there were clearly different styles of use: PO1 and PO2 composed poems with some kind of linear narrative or thought, and used Metaphoria on words they had already written, often finding a suggestion that would finish the line they were working on. In contrast, PO3 input words they thought might be make for interesting metaphors, or words they simply overheard (we met in a coffee shop), many of which never made it into the poem. PO3's use was more like collecting interesting phrases, which they then arranged and edited.

Divergent outcomes. The resulting poems were of dramatically different styles, both due to each poet's differing usage of Metaphoria and their different writing styles. When explicitly asked about the expressiveness of the system, all poets noted that established writers have their own style and the system was unlikely to dramatically change it. Both PO2 and PO3 thought Metaphoria would increase the creativity of amateur poets, who tend to get stuck in cliche language; they thought the unexpectedness of the word combinations was likely to help.

However, PO2 did bring up concerns of ownership. While they did not think that Metaphoria limited them, they were concerned about using suggestions from Metaphoria that were too different from their intention, even if these suggestions were very good. PO3 used Metaphoria most liberally, yet had no such concerns. They drew a comparison between Metaphoria and Instagram, noting that while Instagram has produced a genre of photography that is very recognizable and thus the photos are somewhat similar, it has also produced unexpected and creative artworks. They speculated that Metaphoria might create a genre of Metaphoria-style poems, but would also allow poets to move in new and exciting directions. We analyze these concerns in the Discussion.

7 DISCUSSION

Ownership concerns and cognitive models of usage

Ownership is extremely important to writers. It is essential that writers feel like they own their material, and Metaphoria was designed to augment writer's abilities, not replace them. To tackle this head on, we asked all participants about how much ownership they felt for what they wrote. Each poet in the expert study discussed their relationship to Metaphoria using a different cognitive model:

PO1 was unconcerned about the influence of the system on their writing; they thought of Metaphoria "like a calculator for words." They used Metaphoria as a **cognitive offloading tool**, outsourcing specific moments of word generation and allowing them to focus on other goals like the overall direction of the poem and the flow of the lines.

PO2 was concerned about using Metaphoria when it produced particularly good images. For example, they thought the line 'she is used for currency and jewelry' was "an amazing line of poetry" but "definitely altered the direction of the poem," which worried them. In this case, they treated Metaphoria as a **co-creative partner** who contributed more to the poem than PO2 felt comfortable with.

PO3 used Metaphoria much more liberally–with no particular intended direction, they were more playful and wanted to uncover interesting Metaphoria-style combinations. In this case Metaphoria was used as a **casual creator** [4], an interactive system that encourages exploration in the creation or discovery of surprising new artifacts.

In the novice study, 4 of the 16 participants said that they felt less ownership over the final results because some amount of work was being done by the system; this reaction

PO1's response	PO2's response	PO3's response
My island fills glasses like wine,	Garden Work	Metaphor for restoring quiet
		Use a gun to paint a room
i'ts vines wrap around my	with my mother, her tulips flaming blue	Addiction can clog a sink drain like hair
	and yellow, laboring to bloom beneath	History can win a war
new mouth like grapes.	her palms, the soft lawn grating against	The garden of wasted time
	early spring. We are wasting time, lingering	Fear to extinguish a fire like sand
This new America is used to building things,	under the porch light before dark, flirting	ice is for finding the source of light
	with enemy weeds before my father returns	swimming is like snow. it is for children
anew, strange comfort like the rest of an air-bed	home, drunk and swaying like a storm.	You can use caution to build fear in a movie
		You can use witchcraft to listen to music like an ear
at dusk.	She is used for currency and jewelry	Corruption can outrun you like a horse
	and lighting the pathway. She is for	
How new is new?	making flowers rise up to collide	
	with her hands.	

Table 8: Part of responses from three professional poets working with Metaphoria. Words highlighted in pink were input into Metaphoria by the poets, while words and phrases highlighted in green were suggestions that poets used.

was strongest in those that thought the suggestions were particularly good. In this case, likely they saw Metaphoria as a **co-creative partner** contributing too much to their work.

Thus algorithmic suggestions are used differently depending on the cognitive model users project–a offloading tool that does grunt work (like a dictionary or thesaurus), a true partner that can do too much or too little, or a casual creator that allows the user to explore. Systems designers should be aware of different cognitive models and build tools that support creators without threatening their agency.

Design implications from ownership concerns

All participants in the novice and expert studies acknowledged that they happily accept prompts, ideas, feedback, and edits from people (both teachers and peers) without feeling loss of ownership. For machines to become acceptable co-creative partners, there are two design avenues:

Increased transparency can make the mechanisms of the machine more apparent. This way it feels more like a 'word calculator' than a system trying to outsmart you. Presentation of the suggestions may matter; more studies should be done on how this affects perceived ownership. It could be that for some writers full sentences (even ones constructed naively from templates) are more threatening than a key dangling phrase.

Increased interactivity integrates the person into the creation process. The more interaction, the more the machine can be seen as a causal creator that helps explore new spaces. This interaction with a computational system can give people comfort and agency, similar to how we learn to converse with people offering us advice. Systems could draw suggestions from different contexts or genres that writer can pick or specify, such as a particular novel, technical text, or set of tweets, and include tunable parameters, such as suggestion length, vocabulary sophistication, connotative constraints (like negative/positive), or phonetic features.

Limitations and future work

Interaction with Metaphoria is limited to inputting a source word and requesting a new the vehicle word. This does not take into consideration what a writer has previously written, either the text of whatever they are currently working on or past work that might be relevant. To make systems more personalized, we could highlight how suggestions relate to a writer's previous work, or phrase suggestions in a syntactic style specific to the writer.

Additionally, Metaphoria can be expanded to other domains like journalism. For example, we can provide suggestions to metaphorically explain scientific concepts for lay people. "*CRISPR* can cut *genes* like *scissors* can cut *paper*." We can adapt the system by training a custom word embedding to provide representations for words in specialized domains, like medical research, technology, or law.

8 CONCLUSION

Motivated by past work on user-centric creativity support, we created Metaphoria, an interactive interface for generating metaphorical connections. Our evaluations demonstrated that Metaphoria generates suggestions coherent to context and supports divergent outcomes for writers. We discuss ownership and cognitive models in human-computer collaboration, and present future work for more interactive and transparent systems that can further empower creators.

ACKNOWLEDGEMENTS

Katy Ilonka Gero is supported by an NSF GRF (DGE - 1644869).

REFERENCES

[1] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2010. Soylent: A Word Processor with a Crowd Inside. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology (UIST '10)*. ACM, New York, NY, USA, 313–322. https://doi.org/10.1145/1866029.1866078

[2] Jill Burstein, Beata Beigman Klebanov, Norbert Elliot, and Hillary Molloy. 2016. A Left Turn: Automated Feedback and Activity Generation for Student Writers. In *Language Teaching, Learning and Technology*. 6–13. https://doi.org/10.21437/LTLT.2016-2

[3] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. ACM, New York, NY, USA, 329–340. https://doi.org/10.1145/3172944.3172983

[4] Kate Compton and Michael Mateas. 2015. Casual Creators.. In *ICCC*. 228–235.

[5] Nicholas Davis. 2013. Human-computer co-creativity: Blending human and computational creativity. In *Ninth Artificial Intelligence and Interactive Digital Entertainment Conference*.

[6] Nicholas Davis, Chih-Pin Hsiao, Kunwar Yashraj Singh, and Brian Magerko. 2016. Co-creative drawing agent with object recognition. In *Twelfth Artificial Intelligence and Interactive Digital Entertainment Conference*.

[7] Gilles Fauconnier and Mark Turner. 2008. *The way we think: Conceptual blending and the mind's hidden complexities*. Basic Books.

[8] Andrea Gagliano, Emily Paul, Kyle Booten, and Marti A Hearst. 2016. Intersecting Word Vectors to Take Figurative Language to New Heights. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*. 20–31.

[9] Dedre Gentner. 1983. Structure-Mapping: A Theoretical Framework for Analogy*. *Cognitive Science* 7, 2 (1983), 155–170. https://doi.org/10.1207/s15516709cog0702_3

[10] Katy Gero and Lydia Chilton. 2018. Challenges in Finding Metaphorical Connections. In *Proceedings of the Workshop on Figurative Language Processing*. 1–6.

[11] Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. 2016. Generating topical poetry. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 1183–1191.

[12] Karni Gilon, Joel Chan, Felicia Y. Ng, Hila Liifshitz-Assaf, Aniket Kittur, and Dafna Shahaf. 2018. Analogy Mining for Specific Design Needs. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 121, 11 pages. https://doi.org/10.1145/3173574.3173695

[13] Sam Glucksberg, Matthew S McGlone, Yosef Grodzinsky, and Katrin Amunts. 2001. *Understanding figurative language: From metaphor to idioms*. Number 36. Oxford University Press on Demand.

[14] Eric Horvitz. 1999. Principles of Mixed-initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '99)*. ACM, New York, NY, USA, 159–166. https://doi.org/10.1145/302979.303030

[15] Julie S. Hui, Darren Gergle, and Elizabeth M. Gerber. 2018. IntroAssist: A Tool to Support Writing Introductory Help Requests. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 22, 13 pages. https://doi.org/10.1145/3173574.3173596

[16] Jennifer Jacobs, Joel Brandt, Radomír Mech, and Mitchel Resnick. 2018. Extending Manual Drawing Practices with Artist-Centric Programming Tools. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 590, 13 pages. https://doi.org/10.1145/3173574.3174164

[17] Joy Kim, Justin Cheng, and Michael S. Bernstein. 2014. Ensemble: Exploring Complementary Strengths of Leaders and Crowds in Creative Collaboration. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. ACM, New York, NY, USA, 745–755. https://doi.org/10.1145/2531602.2531638

[18] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*. 957–966.

[19] George Lakoff and Mark Turner. 2009. *More than cool reason: A field guide to poetic metaphor*. University of Chicago Press.

[20] Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2010. Automated grammatical error detection for language learners. *Synthesis lectures on human language technologies* 3, 1 (2010), 1–134.

[21] Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 302–308.

[22] Zach Lieberman, T. Watson, and A. Castro. 2015. openFrameworks. http://openframeworks.cc/about. Accessed: 2018-09-19.

[23] Tal Linzen. 2016. Issues in evaluating semantic spaces using word analogies. *CoRR* abs/1606.07736 (2016). arXiv:1606.07736 http://arxiv.org/abs/1606.07736

[24] H. Liu and P. Singh. 2004. ConceptNet — A Practical Commonsense Reasoning Tool-Kit. *BT Technology Journal* 22, 4 (01 Oct 2004), 211–226. https://doi.org/10.1023/B:BTTJ.0000047600.45421.6d

[25] Neil Maiden, Konstantinos Zachos, Amanda Brown, George Brock, Lars Nyre, Aleksander Nygård Tonheim, Dimitris Apsotolou, and Jeremy Evans. 2018. Making the News: Digital Creativity Support for Journalists. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 475, 11 pages. https://doi.org/10.1145/3173574.3174049

[26] Enrique Manjavacas, Folgert Karsdorp, Ben Burtenshaw, and Mike Kestemont. 2017. Synthetic literature: Writing science fiction in a co-creative process. In *Proceedings of the Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2017)*. 29–37.

[27] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 3111–3119.

[28] Jeffery Scott Mio. 1997. Metaphor and Politics. *Metaphor and Symbol* 12, 2 (1997), 113–133. https://doi.org/10.1207/s15327868ms1202_2 arXiv:https://doi.org/10.1207/s15327868ms1202_2

[29] Kai Niebert, Sabine Marsch, and David F. Treagust. 2012. Understanding needs embodiment: A theory-guided reanalysis of the role of metaphors and analogies in understanding science. *Science Education* 96, 5 (2012), 849–877. https://doi.org/10.1002/sce.21026 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/sce.21026

[30] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I Lead, You Help but Only with Enough Details: Understanding User Experience of Co-Creation with Artificial Intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 649, 13 pages. https://doi.org/10.1145/3173574.3174223

[31] Hugo Gonçalo Oliveira. 2012. PoeTryMe: a versatile platform for poetry generation. *Computational Creativity, Concept Invention, and General Intelligence* 1 (2012), 21.

[32] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[33] James L. Peterson. 1980. Computer Programs for Detecting and Correcting Spelling Errors. *Commun. ACM* 23, 12 (Dec. 1980), 676–687. https://doi.org/10.1145/359038.359041

[34] Casey Reas and Ben Fry. 2004. Processing. http://processing.org. Accessed: 2018-09-19.

[35] Mitchel Resnick, Brad Myers, Kumiyo Nakakoji, Ben Shneiderman, Randy Pausch, Ted Selker, and Mike Eisenberg. 2005. Design principles for tools to support creative thinking. In *NSF Workshop Report on Creativity Support Tools*. Citeseer, 25–36.

[36] Melissa Roemmele and Andrew S. Gordon. 2018. Automated Assistance for Creative Writing with an RNN Language Model. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion (IUI '18 Companion)*. ACM, New York, NY, USA, Article 21, 2 pages. https://doi.org/10.1145/3180308.3180329

[37] Paul J. Silvia and Roger E. Beaty. 2012. Making creative metaphors: The importance of fluid intelligence for creative thought. *Intelligence* 40, 4 (2012), 343 – 351. https://doi.org/10.1016/j.intell.2012.02.005

[38] Robin Sloan. 2016. Writing with the machine. https://www.robinsloan.com/notes/writing-with-the-machine/. Accessed: 2018-09-19.

[39] Jaime Teevan, Shamsi T. Iqbal, and Curtis von Veh. 2016. Supporting Collaborative Writing with Microtasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 2657–2668. https://doi.org/10.1145/2858036.2858108

[40] Tony Veale. [n. d.]. Thesaurus Rex. http://ngrams.ucd.ie/therex3/. Accessed: 2018-09-19.

[41] Tony Veale. 2013. Less Rhyme, More Reason: Knowledge-based Poetry Generation with Feeling, Insight and Wit.. In *ICCC*. 152–159.

[42] Tony Veale and Yanfen Hao. 2007. Comprehending and generating apt metaphors: a web-driven, case-based approach to figurative language. In *AAAI*, Vol. 2007. 1471–1476.

[43] Tony Veale and Guofu Li. 2016. Distributed Divergent Creativity: Computational Creative Agents at Web Scale. *Cognitive Computation* 8, 2 (01 Apr 2016), 175–186. https://doi.org/10.1007/s12559-015-9337-9

[44] Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov. 2016. Metaphor: A computational perspective. *Synthesis Lectures on Human Language Technologies* 9, 1 (2016), 1–160.

[45] Ge Wang. 2008. *The ChucK Audio Programming Language: An Strongly-timed and On-the-fly Environ/mentality*. Ph.D. Dissertation. Princeton University.

[46] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods* 45, 4 (01 Dec 2013), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x