TwitSong: A current events computer poet and the thorny problem of assessment.

by

Carolyn Elizabeth Lamb

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2018

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Ruli Manurung Google Japan
Supervisor(s):	Daniel G. Brown Professor and Director, School of Computer Science, University of Waterloo Charles L.A. Clarke Professor, School of Computer Science, University of Waterloo
Internal Members:	Edith Law Assistant Professor, School of Computer Science, University of Waterloo Dan Vogel Associate Professor, School of Computer Science, University of Waterloo
Internal-External Member:	Dave DeVidi Professor, Dept. of Philosophy, University of Waterloo

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Several parts of this thesis were previously published as academic conference or journal papers with more than one author. In every case, I was the first author listed. The overall design of the research, the literature review, the implementation of systems, the design and implementation of experiments, the interpretation of results, and the writing component were consistently done by me. My supervisors' role was to provide guidance and discussion, to occasionally point me in the direction of additional literature, to help provide structure in order to get me to finish my work on time, and to suggest revisions to my drafted work. No other authors besides myself and my supervisors were included in any publication.

Sections 2, 3.1, 4, 5.1, and 5.2 were written and published with me as the first author and both Dan Brown and Charlie Clarke supervising as described in the previous paragraph. Sections 3.2 and 5.3 describe previously unpublished research undergone by me, with Dr. Brown in the supervisory role previously described. Sections 1, 6, and 7 were written entirely by me, with Dr. Brown and Dr. Clarke's role limited to reading a draft and suggesting revisions.

Nonwithstanding the above, the academic "we" is used throughout this thesis for style and consistency.

Abstract

This thesis is driven by the question of how computers can generate poetry, and how that poetry can be evaluated. We survey existing work on computer-generated poetry and interdisciplinary work on how to evaluate this type of computer-generated creative product. We perform experiments illuminating issues in evaluation which are specific to poetry. Finally, we produce and evaluate three versions of our own generative poetry system, TwitSong, which generates poetry based on the news, evaluates the desired qualities of the lines that it chooses, and, in its final form, can make targeted and goal-directed edits to its own work. While TwitSong does not turn out to produce poetry comparable to that of a human, it represents an advancement on the state of the art in its genre of computer-generated poetry, particularly in its ability to edit for qualities like topicality and emotion.

Acknowledgements

Thank you to Dan and Charlie, my endlessly patient supervisors; to Edith, Dan, Dave, and Ruli, my thesis committee (especially to Ruli, who Skyped in to my defense all the way from Jakarta); to Ming Li for chairing my defense; and to OGS, NSERC, Google, and the University of Waterloo for scholarship and grant money which variously supported me and my supervisors during the research that resulted in this thesis. I also want to thank the ICCC community, particularly Anna Jordanous, Gillian Smith, and Hannu Toivonen, for their enthusiastic engagement and support as I pursued this research.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Can a computer be creative? Could a computer create its own works, such as poetry, that are of interest to humans? And how would we know if it did?

The computational creativity research community has been studying questions like these. Computational creativity is, "The philosophy, science and engineering of computational systems which, by taking on particular responsibilities, exhibit behaviours that unbiased observers would deem to be creative" (Colton and Wiggins, 2012). Computational creativity is of interest not only for its potential applications, which include personalized art creation, new forms of art that are only possible with computers, and co-creative systems that enhance a human's creative learning and ability; but also because, by implementing theories about the cognitive science of creativity in a computational system, we learn about the underlying science of human creativity (Boden, 1990).

Our interest is specifically in computer-generated poetry, which has been produced in some form since the 1950s (Lutz, 1959) but which has been enhanced with artificial intelligence techniques more recently. We set out, at the beginning of our PhD studies, to create a system that generated poetry or song lyrics based on the news. Immediately we were confronted with several questions that arise as the natural result of this goal: How does a computer create poetry? What textual traits do we want a finished poem to have? How does a computer make decisions about these traits? If a computer is meant to generate poetry, how do we measure its success at this task? In particular, how do we tell if a generated poem is any "good"—and how do we tell if it is creative?

Searching for answers to these questions, we have performed a comprehensive survey of current knowledge about creativity that can, or should be, applied to computational creativity evaluation. We find that a bevy of proposed evaluation techniques exist, not all of which draw from the appropriate psychological or humanistic literature, and not all of which are carried out appropriately in practice. Moreover, there are several qualitatively different perspectives from which to judge creativity, including judgment of the creative agent itself, of the creative agent's output, of the process used to generate the output, and of the place of agent and output in the wider society.

We next survey specifically the techniques that are used to generate poetry. We find that many current researchers use AI techniques of optimization and knowledge representation to improve the existing, basic generative techniques for poetic text. Neural networks are another promising technique with which poetry

is currently being written. Although many interesting and promising techniques exist, there are particular challenges for computational poetry, including coherence, for which few existing techniques produce good results.

Leaving these surveys with some questions about poetry still unanswered, we perform a few experiments in the evaluation of both human- and computer-generated poetry. Our experiments add some new discoveries to the computational creativity research field. In particular, we find that non-expert raters (even though they are frequently used to evaluate poetry in practice) are even more inappropriate for poetry evaluation than theory would suggest, largely because most non-expert raters have little exposure to modern contemporary poetry and great difficulty understanding it. We also find that expert raters have limited agreement when it comes to evaluating computer-generated poems. However, by studying the statements made by expert raters about computer-generated poems, we derive a domain-specific, evidence-based set of four criteria (with several sub-criteria) that we hypothesize describes the main desiderata for computer-generated poetry in expert minds.

Finally, we produce and evaluate three successive versions of our own generative poetry system, TwitSong, which writes poems based on the news. TwitSong functions by mining candidate lines from a source text and evaluating those lines on key metrics such as topicality, emotion, and imagery. The third generation of TwitSong also is able to edit its candidate lines based on specific goals, such as revising an off-topic line to be more topical. While (few) other poetry systems exist that can edit their lines in this manner, editing based on semantic criteria such as topicality is an advancement on the state of the art. Results from the second and third generations are mixed, likely because our operationalizations of line evaluation metrics are not as sophisticated as human judgment. Nevertheless, our results show that TwitSong's approach works in principle, that our choices of criteria are appropriate and useful in principle, and that making targeted edits to the lines does improve them, though not always in the manner that we anticipated.

These various research topics are distributed throughout the rest of the paper as follows. In Chapter 2, we have our survey of interdisciplinary methods of computational creativity evaluation, which is the most comprehensive survey on this topic to date. In Chapter 3, we have the remainder of our surveys of related work, organizing the field of computer-generated poetry into a rough taxonomy and giving many examples of the state of the art in the taxonomy's most relevant sections. In Chapter 4, we describe our original experiments in computational poetry evaluation, including a test of non-expert human raters using computational creativity evaluation techniques on human poetry, and a test of expert human raters using human evaluation techniques on computer-generated poetry. In Chapter 5, we discuss the development, design, and evaluation of all three versions of TwitSong, our generative poetry system. Finally, in Chapter 6, we reflect on what we have learned from all four of these branches of research, what still remains to be done, and what we would like to see in future similar poetry systems.

Chapter 2

Related work: Evaluating computational creativity

2.1 Introduction

If we are to build a creative system as anything other than an artistic project, we must have a way of evaluating it. That is to say, we must be able to state whether or not the system met its goals. The challenge of evaluating creative systems is a much-studied topic in computational creativity. In particular, computer scientists in the discipline of computational creativity have preoccupied themselves with the question of how to tell if a system is *creative*. Many creative software systems have been designed both for artistic creativity (including music, visual art, and storytelling) and for other forms of creativity (such as mathematics, design, and code generation (Loughran and ONeill, 2017), and all need to be evaluated. To address this question, we need both a definition of creativity and a methodology for assessing it.

The growing and varied computational creativity community has generated many theories and methodologies for the evaluation of creativity (Aguilar and Pérez y Pérez, 2014; Bown, 2014; Burns, 2015; Colton, 2008; Gervás, 2002; Jordanous, 2012a; Negrete-Yankelevich and Morales-Zaragoza, 2014; Pease et al., 2001; Ritchie, 2001). However, these many proposals mutually contradict, and many of them suffer from a lack of reliability or validity. A major contributor to this situation is a lack of interdisciplinary knowledge about creativity. Psychologists, philosophers, cognitive scientists, and others have attempted to tackle the problem of creativity evaluation from their own perspectives, and computer scientists working on creative systems can learn from their efforts. At present, many proposals for evaluation are not well grounded in the psychology or philosophy of human creativity. (There is, of course, an argument that computational creativity need not resemble human creativity—an argument which we will explore in Section 2.7—but it is inadvisable to reject theories of human creativity without first understanding them.)

This chapter will guide the reader through the major proposals about how to evaluate creative software, with a particular focus on proposals from psychology, philosophy, and other fields that should be used to guide a programmer's decisions. It should be noted that this is not only a summary of existing work, but also a contribution in itself, as we have surveyed interdisciplinary research on the nature of creativity

in humans, much of which is not used, or underused, in computational creativity, and have made recommendations regarding its implications for the field. A prior version of this chapter was published in ACM Computing Surveys in 2018 (Lamb et al., 2018).

In Sections 2.2 through 2.6, we introduce theories of creativity evaluation from four perpectives– including both human contexts and computational ones. We follow this, in Section 2.7, with a discussion of questions as to whether creativity can be evaluated at all. Finally, in Section 2.8, we discuss practical issues specific to computational creativity and some common pitfalls of evaluation in practice. Each section has specific takeaways for the researcher who wants to apply the ideas to their work. These takeaways will be summarized in the chapter's conclusion, Section 2.9.

2.2 Theories of creativity

A variety of theories of creativity evaluation exist. When comparing evaluations to each other, it is useful to group them based on their theoretical perspective. One useful taxonomy of theories, which we use to structure this paper, is known as the four Ps: Person, Process, Product, and Press (Rhodes, 1961).

- **Person** is the human (or non-human agent) who is seen as creative. Person theories study what it is about the agent that makes them creative.
- **Process** is the set of internal and external actions the agent takes when producing a creative artifact. Process theories study what sort of actions are undertaken when creative work is done.
- **Product** is an artifact, such as an artwork or a mathematical theorem, which is seen as creative or as having been produced by creativity. Product theories study what it is about the product that makes it worthy of being called creative.
- **Press** is the surrounding culture which influences people, processes, and products and which judges them as creative or uncreative. Press theories study what it is that leads a culture to view something as creative.

These four Ps originate with Rhodes (1961) and were introduced to computational creativity by Jordanous (2016a). Jordanous suggests the use of the word "producer" in place of "person", to emphasize that the creative agent need not be a human. For this paper, we use "person" so as to match the psychological literature.

We will describe evaluation methods from each of these perspectives in turn. Of course, it is possible to evaluate from more than one perspective. For example, Colton *et al.* (2011) introduce, in the same paper, both the FACE (process) model for assessing the creativity of an act, and the IDEA (press) model for assessing its impact. The SPECS model (Jordanous, 2012a), as we will discuss, contains criteria that arguably cover all four perspectives (Jordanous, 2016c). Many evaluations that are mainly in one perspective also incorporate ideas from others. However, in order to be clear about the ideas behind each perspective, we will for the most part discuss them separately.

2.3 Person perspective

The Person perspective is grounded in psychometrics (the measurement of human mental traits). The goal of the Person approach is to discover what personality, emotional, and cognitive traits distinguish a more creative person from a less creative one. The Person perspective is potentially useful for researchers who want their systems to be viewed as creative agents because of their inherent traits.

For humans, hundreds of psychometric tests for creativity exist (Plucker and Renzulli, 1999), the most famous perhaps being the Torrance Tests (Torrance, 1968). These tests prompt the person to generate many related ideas—for example, by asking how many ways one can use a chair. This constitutes a test of *divergent thinking*—the ability to come up with varied and unusual ideas—which is a popular Person-based definition of creativity (Cropley and Cropley, 2005). However, in many theories, divergent thinking is only useful when it alternates with convergent thinking—the use of conventional knowledge to evaluate and develop the ideas (Csikszentmihalyi, 1996).

Aside from divergent and convergent thinking, Person-based creativity can also be measured through self-report, about a person's perception of their own creativity or about their actual creative achievements, or both. Several scales exist for measuring creativity through self-report, including the Kaufman Domains of Creativity Scale (Kaufman, 2012) and the Creative Achievement Questionnaire (Carson et al., 2005).

Computers can pass some psychometric creativity tests. Olteţeanu and Falomir's system comRAT-C solves a Remote Associates Test by using associative spreading through a database of common bigrams (Olteţeanu and Falomir, 2015). Their system OROC passes an Alternative Uses Test by comparing information about objects' uses and physical properties (Olteţeanu and Falomir, 2016). Gross *et al.* (2012) solve a Remote Associates Test using mined word associations, and discuss how this technique could be used in other creative tasks. Psychometric tests more distantly related to creativity, such as the analogical reasoning of Raven's Progressive Matrices, have also been the subject of computational creativity research (Johner et al., 2015). However, these systems do not do other creative work besides taking tests. Since the structure of a computer system differs from the structure of a human mind, it is not evident that the success of a computer at a psychometric test, *per se*, is a useful proxy for its general creative ability—nor are the above authors making that claim.

Another avenue of Person research is to study famous creative people. This is known as the historiometric approach. Historiometric research reveals multiple successful approaches to creativity (Policastro and Gardner, 1999), with a few general traits in common, such as an ability to distinguish promising avenues of work from less promising ones.

Policastro and Gardner (1999) review historiometric studies and sort creators into four categories, based on two distinctions: creators who work with objects and symbols versus those who work with people, and those who strive for excellence within a domain versus those who challenge the domain's foundations. Each combination of these traits produces a creative personality type. Masters, who achieve excellence within an object- or symbol-based domain, include Mozart, Rembrandt, and Shakespeare. Makers, who create a new object- or symbol-based domain, include scientists like Einstein and Darwin as well as maverick artists like Stravinsky or Joyce. Introspectors, who work for excellence in a people-based domain, use art to express their inner selves; these include Proust and Woolf. Their counterparts, the Influencers, work to change and challenge other people. Influencers include Gandhi, Mandela, and Eleanor Roosevelt. In Policastro and Gardner's theory, those who challenge a domain's foundations are not necessarily more creative than others, only creative in a different way. This is a point that we will return to in Section 2.5.1.1.

Measuring the personal traits of computers is conceptually difficult, but some researchers have attempted it. The Creative Tripod (Colton, 2008) works by attributing traits to a system, but since the Tripod's focus is on convincing people that the system has these characteristics, it will be discussed under Press in Section 2.6.1. A few researchers have attempted to model traits associated with creativity, such as curiosity (Grace et al., 2017). More commonly, as with the Tripod, attributing traits to a computer is something humans do as the result of information from one of the other three perspectives. Unless a system is designed to exhibit specific human cognitive traits, it is likely better not to use the Person perspective for computers.

2.4 Process perspective

The Process perspective includes any theory of *how* creative products are made—that is, what cognitive steps must be taken in order for an activity to be creative. Many of these theories are descriptive rather than evaluative. Since they are created by researchers studying humans, some are built on the assumption that a human cognitive structure, including unconscious reasoning, is already in place. Thus, these ideas have not always been straightforwardly taken up in computational creativity. However, the Process perspective includes many important ideas that can and should influence the design of a creative system. The Process perspective is especially useful for researchers who want to model human creativity, or who want to make an argument that their system is creative because of the kinds of tasks it does.

2.4.1 Conceptual space

The most popular Process theory in computational creativity is Boden's (1990) theory of conceptual space. Boden divides creativity into three types: combinatorial, exploratory, and transformational. Combinatorial creativity occurs when two familiar ideas are put together in an unfamiliar way. The remaining types of creativity depend on the idea of a conceptual space: the space of all things that could be generated according to a set of rules. Exploratory creativity, by testing out the rules' implications, reaches accessible points in the space that have not been reached before. Transformational creativity changes the rules to reach points that were not accessible before. Wiggins (2006) refines Boden's idea of conceptual space and describes types of "aberration" which might lead one to revise the space and/or the way of searching.

Many Process theorists privilege transformational creativity. Several philosophers, for example, define originality as work that changes established rules (Bartel, 1985). To these philosophers, transformational creativity is more original than exploratory creativity. Dorin and Korb (2012) define creativity as "the introduction and use of a framework that has a relatively high probability of producing representations of patterns that can arise only with a smaller probability in previously existing frameworks". The more unlikely the new patterns were under previous frameworks, the more creative they are. Alternatively, a number of computational creativity researchers focus on combinatorial creativity, teaching a computer to blend seemingly unrelated concepts (Cunha et al., 2017; Gonçalves et al., 2017; Veale, 2013b)

Besold (2016) argues that certain processes in machine learning, such as Bayesian theory learning and inductive logic programming, are transformationally creative, as they involve the generation of novel (to the system) ideas. Besold argues that computational creativity, particularly scientific and problem-solving creativity, should use these processes.

2.4.2 Stage- and loop-based theories

The theories described above are vague as to how a human goes about transforming a conceptual space. However, more naturalistic models of the creative process are plentiful. One such theory is Wallas' (1926) four-stage theory: preparation, incubation, inspiration, and verification. The four-stage theory is based on case studies of how scientists have ideas. First, in preparation, the creative person gathers information related to their task. In incubation, the person ponders the information, makes connections, and often abandons the project for a time, while further connections are made unconsciously. In inspiration, something "clicks", and the person gets an idea based on a novel way of looking at the information. Then in verification, the person does a lot of hard work establishing, developing, and polishing their idea into a finished product.

Sadler-Smith (2015) suggests that the case studies on which the four-stage model is based also include a fifth stage, between incubation and inspiration: intimation. At the intimation stage, a solution to the problem exists in fringe consciousness, but has not yet been seized on by the conscious mind. There may be a feeling that a solution is coming, or an awareness of parts of the solution at the edge of the mind. According to Sadler-Smith, expertise helps a creative person progress to the inspiration stage by unconsciously evaluating their unconscious ideas. Csikzentmihalyi (1996) includes between inspiration and verification the stage of evaluation, in which the creative person uses their domain knowledge to decide if their idea is worth pursuing.

In practice, these elements do not necessarily follow each other in a tidy order, but are mixed together (Beardsley, 1965) or can repeat "fractally" as a series of smaller and smaller inspirations about details of the work (Csikszentmihalyi, 1996). This mixing leads researchers to consider creativity as a loop or an iterative process—but what characterizes the iteration?

One iterative theory is BVSR—Blind Variation, Selective Retention. BVSR was first proposed by Campbell (1960) and was further developed by Simonton (2011). In BVSR, a creative agent blindly generates many possible ideas before reflecting and choosing the best ones. For generation to be "blind" means that the probability of generating an outcome is decoupled from its utility: an increase in the utility of an outcome does not cause an increase in the probability of generating that outcome. This is also called the Darwinian theory of creativity (Kronfeldner, 2010) because of its superficial resemblance to natural selection, in which random mutations and recombinations create organisms which are selected for their fitness. However, a close correspondence between BVSR and evolution is not strictly necessary (Simonton, 2011).

Other psychologists and philosophers (*e.g.* (Dasgupta, 2011; Kronfeldner, 2010)) have criticised BVSR. The main criticism involves the notion that generation is blind. Dasgupta (2011) outlines famous cases of creativity where candidate solutions were generated based on schemas, building on partially successful previous ideas to form a progressively more correct solution. These mechanisms are not explained by BVSR. Mechanisms like analogical reasoning and spreading activation, which Simonton lists as examples of blind processes, are not actually blind: they depend on the structure of knowledge in an agent's mind, and on the thinking the agent has already done about the problem, which is likely to guide them towards more useful solutions (Kronfeldner, 2010).

Despite these problems with the concept of blindness, most theories of the creative process still involve a loop between generation and evaluation. (Note that the alternation between generation and evaluation is roughly equivalent to the alternation between divergent and convergent thinking mentioned in Section 2.3.)

Two cases in point are Ward, Smith, and Finke's (1999) Geneplore model and Garcia *et al.*'s (2006) ER model. In Geneplore—a portmanteau of "generate" and "explore"—an agent generates preinventive structures through some means, not necessarily blind: synthesis, transformation, and exemplar retrieval are mentioned (Ward et al., 1999). The agent then evaluates the preinventive structures and explores their properties and implications. This exploration leads either to refining the preinventive structures or discarding them and generating new ones. In the ER model–standing for Engagement-Reflection—the agent brainstorms ideas for solving a problem in the Engagement stage, then evaluates them at the Reflection stage (García et al., 2006). Ideas at the Engagement stage are not necessarily blind, but they are not evaluated at this stage. Ideas at the Reflection stage are discarded if their prerequisites are not met, or modified to make them more acceptable. They then form a partial solution to the problem. This partial solution is fed back into Engagement for elaboration, which is again reflected on, and so on, until the solution is complete.

Dahlstedt (2012), similarly, defines the two phases as "implementation" and "re-conceptualization". During implementation, an artist has two ideas: a conceptual description of the desired end product, and an instruction for how to get there. In some genres, such as improvisatory theatre, the description may be missing. In re-conceptualization, the artist compares their work to their conceptual description, and changes either the product or the description to bring them closer to each other. These changes can take various forms, such as additions, expansions, generalizations, mutations, new constraints, replacements of material, or even a new conceptual representation from scratch. When the work is finished, the conceptual representation is hidden; each person in the audience then makes their own conceptual representation of what they think the art is about.

The cultural psychologist Glăveanu (2015) defines these loops as being based in perspective-taking: an artist must switch between the perspective of a creator and the perspective of an audience viewing the art. An artist has interactions with real audience members throughout their career, which allows them to broaden the range of perspectives they can take on when evaluating their work. Basing the creative loop on perspective-taking defines it as a social phenomenon which is inextricably connected to the Press perspective. A very simple form of perspective-taking takes place in systems which evaluate their work based on data about what humans prefer. More advanced social reasoning would take place in a system that can actively interact with critics; we describe a few such systems in Section 2.6.6.

An important thing to note about all of these loop-based theories is that the evaluate portion of the loop is assumed to constitute a decision making process with some reasoning/inference involved. This reasoning may be relatively simple (as in the case of a genetic algorithm recombining prototype products based on their scores on a simplistic metric) or relatively advanced (as in Glăveanu's example). The fact that reasoning is involved - that is to say, that the system decides what to change or improve based on some logic, rather than arbitrarily - is the reason why these loops are considered creative and useful.

Amabile's (2012) Componential Theory breaks the creative process into five steps: Task Identification, Preparation, Response Generation, Response Validation, and Communication. Some or all steps can be repeated if more progress is needed, and each step is affected by the personality and environment of the creator.

Systems based on these loops are common. A genetic algorithm naturally switches between generation of artifacts and evaluation of the current population, and these algorithms are popular (Galanter, 2012). The ER model has been used to generate stories, do geometry problems, and model early sensorimotor development. A few researchers (Gervás, 2013a) incorporate more critical evaluation, attempting to detect specific mistakes in their work and correct them. This type of evaluation, while difficult to implement,

is more sophisticated than the random recombination of a genetic algorithm, and we recommend its use when possible; we use it ourselves in one of the versions of our TwitSong system (Chapter 5.3).

We are unaware of any systems that explicitly use Wallas' (1926), Sadler's (2015), or Amabile's (2012) multi-stage models. The results of a study replicating these methods in a computer system might be enlightening.

2.4.3 The process of professional artists

These theories are supplemented by naturalistic studies of artists at work. Fayena-Tawil *et al.* (2011) summarize studies of artists and nonartists asked to create a drawing in the laboratory. Compared to nonartists, artists spend more time examining, selecting, and rejecting available objects; reworking their drawing; developing an overall composition; stating large-scale goals; and making large-scale evaluations. Nonartists spend most of their time trying to reproduce visual details.

It should be noted that the creativity of the artists in this and similar studies is not measured; rather, it is assumed that working artists are creative. By the Four Cs model described in Section 2.5.1.1, working artists by definition are Pro-C creative, but not necessarily Big-C or H-creative. From most theoretical perspectives, there is no problem with this; however, from the perspective of some Person-based historiometric theories, which restrict themselves only to historically eminent creative people (Section 2.3), these results may or may not be relevant. Nor do these studies address the question of whether some working artists are more creative than others, and whether this difference is characterized by differences in process.

Mace and Ward (2002) study self-reports by artists making art for a gallery. They divide the process into four stages: Conception, Development, Creation, and Finishing. In Conception, the artist gets an idea, either spontaneously, or by expanding on previous ideas. In Development, the artist does preliminary sketches, and enriches the concept through further associations, until they have a specific sense of what the work will look like. They then, in Creation, physically construct the artwork itself. This involves a generate-evaluate loop, with frequent restructuring in response to mistakes or new ideas. Finally, in Finishing, the artist prepares the art for display by framing, mounting, *etc.* Each stage includes the possibility of shelving the idea or returning to earlier stages.

The process may be different for different domains. Bourgeois-Bougrine *et al.* (2014) study the process of French screenwriters. For these writers they identify three stages: Impregnation, Structure/Planning, and Writing/Rewriting. A screenwriter's Impregnation differs from an artist's Conception because screenwriters do not typically have the ideas for their own films: instead, a director hires them to create a screenplay on a specific topic. The Impregnation phase also includes rest or seemingly unrelated tasks. During this time, the screenwriter does not appear to be creating anything, but is actually allowing the vital incubation stage to happen. Next, during Structure/Planning, the writer decides on a tentative structure for the film. Finally, the writer actually writes the screenplay. Scenes are constantly rewritten as the writer finishes other scenes and traces their implications. Rewriting goes on throughout the filming, so there is no Finishing stage.

In spite of their differences, one can see a parallel in the progression from idea to structure to specific creation in both studies. It appears to be typical for most creative humans to progress in this way, beginning with a simple inspiration, then planning a structure, then implementing. However, this likely

does not apply to every form of creative work. For example, improvisational performances may not involve a structuring or planning component (Kalonaris, 2018).

2.4.4 Autonomy

A creative system need not be a slavish imitation of human creativity. Researchers may instead build a system whose process emphasizes the strengths of computers (Gervás, 2010). But one must also address the weaknesses of computers. One issue here is autonomy: the ability of the machine to decide what to do for itself.

A lack of autonomy is a major criticism both of general AI and computational creativity specifically. Mumford and Ventura (2015), surveying public opinion about creative computers, found that autonomy was one of the biggest issues, and is a particular issue for skeptics. Current systems possess some autonomy—being able to add to their knowledge base, for instance—but cannot define or refine their own processes. Guckelsberger *et al.* (2017) describe a thought experiment asking a system why it made a creative decision—and then asking "why" again, recursively; all existing systems would eventually have to answer "because my programmer told me to".

Negrete-Yankelevich and Morales-Zaragoza (2014), in their Apprentice Framework, suggest moving a system through four stages, each with greater autonomy than the last: the toolkit (a set of processes to be used by human artists), the generator (a machine that can make finished or partially finished work), the apprentice (a generator which makes novel and/or valuable work at least sometimes, with some human curation), and the master (a generator which always makes finished, acceptable-to-experts work on its own). Colton (2012) describes this autonomy-based progression as "climbing the meta-mountain". One looks at the decisions humans make for the system and automates those; then looks at the decisions humans make for the revised system, and automates those; and so on, potentially *ad infinitum*.

Colton *et al.* (2014) suggest drawing diagrams of a system's process so that judges can see what the system is responsible for, what choices it makes, and what is left to humans. Current diagrams could be compared to older ones to show an increase in autonomy over time. Mogensen (2017) uses diagrams of this nature to analyze the creativity of the Voyager and Favola music systems, although his analysis does not involve judges and is not solely concerned with autonomy. Cook and Colton (2018b) go further, designing a video game creation system which demonstrates its autonomy by continuously working on multiple projects at once, choosing which project to work on at a given time, and communicating with the public about its choices (much as human artists sometimes do through blogging or tweeting about a work in progress). Communicating with the public in this way causes demonstrations of autonomy to veer over into the Press perspective, which is discussed more fully in Section 2.6.

Complete autonomy may not be readily achievable. Relevant to this discussion is Smithers (1997), who argues that the term "autonomy" in AI is misused. "Autonomy" in AI is used to describe mobility, lack of a direct controller, self-regulation in a narrow sense, or the ability to do certain things automatically. Smithers argues that this definition is too narrow: computer systems are not autonomous unless they possess self-lawmaking and self-identity. That is, autonomous systems are governed by rules that they create through interaction with their environment. Autonomy as self-lawmaking is congruent with the way the word "autonomy" is used in law, politics, medicine, and biology. If we accept Smithers' argument, then creating an autonomous creative system would require climbing many levels of Colton's meta-mountain indeed.

Some researchers have taken up this quest for high autonomy. Guckelsberger *et al.* (2017) define adaptive creativity, in which a truly creative agent must develop its goals on its own the way living organisms do. The agent must be embodied and have a precarious existence, which requires it to continuously interact with its environment, preserving itself by modifying itself or the conditions around it. (It is unclear if a non-embodied agent could have this type of precarious existence; for the sake of argument, it is assumed that embodiment is necessary.) Only values generated through these attempts at self-preservation, in Guckelsberger *et al.*'s view, can truly belong to the agent. The most successful agents exhibit behavior that is novel and valuable (see Section 2.5.1.4): to survive, they must respond flexibly to unexpected threats.

As a metric for evaluation, Guckelsberger *et al.*'s ideas have some weaknesses. As Guckelsberger *et al.* admit, the set of systems which are adaptively embodied is quite different from the set of systems which most human observers would intuitively deem creative. In fact, it is uncertain if human creativity would meet Guckelsberger *et al.*'s standards. Humans are embodied agents with a precarious existence whose creative predilections arise from natural selection pressures. But human creativity is social in nature, and often far removed from preserving viability in the moment. Given the culturally constructed nature of many human endeavors, it is unclear if a human would pass Guckelsberger *et al.*'s recursive "why?" test; most humans do at least some things, including some creative things, because another human told them to. Creative actions in humans can also be a response to physical, psychological, or social limits which have little to do with viability as such. This does not mean that Guckelsberger *et al.*'s theory is not useful for low-level creativity in embodied agents. But it does rest on strict assumptions which may not be useful for every type of creativity. Researchers uninterested in building embodied agents can still achieve interesting things by using weaker autonomy requirements.

2.4.5 Specific evaluation techniques

All of this Process research provides sound advice for system design, but not all of it is clearly applicable to evaluation. Comparing one's system's processes to theoretical processes is useful as a check by the researcher, but it is not a formal evaluation. Fortunately, several evaluation methods exist which are process-based, in that the evaluator takes into account the system's inner workings.

One such method is the FACE model (Colton et al., 2011), which distinguishes different things that a system could create: concepts, expressions of concepts, aesthetic measures, framing information, or methods for generating any of these. Colton *et al.* suggest a number of ways FACE could be used to evaluate creativity, including a "cumulative way", in which a system is more creative if it performs more types of generative acts, thereby taking control of more of its own decisions. Or, in the "comparative way", some components are considered more creative than others. A third option is the "process way", in which a system with a more creative process is deemed more creative. What these authors think of as a more creative process is not clearly defined, although in one example, they state that they prefer more autonomy. This lack of definition means the process way verges on a tautology. In fact, all the ways of evaluating using FACE are vague. Jordanous (2012b) found that the FACE model ranked musical improvisation systems in the opposite order to other evaluation methods, and its categories are not necessarily relevant to what researchers want to achieve, which calls into question FACE's utility.

Jordanous's SPECS model (2012a) evaluates systems based on fourteen factors identified through study of how humans define creativity. These include active involvement and persistence, dealing with uncertainty, domain competence, general intellect, generation of results, independence and freedom, intention

and emotional involvement, originality, progression and development, social interaction and communication, spontaneity and subconscious processing, thinking and evaluation, value, and variety, divergence, and experimentation. Many are impossible to evaluate without knowledge of a program's inner workings. (Bhattacharjya (2016) divides the SPECS criteria among all 4 Ps; Jordanous (2016c) says many criteria cross more than one perspective.)

The SPECS model leaves it up to researchers how they will use the criteria, but Jordanous provides a case study in which musicians, after some training in how to think about creativity, were given information about musical improvisation systems and examples of products the systems had produced and asked to rate each program on the fourteen criteria. The criteria are not combined into a single creativity score, but instead provide a nuanced picture of what the system does and doesn't do well (Jordanous, 2012b).

A more hierarchical model is Ventura's (2016), which categorizes creative systems into seven categories, each more creative than the last. Randomization (in which output is completely random) and Plagiarization (in which output is copied from existing artifacts) are the least creative. In Memorization, the system modifies existing artifacts, and in Generalization, it creates new artifacts based on rules that can be programmed in or discovered by the system. In Filtration, the system can evaluate its output with a fitness function; any system using a generation-evaluation loop is at least at the Filtration level. Even more advanced, in Inception, the system uses a knowledge base to inject deeper meaning into its artifacts. Finally, in Creation, the system has perceptual abilities and is able to create artifacts based on what it sees, hears, and experiences.

Ventura's model is not an evaluation technique, but a quick-and-dirty evaluation could be performed by asking an expert to place a system in one of the seven categories. The advantage of this model is that, being an ordinal scale of levels of creativity, it is very simple to use to argue that one system is more creative than another. The disadvantage is that it lacks nuance. If two systems are at the Filtration level, for instance, Ventura's model says nothing useful about the distinction between them. Ventura's model also cannot show that filtration (or any other process) is done well.

As we saw in Section 2.4.4, further process evaluations can be done by placing a system in the Apprentice Framework (Negrete-Yankelevich and Morales-Zaragoza, 2014), or asking experts to analyze a diagram of the system's responsibilities (Colton et al., 2014).

Ventura's model and the FACE model are developed *a priori* from theory, and SPECS is based on a linguistic analysis of what humans associate with creativity. We do not know of an evaluation technique that comes at process from a psychological perspective—that is, one that systematically compares a process to the human creative processes described in Section 2.4.2. Although a creative computer need not exactly resemble a human, such a technique would still be an extremely interesting contribution.

Care must be taken to match the theory underlying a system to the theory underlying its evaluation. For instance, if one's goal is to make a system as autonomous as possible, one may not be interested in Ventura's model, which deals only indirectly with autonomy.

Some process evaluations can, with a bit of tweaking, be used to distinguish creative from uncreative processes. For example, in a hierarchical model like Ventura's or the Apprentice Framework, a researcher could state that the boundary between creative and uncreative processes lies along one of the boundaries between levels of the model. However, the best use of process theories may not be a binary evaluation of a system as creative or not; most forms of process theory do not work that way. Instead, process-based evaluation forms a nuanced picture of what a system does and does not do for itself.

2.5 Product perspective

Rather than Person or Process (or Press), many evaluations in computational creativity focus on Product. Even a system with a very sophisticated process must at some point produce a creative artifact. Intuitively, if a system is creative, we would expect it to produce some artifact that is somehow "good", interesting, or otherwise demonstrative of the system's competence at creation. From some perspectives this is less important; for example, a cognitive scientist trying to simulate the creativity of everyday humans will be less interested in the quality of the product, and more interested in whether the system's processes are psychologically sound (Pérez y Pérez, 2018). But for researchers with a more engineering-oriented approach, and for the public at large, product is undeniably important. We can evaluate the product itself as creative or uncreative by assessing its traits.

Note that, by "artifact", we do not only mean artworks: a mathematical theorem, scientific hypothesis, business plan, or engineering design is an artifact just as poems, visual artworks, and pieces of music are. Even a way of adapting to the environment is arguably a creative artifact (Guckelsberger et al., 2017; Aguilar and Pérez y Pérez, 2014). However, it is assumed in the Product perspective that all creative systems, in some way, produce something.

The Product perspective is especially useful for systems that have the goal of producing something appealing to humans.

2.5.1 Novelty and value

A very common set of creativity criteria are novelty and value. That is to say, if and only if a product is both novel and valuable, it is considered creative. This is the leading definition of creativity among philosophers and psychologists, and has a long history (Gaut, 2010); it was introduced to computational creativity by Boden (1990). Sometimes, synonyms for novelty and value are used: "originality" and "quality" or "usefulness", for example. Value is particularly prone to being represented by various synonyms, such as "appropriate", "significant", and "adaptive" (Jordanous, 2018), though the appropriateness of "usefulness" as a synonym for value may vary based on the domain (Cropley and Cropley, 2008). Regardless, definitions of creativity are typically considered equivalent to novelty and value if they boil down to the combination of two traits, which specify that the product should be somehow new and also somehow good.

Jordanous (2016a) points out that novelty and value can be applied to each of the four Ps: a system can employ a novel process, interact with the press in novel ways, or exhibit novel personal traits related to creativity. However, we will focus here on novel and valuable products. We will look at both concepts in depth before discussing how they have been brought together in practice.

2.5.1.1 Novelty

Novelty has several definitions. Boden (1990) distinguishes between P-creativity—that which is novel because its creator has not thought of it before—and H-creativity—that which is novel because no one has thought of it before. Boden also raises the issue of "mere novelty"—products which are novel, but in a trivial or uninteresting way. Boden suggests the use of *surprise* in addition to novelty: an idea is more surprising when it requires a more fundamental change to conceptual space. Note that this definition, and others, privilege transformational over exploratory creativity! Simonton (2011) similarly deals with mere

novelty by citing U.S. patent law: a product is creative when it is novel, valuable, and non-obvious (based on pre-existing domain-specific knowledge). The use of surprise or non-obviousness as a factor independent from novelty is supported by psychological studies (Acar et al., 2017).

Along with the distinction between P-creativity and H-creativity comes the finer-grained model of the Four Cs, developed by Kaufman and Beghetto (2009). This model splits creativity hierarchically into Big-C, Pro-C, little-c, and mini-c creativities. Big-C creativity is the H-creative work of master creators who are eminent in their field. Pro-C creativity is the work of creative professionals which is not historically significant, yet is successful enough to provide for a creative career. Little-c creativity is the work of ordinary people, inventively solving problems in daily life or producing creative artifacts as a hobby. Mini-c, finally, refers to the creative work of children. Each of these forms of creativity may be evaluated in a different way. Note that some of the four Cs imply something at work in the Press perspective: Big-C creative people are identified by their fame and influence, and Pro-C creativity depends on enough people buying a creative product to provide the creator with income. Without a Press to influence and reward them, all creative adults would be little-c creative. The four Cs also depend on an implicit Person perspective, since they are used to describe a creative agent and their entire present career (at least in a specific domain) rather than a single product or single instance of process.

Sternberg (2017) defines three types of creativity in terms of their type of novelty, or 'defiance': according to Sternberg, creativity can be self-defying, other-defying, or Zeitgeist-defying. An act of Zeitgeist-defying creativity is one that fundamentally changes the unconscious assumptions of the creative person's surrounding culture (which, of course, is a Press-based measure). Zeitgeist-defying creativity is therefore a special case of H-creativity, which is similar to Big-C creativity, or to the domain-changing creativity in Csikszentmihalyi's Press-based theories, which we will discuss in Section 2.6. The three types are also not fully independent; intuitively, since self and others are subject to the Zeitgeist, to defy the Zeitgeist really requires the creative person to defy all three.

Bartel (1985) objects to defining novelty as the transformation of conceptual space. He points out that one can imagine conceptual spaces being changed in trivial, random, or uninteresting ways. Bartel discusses the distinction between the terms "unique", "different", and "original", and proposes a definition of original works as those which are an *origin*. That is, a work is original if it is the first to display some unique or different attribute which is then adopted by other works. (Again, this implies that Product-based novelty is dependent on actions taken by the Press.) Merely novel works are not copied because the first such work already exhausts its own interesting possibilities. Note that, while this definition suggests H-creativity, it is not difficult to adapt to P-creativity; artists can and do copy themselves. Dasgupta (2011) defines A-creativity (standing for *antedecent*, any H-creative product which has never been seen before) and C-creativity (standing for *consequence*, an H-creative product which influences others). While A-creativity can be assessed immediately, assessing C-creativity requires a Press-based historical perspective.

2.5.1.2 Value

The idea of value comes with questions of its own. Who decides if a work is valuable or not, and on what grounds? An obvious answer is that the wider culture, the Press, will decide. But it is easy to think of creative people (*e.g.*, van Gogh or Mendel) who were ignored in their day, and assigned great cultural value posthumously. Does this mean that they became creative posthumously? Some artists have multiple phases of greater or lesser popularity after death. It is difficult to imagine that changes in the dead artist's actual creativity are behind these shifts (Weisberg, 2015). But Csikszentmihalyi (1996) argues that a

person's creativity can and does change after their death—because Csikszentmihalyi views creativity as a Press phenomenon, residing not in the creative person but in their interactions with the wider culture, which naturally changes over time, even when the creative person is dead. Dorin and Korb (2012) also bring up the example of cultural biases. Women's art has historically been overlooked due to systemic sexism. Does this mean that women's art is actually less valuable? Less creative?

Philosophers such as Gaut (2010) bring up the idea of "negative value". For example, imagine a terrorist who comes up with a new, unusual, and effective way of carrying out attacks. Is the terrorist creative? If so, who is assigning value to the terrorist's work? One solution is to deny that destructive acts are ever creative; another is to define them as valuable if they effectively serve the creator's goals. Cropley *et al.* (2008), who study this type of malevolent creativity, define acts as "subjectively benevolent" if they are beneficial to one person or group at the expense of another. In other words, a creative terrorist attack is assigned value by the group of other terrorists sharing the same political aims, even if they are not valuable to society at large. (Fake news, creative ways of cheating on an exam, creative phishing schemes, and many other examples of malevolent creativity can be evaluated in a similar way.) "Value" is sometimes phrased as "usefulness" or "appropriateness to the task", which accords with the intuition that even negative or pointless tasks could be done creatively (Kaufman and Baer, 2012).

Weisberg (2015) argues that value should not be a criterion for creativity. In addition to artists becoming more creative after their deaths, a changeable definition of "value" means that it is difficult to track the validity of creativity research. One generation might consider Group A's work valuable, and Group B's work not valuable. A second generation might reverse these judgments, considering only Group B's work valuable. The previous generation's creativity research therefore becomes invalid, unless it can be shown that the conclusions drawn about Group A also apply to Group B. Weisberg suggests defining creativity as "intentional novelty". Harrington (2018) objects to Weisberg's argument, stating that removing value from the definition of creativity does not reflect the typical goals of creators or audiences in regards to creative work. Also, removing value does not remove the problem of cultural or historical specificity: the way in which humans judge novelty, and the areas in which humans especially value novelty, may also change over time.

Bown (2012) also argues against the use of value, citing processes like plate tectonics, which produce novel artifacts without a goal. Bown calls these processes "generative creativity". Humans arguably engage in generative creativity at times, such as when brainstorming. Psychologists also identify processes such as daydreaming, as well as ignorant mistakes, to involve novelty without value, sometimes referring to these processes with names like "pseudocreativity" or "quasicreativity" (Cropley and Cropley, 2008). For some applications, such as a system that engages in brainstorming, requiring this kind of quasicreativity judged by novelty alone may be appropriate. However, our view is that most applications do require value. A work of art or science is judged by its value when created by humans—so a computer system working in these domains should be held to the same standard.

2.5.1.3 Typicality and Ritchie's Model

Ritchie (2007) argues that typicality should be used rather than novelty. His argument is that, while novel output of some type by a computer is conceivable, the computer would first need to generate products that are actually of the given type. A novel string of incomprehensible letters would be easy to produce, but if that was all the computer produced, it would be difficult to argue that it is successfully writing poetry. Therefore, creative systems should reliably generate both typical and valuable output. In Ritchie's view,

the most useful way to measure this would be to look at statistical properties of a system's output over time, and to compare them to the properties of an inspiring set—a set of exemplars the system uses to define its task. Ritchie lists many criteria that one could test for based on these statistical properties. For example, one might wish for the average typicality to be within a certain range, or for a certain proportion of the output to be above a given threshold for value. Which of these criteria, and what ranges and threshold values, are appropriate for a given system is an unsolved question. While typicality might appear to be the opposite of novelty, Ritchie's framing of them as separate criteria is supported by experiments such as Hekkert *et al.*'s (2003), who find that novelty and typicality in design both separately contribute to human preference.

Ritchie's model has fallen out of favour in recent years (Jordanous, 2012b), but it has been used to evaluate computational systems. Jordanous (2012b) found that Ritchie's model was cumbersome to implement, because of the need to give ratings to a large number of outputs. Its results are abstract and need to be rephrased to be useful to the researcher. Pereira *et al.* (2005) report that appropriate threshold values are usually unknown, and the criteria involving threshold values are very sensitive to changes in these values. Poorly chosen threshold values can invalidate other criteria by, for example, creating a scenario in which there are no valuable but atypical items. Defining typicality is challenging in practice, as it can be based either on content or structure; it is unclear what to do with, for example, an unoriginal artifact with structural errors.

Ritchie (2007) is ambivalent about attempts to use his criteria, stating that many of them are implemented incorrectly, particularly in the inspiring set. Many studies either choose an inappropriate inspiring set, or invalidate many composite criteria by basing typicality scores on similarity to the inspiring set, so that the typicality of generated work and the typicality of the inspiring set cannot be compared. RASTER, a thought experiment meant to demonstrate the insufficiency of Ritchie's criteria, similarly renders the criteria involving an inspiring set inapplicable (Ventura, 2008). Pereira *et al.* (2005) found that this invalidation occurs when there are no atypical items in the inspiring set. RASTER also defines typicality and quality identically, which invalidates even more criteria (Ventura, 2008).

When not partially invalidated in this way, there may still be some value to Ritchie's model, and to the idea of typicality as a prerequisite to meaningful creativity. Other evaluations, often by authors who do not use the word "typicality" nor Ritchie's mathematical criteria, implicitly contain the idea of typicality as necessary for creativity. For example, some music systems are evaluated by their adherence to music theory (Sturm and Ben-Tal, 2017; Lattner et al., 2018; Opolka et al., 2015; Scirea et al., 2015) or by the accuracy of machine learning at classifying them as a part of the appropriate genre (Videira et al., 2017). Certainly, it is intuitive that a creative system's output should, in some way, be appropriate to its intended domain.

2.5.1.4 Implementing novelty and value

If novelty and value are used to evaluate a system, then a specific procedure for their evaluation is desirable. It is especially desirable if novelty and value can be calculated by the computer itself—for use in the evaluation portion of a generate-evaluate loop (see Section 2.4.2)—although a computer system should not be judged only by its own self-assessment.

Pease *et al.*(2001) describe a number of possible methods for calculating novelty and value. For novelty, this includes difference relative to an inspiring set, level of transformation of conceptual space, relative

complexity, membership in a fuzzy set based on an archetype, Bayesian improbability, and novelty perceived subjectively by humans. For value, it includes the emotional response of humans, or the degree to which the product achieves the creator's goal.

França *et al.* (2016) define novelty as Bayesian surprise, and value as *synergy*—a metric based on expert judgments of the value of pairs of components. Bhattacharjya (2016) defines quality in terms of a preference model which mathematically combines different aspects of a subjective judgment.

Elgammal and Saleh (2015) measure novelty in art history by creating a directed graph of influences based on semantic and perceptual traits—recalling Bartel's definition of originality as the creation of an origin—and found good correspondence between their system's behaviour and the opinions of art historians. Shrivastava *et al.* (2017), measure "influence" in film in a similar way, creating their graph based on features learned through applying word2vec and Principal Component Analysis to plot summaries and other textual information about the film. Shrivastava *et al.* do not equate novelty with influence; instead, their scoring system combines "novelty" (the presence of traits that haven't been seen before), "influence," and "unexpectedness" (the presence of traits that, based on a predictive model, were not likely to occur). Note that unexpectedness, in this definition, is distinct from novelty; a trait that has been used before can be used unexpectedly by using it in a time or context in which it is not typical. By combining novelty, influence, and unexpectedness, Shrivastava *et al.* generate ratings for films that correlate well with expert film ratings on Rotten Tomatoes and IMDB.

Novelty and value can also be based directly on human opinion. Some researchers in this vein (Llano et al., 2014; Riedl and Young, 2006; Veale, 2015) have found that human assessments of novelty or surprise are negatively correlated with most other desired criteria. Indeed, in Llano *et al.*'s study, random statements are judged more surprising than either human or computational outputs, yet do poorly on every other criterion. Such "mere novelty" suggests that Ritchie was correct to require typicality as a prerequisite for novelty, but they are inconsistently reproduced. Some researchers supplement novelty and value with *surprise*—Boden's (1990) word for meaningful, significant novelty.

Mckeown and Jordanous (2018) analyze the relationship between novelty, randomness, and creativity in narrative generation. They find that relaxing constraints leads to more novel output, which is rated by humans as more creative, up to a point—but that, if constraints are relaxed completely (to the point of violating human knowledge about how the world works), the resulting narratives are judged as much less creative.

2.5.1.5 Combining novelty and value

It is also worth asking whether novelty and value, when used together, are of equal importance, and how they should be combined. In Section 2.5.1.2 we have already seen theoretical justifications for why some researchers dispense with value and use novelty alone. Other researchers privilege novelty over value, while still requiring at least minimal value to exist. Smith and Smith (2017), for instance, suggest a "1.5 criterion" model: a creative product must be definitely novel, but it need not be definitely valuable, only *potentially* valuable. That is to say, a creative idea needs to make sense, and novel ideas that are obviously wrong are not creative; but there is no need to fully implement and test an idea, to see if it "really" works in a valuable way in practice, before deeming it a creative idea. This, of course, implies a Press perspective in which some group of humans or other agents use their intuition to judge an idea's potential value.

Acar *et al.*'s (2017) study suggests that novelty is more important; ratings of novelty explain far more of the variance in ratings of creativity than ratings of value, aesthetics, or surprise. On divergent thinking tasks, novelty and value often negatively correlate. Diedrich *et al.* (2015) found that, on such tasks, human judges' ratings of creativity correlated with novelty but not usefulness. Of course, this result may be specific to divergent thinking tasks, in which generating novel ideas without much evaluation of their usefulness is the whole point. Smith and Smith's 1.5 criterion model (2017), as well as Acar *et al.*'s, also draw largely from the results of divergent thinking tasks.

On the other hand, it appears to be common for some researchers to evaluate their systems *only* by the value of the output and not by its novelty. Jordanous (2018), surveying recent computational creativity papers, finds that evaluating in this way is more common than using both novelty and value, yet that researchers often still refer to themselves as evaluating the system's creativity. Unless the researchers have a strong theoretical justification for this, it is to be avoided.

2.5.2 Other criteria

Besides simple novelty and value, other product-based sets of criteria exist. Cropley and Cropley (2005) define a hierarchical model specifically for "functional creativity"—the kind used in engineering. Their criteria are Effectiveness (the product, at a basic level, solves the problem it was intended to solve), Novelty, Elegance (the product is pleasing, or goes above and beyond mere correctness, such as by being more cost-effective than previous solutions), and Generalizeability (the solution can easily be applied to additional problems). Cropley and Cropley's model requires previous criteria in their list to be met before the next becomes relevant. Without novelty, an engineering product is not creative; but if a product is not effective, its novelty does not matter to an engineer. The remaining criteria of elegance and generalizeability define higher levels of functional creativity. Cropley and Cropley later expand their model into one that they claim is domain general, changing Generalizeability to Genesis and adding various sub-criteria. Several similar models, starting with novelty and value and adding other criteria, also exist in the psychological literature (Cropley and Cropley, 2008).

Domain general sets of criteria without novelty and value have been proposed. A popular set is Imagination, Appreciation, and Skill, the Creative Tripod (Colton, 2008). However, subtleties in the thinking behind the Tripod mean it fits best with the Press perspective, and will be discussed in Section 2.6.1. Lehman and Stanley (2012) evaluate products by their Impressiveness: products are impressive when they are easy to appreciate, but difficult to recreate.

Some Product-based criteria for creative systems are entirely domain-specific, such as Potash *et al.*'s (2016) proposed criteria for ghostwriting song lyrics: stylistic similarity to a source text (which is arguably a form of Typicality), combined with novel content; or Gervás's evaluation of plot generation systems by their adherence to Vladimir Propp's Morphology of the Folktale (Gervás, 2017). Criteria can also be specific to a generation method, such as cross-entropy and other statistical criteria that are used to test the output of many neural networks (Sturm and Ben-Tal, 2017; Wang et al., 2016a; Yan, 2016; Zhang and Lapata, 2014). Care should be taken with such methods, as some can fall afoul of the problem Jordanous (2018) points out, measuring only the output's quality and not its creativity as such. Of course, that problem can be avoided if both a domain-specific quality criterion and a measure of novelty are used.

An extremely common technique is for researchers to use *ad hoc* domain-specific product criteria. The rationale for domain-specific criteria will be discussed in Section 2.7, and some examples given in Section

2.8.1. Unlike the domain-specific criteria discussed above, *ad hoc* criteria are not clearly grounded in a previously developed theory or in experimental evidence. This means that *ad hoc* criteria are not ideal from a validity perspective, but the intuition behind them is sound. A given system may have the goal of producing work with certain traits, such as an elegant theorem, a poem fitting a given form, or a humorous Internet meme. Criteria testing if these goals have been met can easily be added to a more general test of creativity, or if using novelty and value, criteria like these might be used to test value.

2.5.3 The modified Turing test

A simple product evaluation without criteria is the modified Turing test: human subjects are given human-created and computer-created products, and are challenged to figure out which is which. If they cannot do it, then the computer system is claimed to be creative. It is important to note that this is a *modified* Turing test. The original Turing test, for general intelligence, involves judges who can question the system however they like, while the modified test involves only static, finished products. Thus, it could not be said that a system which passed such an evaluation was intelligent, only that its products were indistinguishable from human products. However, the modified Turing test is frequently used in computational creativity (Schwartz and Laird, 2015; Elgammal et al., 2017; Pearce and Wiggins, 2001; Agirrezabal et al., 2013; Loller-Andersen and Gambäck, 2018). Some researchers also refer to their evaluations as modified Turing tests when they are not strictly a test of distinguishability: for example, when both human and computer products are rated on their enjoyability by an audience that doesn't know which ones are computer products (Videira et al., 2017). Or these tests can be combined with a distinguishability test (Collins and Laney, 2017; Pearce and Wiggins, 2007; Rashel and Manurung, 2014a).

Pease and Colton (2011) criticize the modified Turing test on several grounds. First, the test encourages pastiche and superficial imitation of human work. Second, the lack of interaction between judge and creator calls into question the modified test's validity. Third, the test limits access to framing information, such as how and why the product was created, which Pease and Colton consider an important part of evaluation. Finally, according to Pease and Colton, no current system can pass a modified Turing test, so evaluations are needed which can identify the strengths and weaknesses of systems even if they are only partially creative.

Other authors have also criticized the modified Turing test. Sturm and Ben-Tal (2017) say that the idea of computers trying to fool humans "trivializes" computational creativity. Fooling other people into mistaking them for a "real" artist is not generally the goal of creative humans. Additionally, it is not at all clear how to choose the judges for a modified Turing test, the human works against which the system's products are tested, or the sample of the system's products which are included in the test. Small changes in any of these parameters could drastically affect the test's results. Similarly, Clements (2016) argues that some of the most interesting computer-generated creative products are recognizable as a computer's products, using a computer's strengths to create something in a computer-like way which is still nonetheless valuable to humans. Such products would be creative and desirable, but would fail a Turing test.

We have some quibbles with some of these arguments: for instance, pastiche may be the goal of some systems. However, we acknowledge that a modified Turing test is not appropriate unless the system is meant to create work which closely resembles existing work by humans.

2.5.4 Consensual assessment

Another criteria-free Product evaluation has a much stronger pedigree. This is the Consensual Assessment Technique (CAT) developed by Amabile (1983b). The idea in consensual assessment is that accurate judgment of creativity is not formulaic but relies on the expertise of people who are recognized as creative in a domain. In a consensual assessment, human subjects create something—for example, a collage. The subjects should be matched for their level of experience with the chosen medium. A team of experts in the relevant field then looks at the products and each decides how creative each product is. An important point is that the experts are *not* given criteria such as novelty and value, or other instructions in how to make their judgments, except that they must use the full scale. They also must make the judgments on their own, rather than consulting with each other. If most of the experts agree (that is, if they have good interrater reliability), the consensual assessment is successful.

Even though the judges may be making different kinds of assessment, the CAT has good interrater reliability in practice as long as the expertise of the judges is high enough (Kaufman and Baer, 2012). However, finding judges with such expertise can be difficult. Our own experiments with the CAT, described in Section 4.2, suggest that for computational creativity there are two major roadblocks to administering the CAT. A highly experimental and diverse domain such as computational art (or in our case, poetry) may not have a stable group of experts to evaluate it, and products made by wildly different computational processes may be difficult even for experts to compare. However, in stable and well-defined categories of computational product, such as Wiggins' (2006) experiments with chorale melodies, the CAT remains a sound methodology.

The use of experts regardless of their preferred method of judgment accords with how creative artifacts are evaluated in the real world, by critics, editors, peer reviewers, gallery owners, prize committees, *etc.* (Kaufman and Baer, 2012)—though, of course, jurors, critics, and editors do talk to each other in the process of judging. Like the modified Turing test, a computer cannot use the CAT to evaluate its own work, since it requires a group of expert assessors.

2.5.5 Product evaluation, mk. II: Computational aesthetics

In the arts, especially the visual arts, many researchers have attempted to create domain-specific criteria with a deeper perceptual grounding. While being aesthetically pleasing may not be equivalent to being creative, it is still useful to use well-grounded aesthetic measurements as part of a judgment of value or similar components of creativity. This leads us to the field of computational and psychological aesthetics.

Galanter (2012) summarizes computational aesthetic research in visual art. Many of the aesthetic rules taught to humans in art schools, such as color theory, are difficult to encode in a way that matches human perception. However, many perceptual rules exist which are mathematical in nature, and can therefore be calculated. These include complexity, Zipf's law, JPEG compressibility, and prototypicality. Many of these do correlate at least somewhat with human preferences. Alternatively, a neural net can be trained to evaluate art without explicit rules. This need not be a binary classification. Systems such as DARCI learn to recognize many subjective qualities in artwork, such as "happy", "fiery", or "lonely", based on humans' descriptions (Norton et al., 2010).

How humans process visual art is a challenging field of study for human psychologists. Human responses to art can be highly individual (Dubnov et al., 2016; Juslin et al., 2016) and affected by context (Leder

and Nadal, 2014). Art processing involves the entire brain, rather than a specific art processing region (Leder and Nadal, 2014). Still, psychologists are at work on theories outlining how this processing occurs.

One of the theories with the most explanatory power is the theory of representational fit (Sammartino and Palmer, 2012). In this theory, images are preferred when they transparently reflect the work's intended meaning. In simple cases, this means humans will prefer images that are easy to process; but representational fit also explains complex and difficult artworks, such as ambiguous photographs. If ambiguity is the best way to make a point, and the audience understands this point, then they will appreciate the ambiguous artwork.

Alternately, Silvia (2005) suggests an emotional appraisal theory of aesthetics. If a human appraises an art object as novel and not understood, but believes themself capable of understanding it, they will take interest in the art. Appraisal theory explains why novices prefer art that is easier to process, but also why art experts prefer more difficult work.

Leder *et al.* (2004) combine these and other theories into a model of the stages of aesthetic judgment. In Leder *et al.*'s model, the ultimate goal of looking at art is "cognitive mastery", in which the audience feels they have figured out what the art means and what they think of it; different people will seek to achieve this mastery in different ways.

Art appraisal is also influenced by context. An artwork's price, title, setting, the artist's name, reported approval or disapproval of the artwork by other social groups, or even the viewer's current physiological state can influence human processing of art (Lauring et al., 2016). Therefore, a full accounting of human processing of visual art would also have to include contextual and bodily effects as well as social factors.

Similar work is available for other domains, often indicating that different domains have different aesthetic rules. Augustin *et al.* (2012), for example, find that humans value different qualities in visual art, film, and music. Jacobs (2015) identifies two distinct aesthetic processes that occur in literary writing– a "background" process which creates suspense or empathy through the situations that are described, and a "foreground" process which creates aesthetic appeal through the style in which they are described. To facilitate this type of research across different domains, Blijlevens *et al.* (2017) have developed a reliability and validity tested scale for measuring aesthetic pleasure without making assumptions about its cause.

None of these theories can yet be encoded precisely enough for a computer. However, a researcher without intensive domain training might do well to refer to these or other theories of aesthetics when setting qualitative goals for the characteristics of their system's output.

2.6 Press perspective

After a person uses a process to create a product, the Press—other people—then receive it. The Press perspective studies how this reception occurs, and what kind of social effect a product needs to have to be called creative. To some extent, a Press perspective is necessary to every other perspective: the act of judging people, processes, or products is always done by people in a cultural context. (Or, if it is done by a computer, that computer was programmed in a cultural context by people.) The pure Press perspective is especially useful for researchers who want their system to make an impact on society, influence human opinion, and be recognized for its work.

In Rhodes' (1961) formulation, Press refers not only to "the press", as in cultural agents responding to a creative work, but to the general environment "pressing in" on the creative person. This includes social

responses to creative products and the social causes of their creation. The social environment influences the values and beliefs of a creative person, determines what forms of education and training are available, and can direct the form of creativity through incentives and commissions (Csikszentmihalyi, 1996). However, research from the press perspective has increasingly focused on responses and judgments.

Many scholars argue that creativity cannot be studied independently of its context. There are weak and strong versions of this claim. The weak version, advanced by Boden (1990), is that certain criteria are contextual. For example, H-creativity is contextual since it depends on what other people did and didn't do before, and value is subjective because it depends on what people find valuable. Nevertheless, we could objectively study value by, for example, making a computer model of the values of a certain group and testing how well new works fit the model.

The strong version, advocated by Csikszentmihalyi (1999)—also known as a systems perspective—situates creativity entirely outside of person, process, and product. Since all judgments are subjective, in the strong Press view, it is epistemically and perhaps ontologically impossible to separate creativity from people's judgments of creativity. Therefore, creativity in practice is not separate from the people judging it. It is situated, not in the creator, but in an interaction between creator and audience. Even aspects such as skill and the ability to self-evaluate are in some sense internalizations of domain knowledge and field attitudes, which did not originate with that person (Csikszentmihalyi, 1996).

Csikszentmihalyi breaks down some factors involved in Press-based creativity. These include the individual, domain, and field. The domain is the cultural and symbolic aspect of creativity: for example, what works have been produced in this genre, and what are its conventions and rules. The field is the social aspect: fellow creators, editors, curators, and critics who serve as gatekeepers. The word "genre" is used here not only to denote artistic genres, but also branches of science, business, activism, or any other creative activity. Even abstract tests such as the Torrance Tests have a domain (questions about chairs, *etc.*) and field (the scientists designing and scoring the tests). To be successful, in Csikszentmihalyi's view, a creative individual must change the way individuals in the field think, feel, or act (Csikszentmihalyi, 1999). This means that Csikszentmihalyi is primarily interested in big-C, H-creativity, not in creativity's everyday forms (Csikszentmihalyi, 1996).

As mentioned in Section 2.5.1.2, a field is not necessarily free from bias. Racism, sexism, and other prejudices on the part of the field could lead to creative work from certain groups being systemically underrepresented. Thus, such individuals could be "less creative" from the Press perspective, even if their Products, Processes, and Persons are as good as (or better than!) those of white men. For researchers such as Dorin and Korb (2012), this is a reason not to uncritically use the strong Press definition of creativity. However, if the strong Press definition is taken literally, then to call marginalized creators "less creative" is not a value judgment, but a factual description of what occurs when they attempt to interact with a prejudiced field.

The question of who performs the evaluation—of who constitutes the field—leads into the cross-cultural study of creativity. A review of cross-cultural studies by Lubart (1999) finds that not all human cultures agree on what constitutes creativity. He states that Westerners focus on originality and the production of something new, while Eastern cultures focus on the expression of insight, growth, and inner truth. He also describes differences in where creativity is thought to be situated. In Bali, for example, musical groups can be distinct from one another, but individual musicians within the groups may not. Therefore creativity occurs here only on the group level. Blijlevens *et al.* (2017) find that parts of their model of aesthetics are weighted differently by participants in different countries, but the cause of this is not clear.

For computational creativity, Press evaluation is a challenge since most people are not used to seeing computers as creative. Some researchers study how exactly these people respond to creative artifacts made by a computer, or on effective methods for convincing them that the computer's work is worth consideration. These issues of perception will be discussed more fully in Section 2.8.2.

2.6.1 The Creative Tripod

Some researchers design systems with an explicit persuasive element, with the goal of convincing humans that the system is creative. Colton (2008) leads this trend. His Creative Tripod appears to be Person-focused at first: it presents several traits that a creative system should have, namely Skill, Imagination, and Appreciation. Colton's assertion is not that a creative system must possess these qualities, but that a creative system must *appear to possess* these qualities. Much of Colton's research involves making machines more persuasive in convincing an audience that they have these creative qualities. This includes work on the framing of artifacts, taking advantage of contextual effects that can sway human judgment (Charnley et al., 2012).

Certain objections to the Creative Tripod have been raised. Colton *et al.* found that, to overcome skeptics' objections to computational creativity, the original three criteria were not enough. They added the new criteria of Learning, Intentionality, Accountability, Innovation, Subjectivity, and Reflection (Colton et al., 2014). Not much justification is given for these specific additions, except a statement that they addressed the most common objections to characterizing a system as creative. The same research team later added the much better-defined criterion of *authenticity*: to be authentic, a creative system must not be perceived as misleading its audience or as trying to write about matters that are too far outside its experience (Colton et al., 2018).

Bown (2014) objects to the Tripod because the terms are not given clear definitions; therefore, they "cannot be distinguished from trivial pseudo-versions of themselves." A system can be argued to possess imagination, for example, if the programmer put something into it which the programmer believes is related to imagination—which bypasses any falsifiable inquiry into whether this actually constitutes imagination. Many papers perform exactly this kind of trivial evaluation, and even obviously uncreative systems can appear to pass the Tripod with the right argument. For example, Ventura's thought experiment, RASTER, generates the pixels of images at random, and outputs the images if a similar image can be found online. Ventura (2008) describes RASTER as meeting Tripod criteria: *imagination* because it engages in random search, *appreciation* because it uses a (simplistic) fitness function, and *skill* merely because it produces images. (Incidentally, we have found that researchers using the FACE model, which was developed by the same research team, also tend to make trivial or rhetorical arguments as to why their systems satisfy the model, e.g. (Colton et al., 2012; Misztal and Indurkhya, 2014; Oliveira and Alves, 2016))

Smith, Huntze, and Ventura (2014) give their own working definitions to the Tripod. Skill is the ability to produce something useful; imagination is the ability to search the conceptual space and produce something novel; appreciation is the ability to self-assess and produce something of worth. (Note the implied links between the Tripod and a Product perspective, as well as Boden's work.)

Jordanous (2016b) identifies Colton's tripod with three aspects of the SPECS model: Skill to *domain competence*, Imagination to *variety, divergence, and experimentation*, and Appreciation to *thinking and evaluation*. Jordanous (2012b) argues that evaluating work based on the tripod requires process information about the system's behavior over time. Product, Person, and Process concepts frequently creep into

Press evaluation, because the humans reacting to a creative system are assumed to be using these concepts themselves.

There is nothing wrong with the Creative Tripod criteria when they are defined in this way. However, inappropriate uses of the Tripod serve as a cautionary tale: for any criteria, Press or otherwise, if we have not specified just what we mean by each of our criteria, our evaluation becomes meaningless.

2.6.2 Impact on the domain and field

Given Csikszentmihalyi's (1999) contention that creativity must change the domain or field, one might reasonably ask how to measure such a change. As with Process creativity, several theories exist which place creative achievements into rough categories based on their type or amount of impact on their domain.

Some of these theories were discussed in Section 2.5.1.1, as they can be conceptualized as either Press or Product measures. H-novelty is inherently a Press measure in some sense, as it depends on what has happened in the domain previously. So, for example, Sternberg's (2017) classification of types of creative defiance is both a Product-based theory of novelty and a Press-based theory of divergence between a product and its existing domain. Similarly, Bartel's (1985) definition of originality as an origin is heavily entwined both with a Product view of novelty and a Press view of impact on the domain. If Bartel's theory is accepted, then measurements like Elgammal and Saleh's (2015) in visual art, or Shrivastava's (2017) in film, can be made to precisely determine the amount of influence a given work has on subsequent works in its domain. However, making these measurements requires a historical view, as it will take time for works influenced by a current work to appear.

Sternberg *et al.* (2001) also have developed the propulsive theory of creativity, in which a domain and field are construed as moving in one direction, and art is placed into one of eight categories based on how it relates to that direction. For example, art can move forward by introducing new elements and ideas, or stay in place in order to develop the most excellent examples of the current state of the art, or it can radically reimagine where the domain is supposed to be. Sternberg *et al.* note that one of these categories is not necessarily more creative than another (recall Policastro and Gardner's similarly value-neutral categorization in Section 2.3). Some are, on average, more novel than others, but the category system says nothing about the work's value. Value is, instead, to be determined by cultural impact and reception, though for some forms of creativity, when an artist is ahead of their time, that reception will be delayed.

Some further attempts at measuring impact on a domain are discussed in Section 2.6.7.

2.6.3 Measures of audience impact

Instead of measuring a work's long-term impact on its domain, one can also try to measure its immediate impact on an audience. Several researchers are interested in measuring audience impact. These include Colton *et al.* (2011), who describe the interaction between a creator's work and an idealized audience with the IDEA model: Iterative Development, Execution, Appreciation. Colton *et al.* describe development moving through stages based on how novel it is, from completely derivative work to humanlike work to work so novel that humans cannot comprehend it. Audience impact at any stage is measured with two variables: *wellbeing* (how much the audience likes the work) and *cognitive effort* (how prepared the audience is to spend time trying to understand it).

Like Ritchie's criteria, these variables can be combined in a variety of ways. For example, a work with a high standard deviation in wellbeing would be considered "divisive". IDEA is a descriptive model, and it is up to the researcher to decide which of the possible adjectives applied to their system are desirable.

Burns' (2015), EVE' model measures the mental processes of the audience in another way, defining creativity as surprise with meaning. In the EVE' model, which has been applied to jokes, simple visual art, advertisements, and poetry, an expectation (E) is set up, then violated (V). The violation is accompanied by a new explanation (E') which accounts for the unexpected events. This explanation may be overt, or may happen implicitly as the audience retrieves contextual information from long-term memory (Dubnov et al., 2016). If the audience is surprised by a work, *and* can make meaningful sense of it, they approve. In experiments, ratings of surprise multiplied by ratings of meaning accounted for 70% of the variability in ratings of creativity (Burns, 2015).

There is some theoretical support for the EVE' model. As Dubnov, Burns, and Kiyoki (2016) point out, it is compatible with the appraisal theory discussed in Section 2.5.5. A surprising stimulus is appraised by an audience as novel and not understood, and the subsequent explanation causes the audience to appraise themselves as able to understand it. According to Jacobs (2015), surprise followed by explanation is what constitutes foregrounding in literature. In a longer work, oscillation between surprise and explanation is constant. However, it remains to be seen experimentally if the EVE' model can be applied straightforwardly to creative works with longer, more complex content.

Depending on the domain, any of these methods may be appropriate for testing if a creative system is having the desired effect on its audience.

2.6.4 Interactive art

Bown (2014) suggests that computational creativity researchers should perform evaluation through the lens of interaction design. The designers of a creative system must consider how the audience will interact with their system and what effect they wish it to have. Fortunately, there is already a great deal of research on creative interaction design in the domain of interactive art, studied by human-computer interaction researchers, museologists, and others.

Candy and Bilda (2009) describe three types of audience engagement: immediate (catching attention), sustained (attending to the art for a period of time), and creative (having a lasting effect that somehow changes the audience). Similarly, Bollo and Dal Pollozo's (2005) model of museum exhibits involves variables such as Attraction (the percentage of visitors who look at an exhibit) and Holding Power (the amount of time an average visitor spends looking at the exhibit). Edmonds *et al.* (2006) relate these traits to parts of a display. Attractors increase attraction, Sustainers increase holding power, and Relators encourage the visitor to keep thinking about the artwork and return later.

HCI researchers also make use of qualitative, descriptive methods in assessing audience response. Her *et al.* (2014) review many of these methods.

All of this work is helpful for researchers crafting an interactive system. It is more difficult to apply theories from HCI to a system that creates something static, such as a painting or poem. It is even more difficult to apply them to a system that performs scientific, mathematical, or some other form of creativity in which the goal is to soberly present a theory that experts recognize as meaningful. However, if interaction is one of a researcher's goals, then care should be taken to consult the literature on interaction which already exists.

2.6.5 Creativity support tools

Another application of interaction design is to co-creativity or creativity support tools. Rather than producing an artifact by themselves, these tools make it easier for a human to be creative. It is possible to evaluate a creativity support tool using any of the four perspectives, but the method that usually makes sense is to evaluate the quality of the user's interaction with the tool. This is Press evaluation, not necessarily in the sense of a tool becoming culturally successful, but in the sense of evaluating the tool by evaluating the interactions and impacts that it has on its users.

Evaluation of creativity support tools is well studied, and includes empirically validated rating scales measuring the extent to which a tool supports a particular creative goal (Carroll and Latulipe, 2009). Apart from these scales, evaluation of co-creativity tools can focus on usability, enjoyability (Kantosalo et al., 2015; Waller et al., 2009), usefulness to creative professionals (DiPaola et al., 2013; Kantosalo et al., 2015), or the quality of output created using the system (Lee et al., 2016; Shibata and Hori, 2002; Waller et al., 2009). Kantosalo and Toivonen (2016) also classify co-creativity systems as *alternating* (where the human and computer take turns modifying an artifact) or *task-divided* (where the human and computer are responsible for different subtasks). A researcher making a co-creative system should generally use existing and well-validated scales to evaluate such work.

2.6.6 Artificial social systems

In Section 2.4.2 we mentioned Glăveanu's (2015) comment that the evaluation phase of the human creative process is based in perspective-taking. Now that we have seen Csikszentmihalyi's theory, we can be clearer about the humans whose perspectives are important: they constitute the field. Several researchers have asked what happens if creative systems serve as each other's field. They have created multi-agent systems, either of robots or of software modules, which influence and learn from each other's work (Hantula and Linkola, 2018; Kirke and Miranda, 2013; Linkola et al., 2016; Saunders et al., 2010). Corneli *et al.* (2015) imagine how a computer could go through a writing workshop, in which drafts are critiqued by other computers trained on similar tasks. Systems also exist which simulate groups of musical performers improvising together (Eigenfeldt et al., 2017; Puerto and Thue, 2017). The precise method for evaluating the interactions between systems in such groups remains unclear.

2.6.7 Cultural success

A final method for press evaluation is to publish the creative product as a human would publish theirs. In visual art, this means submitting the work to an art exhibition or gallery. It is not uncommon for HCI researchers creating interactive art to do exactly this. The artwork is judged a success based on the opinion of the audience, gallery admissions, statements of interest by curators, or whether the artist was invited to submit work to further exhibitions (DiPaola et al., 2013; Sheridan et al., 2005; Tresset and Deussen, 2014). Researchers in creative music systems, similarly, can evaluate their work by performing it as a concert (Sturm and Ben-Tal, 2017).

An artwork's success with these gatekeepers can be a useful definition of press-based creativity; it naturalistically reflects the metrics most working human artists use to judge their own success. Nor is this a method restricted to art: domain-appropriate methods may exist for other domains. Scientific

advancements are evaluated through peer review, publication, and citation; entertainment for a general audience is evaluated by its commercial success; Internet memes are evaluated by how often they are shared.

Gatekeepers accepting interactive art for an exhibition generally know the role of the computer in the art, but it has happened that computer-generated works have achieved cultural acceptance from gatekeepers who do not know that they were made by computers. For example, computer-generated poems have been accepted for publication by magazine editors who believed they were by a human (Clements, 2016). Fooling a gatekeeper in this way is sometimes referred to as "passing the Turing test", though this is imprecise: it has even less to do with the methodology of the original Turing test than the modified Turing tests mentioned in Section 2.5.3. No experimental constraints are in place to test how often and to what extent the systems in these cases are capable of fooling people, and, as explored in Section 2.5.3, fooling people may not be the most appropriate creative goal. Still, being accepted by a gatekeeper in this manner is a measure of at least some demonstrable Press-based success.

Jordanous has quantified some of these forms of success, calculating a musician's cultural value by the number of comments they receive on a digital music site (2015) and a computational creativity researcher's impact by the number of non-self citations in the 5 years following publication (2016b). Interestingly, the number of non-self citations of a system "roughly aligns" with expert judgments of the system on metrics like Ritchie's model or the Creative Tripod—but not with the judgments of experts who were asked "how creative is this?" Vartanian *et al.* (2017) measured architects' creativity based on a poll of the popularity of buildings they had created, and found that these measurements correlate well with expert assessments of the same architects' creativity that were made 50 years earlier. To some extent, these correlations may reflect a self-fulfilling process: if a creative product is accepted by experts though the measures that experts are used to, then it will be cited and promoted by more experts and its success will be more likely to eventually catch on with the general public.

Cultural success is the ultimate form of Press evaluation, and is very similar in principle, if not in methodology, to the Consensual Assessment Technique (Section 2.5.4). Cultural success is a goal that takes a creative product seriously in its entirety—and that subjects it, however indirectly, to the same career pressures that would be applied to a creative human. For systems that are meant to create work that fits in to human movements and genres, a properly quantified measure of cultural success may be one of the most interesting evaluations possible.

2.7 Arguments against evaluating creativity

Now that we have looked at all four perspectives, it is time to mention some important counterpoints from researchers who question whether creativity evaluation is possible at all.

2.7.1 Domain specificity

One surprising argument against the existence of creativity is Baer's (2012) theory of domain specificity. Baer states that there is no such thing as creativity—or, rather, that there are many creative skills, but there is no underlying process which informs them all. Being creative in one domain does not imply the ability to be creative in other domains; therefore, to call a person or process creative without specifying the domain is not scientific. Baer writes,

> "It is sometimes useful to group together beautiful artifacts, fascinating ideas, brilliant designs, and ingenious theories and call them all creative, but that does not mean that they share any underlying unity."

Baer's theory, while disheartening, is supported by plentiful evidence. The Consensual Assessment Technique requires domain-specific experts, and a subject's CAT results in one domain do not significantly intercorrelate with their results in another domain. The only times there have been even modest correlations are in highly related domains, like different kinds of stories or visual art (Baer, 2012).

Similarly, the Torrance Tests predict creative achievement only in their associated domains; performance on the verbal Torrance Test does not correlate with the figural Torrance Test (Baer, 2012). Such tests also correlate only modestly, with actual creative achievement either at the time of taking the test or later in life (Baer, 2011). Personality-based correlates of creativity are different across domains (Baer, 2012), as is the influence of environment on creative performance (Sternberg, 2018). Direct measurements of creative achievement in different domains do not correlate (McKay et al., 2017). The relationship between divergent and convergent thinking, and between verbal and figural domains, also appears to be different in different cultural contexts (Storme et al., 2015). It is difficult to explain these results without acknowledging that some component of creative thinking is different for each domain.

Some results from previous sections support domain specificity. Readers will recall Augustin *et al.*'s (2012) discovery that different Product qualities are important in different artistic forms. Mace and Ward (2002) and Bourgeois-Bougrine's (2014) studies of working artists, while roughly parallel, show differences in process, especially in how a work is completed. Product-based theories are frequently domain specific, such as Cropley and Cropley's (2005) theory of functional creativity, or Jacobs' (2015) theory of foregrounding and backgrounding in literary work. Press-based HCI studies of interactive gallery artwork are often difficult to apply to any other domain. Sternberg (2018), additionally, argues that creativity will become even more domain specific the more eminent and transformational it is.

There is also some evidence that aspects of creativity are domain general. Eminent human creativity often involves cross-applying knowledge from one domain to another (Csikszentmihalyi, 1996). Jordanous (2012b) found that different SPECS criteria are considered important in different domains, but four of the fourteen criteria are very important for every domain: Generation of Results, Originality, Spontaneity and Subconscious Processing, and Value. Since "novelty" and "originality" are synonyms, this would imply that the "novelty and value" definition of creativity, and perhaps others, are domain general. McKay *et al.* (2017) also found that Humor appears to be domain general. Jordanous (2012b) thinks of creativity as partly domain general and partly domain specific.

The evidence for at least partial domain specificity is very strong. Computational creativity researchers must take note of this, and ensure that evaluation techniques are appropriate to the domain. If possible, generalized evaluation techniques should be tested for their applicability to specific domains. However, it does not follow that there is no creativity evaluation. Researchers can continue to evaluate systems that create art, music, mathematics, *etc.*—on the understanding that the evaluations for these different systems will also be different.

2.7.2 Other arguments

Beyond domain specificity, there are other arguments against measuring general creativity. One such argument is that creativity should not be quantified. Boden herself (1990) takes this angle, preferring

to ask "what parts are creative and why?" rather than "how creative?", as this produces more nuanced information for the system's creators. Nake (2012) argues that the quantification of creativity is an American invention, and risks commodifying creativity by framing it as an object one must have a certain amount of, rather than a quality that emerges in a social context. However, many computational creativity evaluations are not quantitative; SPECS is purely qualitative, for instance (Jordanous, 2012b).

Related to this is the argument that *computational* creativity should not be measured by *human* standards. Loughran and O'Neill (2016) typify this argument, stating that humans can already produce creative artifacts which are pleasing to humans. To Loughran and O'Neill, it is more interesting to see what computers produce according to their own, non-human standards. However, most advances in computational creativity will require at least some evaluation: otherwise it is difficult to show that an advance has been made. If one wants to judge computers by non-human standards, one can still make statements about the computer's success performing to these standards, perhaps by using an autonomy-based model such as Guckelsberger's (2017) or a categorization like Ventura's (2016) of the tasks the computer takes on.

Finally, some say creativity is inherently human and can never be present in computers. Boden (1990) lists and refutes lines of argument here, as does Minsky (1982).

First, there is the argument that human creativity is an inexplicable gift which cannot be modeled computationally. Minsky (1982) convincingly refutes this argument, stating that creativity seems to be a combination of ordinary cognitive processes. Collecting domain knowledge, generating ideas, evaluating them, and revising can all be in principle done by a computer, as can the progression from an idea to a plan to a finished implementation. The luck and social factors that lead to a creative achievement being accepted by its field (Csikszentmihalyi, 1996) are also not impossible for a computer, assuming that the field is open to the possibility of a computer achieving something.

If we accept for the sake of argument that human creativity is somehow not computational, we can still get useful results from computational creativity by focusing on product or press. If computers produce believable creative work and influence human culture, then they are doing something useful regardless of their means of doing so. Furthermore, computer systems can still provide evidence for and against hypotheses about parts of the process of human creativity (Boden, 1990).

Second, there is the argument that, due to a computer's lack of richly embodied life experience, its performance will never match that of the greatest creative humans. Boden (1990) agrees with this, but says there are many other reasons why modeling creativity with computers is useful. Mogensen (2017) avoids the problem of matching human levels of creativity by referring to creative systems as having "partial creativity"; the theorists mentioned above, who argue that computational creativity *should not* resemble human creativity, also avoid this problem. Several other proposed solutions exist. Colton (2008) lists the embodiment problem as a reason for providing framing information, to create the illusion of humanlike experience. Moreover, a number of creative systems are actually embodied, either as robots (Infantino et al., 2016; Tresset and Deussen, 2014) or in a virtual world (Aguilar and Pérez y Pérez, 2014).

Third, there are arguments that even if a computer had humanlike processes and products, it could not be "really" creative (Boden, 1990). These arguments can stem from appeals to the non-biological nature of computers, or to variants on Searle's (1980) Chinese Room argument, or to a lack of consciousness on the computer's part, or finally, to the simple belief that creativity is a property only of humans. Chinese Room-style arguments can be countered by the argument that understanding is an emergent property of the system as a whole, or Minsky's (1982) argument that humans do not "really" understand things either:

our commonsense knowledge consists of imprecisely defined concepts induced from sensory perception, and it is possible for a computer to do this sensory processing as well. As for consciousness, Minsky (1982) argues that this is merely an ability to monitor oneself. Many creative systems do have a rudimentary ability to monitor their own work, and in principle, nothing stops this self-monitoring from becoming more sophisticated. Linkola *et al.* (2017) discuss one possible framework for this kind of self-monitoring.

Arguments about "real" creativity, "real" understanding, and "real" consciousness can be pernicious due to the ill-defined nature of these terms. This lack of definition leads to moving goalposts and unfalsifiable arguments. McCormack and d'Inverno (2014) suggest that, once a computer can perform a task, humans will no longer see that task as creative—even when humans do it. As a result, creative humans will concentrate on whatever tasks computers have not yet achieved. There is some precedence for this in, for example, the movement away from photorealism in painting following the invention of the camera.

In our opinion, none of these arguments are reasons to do away with evaluation. However, researchers should be aware that these viewpoints exist. For some members of the audience, "real" creativity may be a moving goal that a computer, no matter how sophisticated, can never quite meet.

2.8 Issues in computational creativity evaluation

So far, we have seen many theories of creativity and methods for evaluating creativity. We have seen potential issues arise, such as debates about definitions of terms, a lack of autonomy in existing systems, the cultural specificity of many judgments, and the potentially domain-specific nature of creativity. Some of these issues have obvious implications when applied to evaluation in practice. Others are unsolved problems.

We now turn to some issues that arise due to the practicalities of computational creativity evaluation. Some of these have to do with problematic evaluation methods, or ones not well-supported by theory and evidence. Some are practical questions to think about, while others are common pitfalls. Discussing problematic evaluations necessitates discussion of meta-evaluation: how to evaluate evaluation methodologies.

2.8.1 Implementations of models and *ad hoc* tests

It is one thing to propose criteria for creativity, and another to operationalize them to be used in practice. A number of researchers have operationalized models such as Ritchie's criteria (Gervás, 2002; Tearse et al., 2011) or the Creative Tripod (Chan and Ventura, 2008; Monteith et al., 2010; Norton et al., 2010; Smith et al., 2014), either by creating a questionnaire based on the model's criteria, or by somehow automating the judgments. Others have created their own questionnaires *ad hoc*. Operationalizations of novelty and value (Section 2.5.1.4), Ritchie's criteria (Section 2.5.1.3), and the Creative Tripod (Section 2.6.1) have already been discussed; we now turn to the issue of questionnaires which are not, or only partly, based on such models.

Some researchers evaluate only part of a model, or combine criteria from multiple models. Karampiperis *et al.* (2014) combine novelty and surprise (but not value) with Lehman and Stanley's (2012) impressiveness. A few systems are designed specifically for the Appreciation portion of the Creative Tripod (Norton et al., 2013).

Questionnaires used to evaluate creative systems do not necessarily adhere to an established model. Neither do the criteria used in a system's internal fitness measures, if applicable. More commonly researchers evaluate systems according to *ad hoc* criteria. We have encountered dozens of such criteria, and Jordanous' (2012b) meta-analysis describes them as one of the most common forms of evaluation, but for space reasons, we make no attempt to list them all. A few illustrative examples of domain-general *ad hoc* criteria are meaningfulness (Das and Gambäck, 2014), interestingness (Román and Pérez y Pérez, 2014), and coherence (Harmon, 2015); while domain-specific *ad hoc* criteria include grammaticality for poetry (McGregor et al., 2016) or whether a respondent would use a generated image as desktop wallpaper (Norton et al., 2013). Many of these studies include both domain-general and domain-specific criteria, as well as novelty, value, surprise, or a modified Turing test.

With *ad hoc* criteria, there is often little or no justification for why these criteria were used, except that they were the criteria the researcher happened to be interested in. Arguably, since part of creativity is likely to be domain specific, *ad hoc* criteria which are domain specific may be a better fit for a given project than standardized criteria. However, if a researcher argues that their system is creative—as opposed to merely retweetable or humorous—then such criteria must be based on experimental evidence with regards to the system's domain and rigorous theory.

Both *ad hoc* and theoretical criteria should be tested for properties such as construct validity, in the same way as other psychological criteria. The only model we are aware of that has been validity tested in this way is Carroll and Latulipe's (2009) Creativity Support Index, which is used for creativity support tools and is inappropriate for non-co-creative systems.

Some preliminary results indicate that most creativity criteria are not independent. Tapscott *et al.* (2016) found that measures of quality and narrative potential are interdependent; Pereira *et al.* (2005) found that ratings of typicality and quality under Ritchie's model are not independent. Lamb *et al.* (2015a) found similar results with Ritchie's model, the IDEA model, and the Creative Tripod when assessed by non-experts.

A related problem is that some criteria are operationalized simplistically; defining "poeticness", for example, as the use of a meter and rhyme scheme (Das and Gambäck, 2014), when the majority of contemporary English poetry does not rhyme. Beginning with such goals in the early stages of a project is defensible, but for a system's output to be taken seriously, effort must be made to evolve towards criteria that resemble the actual criteria applied to human work.

2.8.2 Opinion surveys, non-expert judges, and bias

Aside from the criteria on a questionnaire, several other issues can arise. One issue is rater expertise. Intuitively, people who know little about a kind of creative artifact might not be good judges of those artifacts. Some researchers have expressed this intuition in their published work: for example, Gervás and Veale both independently worry that humans will rate their systems' output too highly because they do not understand it (Gervás, 2002; Veale, 2015). The intuition is supported by considerable evidence. The Consensual Assessment Technique requires domain expert judges for a reason: only domain experts can be trusted to judge artifacts in that domain (Kaufman and Baer, 2012). Non-expert judges lack interrater reliability. Even when they agree, the validity of their judgments is in serious question, because they fail to correlate well with the judgment of experts.

This should not be surprising; it is well known in cognitive science that experts perceive the subject of their expertise differently from novices, "chunking" and analyzing patterns that are invisible to a novice (Gobet and Simon, 1998). In art evaluation, experts evaluate art differently from novices in several ways. Photography experts carry out less simplistic evaluation than non-experts, and prefer more unfamiliarity and uncertainty (Galanter, 2012). The relationship between novelty and typicality is different for experts and non-experts, with experts showing a stronger preference for novelty (Hekkert et al., 2003). Art experts focus more than novices on relationships between properties, and less on the properties themselves; they rely more on domain-specific knowledge, while novices rely more on their general life experience (Kim et al., 2011). Visual art experts show less pronounced emotional responses to visual stimuli than novices, rely less on emotion in their judging artworks, and are more tolerant of negative emotions in art (Leder et al., 2014). Experts and novices even move differently, with novices either losing interest or hovering around the art without a mental framework for interpretation (Ryokai et al., 2015). Rather than being a weaker version of expert judgment, novice judgment tends not to correlate with expert judgment at all (Kaufman et al., 2008) and can even run in the opposite direction (Lamb et al., 2015a).

Non-expert raters in computational creativity face an additional problem: they typically do not know how to apply the concept of creativity to a machine. As mentioned earlier, many people are reluctant to attribute creativity to machines; and experiments suggest that those who are willing are frequently unsure how to do it. Jordanous (2014) found that participants in a survey, asked how creative a system was, expressed confusion as to what definition of creativity to use, and admitted they were likely to conflate creativity with other factors. Norton *et al.* (2011) found that art students were reluctant to evaluate a computer's creativity without knowing more about its process.

Some researchers (Colton, 2008) worry that this reluctance will lead to bias against creative computers. The evidence for such bias is patchy. Moffat and Kelly (2006) found that musicians and non-musicians are biased against computer-generated music, but their sample size is quite small, and other ways of analyzing their data did not yield this result. Other researchers have generally not reproduced Moffat and Kelly's results. Friedman and Taylor (2014), Norton *et al.* (2015), and Pasquier *et al.* (2016) found that, while individuals differ, there is little overall bias against computational creativity in the general population.

Gade *et al.* (2017), in a more nuanced experiment, gave the same computer-generated art to two groups of participants, with one group being told a truthful account of its generation, and the other given deceptive framing information in which the artist was described as a specific human with specific reasons for creating the work. Both groups were given a detailed questionnaire of different statements about the art and the artist. Gade *et al.* found that, overall, there was no statistically significant bias against work known to have been created by a computer—except for three questions that dealt with *Person*-based creativity, such as, "I would describe the creator of this piece as an artist." However, a minority of participants in the deceptive group changed their answers across the four Ps to be significantly more negative after it was revealed that the artist was not a human.

The mixed results of these different researchers seem to indicate that bias against computers exists for some individuals, and in some types of judgement, but is not generally pervasive. However, the idea of pervasive bias against computers is still widely referenced. Colton *et al.* (2008) recommend the use of framing information to combat bias, but McGregor *et al.* (2016) found that framing information does not significantly affect human ratings of computer-generated artifacts, and Gage *et al.*'s study, as mentioned, found a significant difference only for Person-based ratings. A more subtle question is if humans are biased towards familiar and humanlike forms of creativity. Guckelsberger *et al.* (2017) raise this question when discussing embodied creative agents whose bodies might be very non-anthropomorphic. Framing

information might be useful to help humans understand an agent who is unlike them; but this remains to be tested.

Some researchers argue against opinion surveys altogether. Colton (2012) worries that evaluating systems by surveying groups of humans would lead to "creativity by committee"—which would presumably be bland or otherwise undesirable. However, no real evidence underlies this claim, and some evidence—the success of the CAT, for instance—suggests that groups of experts can evaluate competently.

Jordanous (2014) argues that a lack of reliability renders opinion surveys unsuitable as an evaluation method; she recommends a detailed evaluation by a single expert. However, Jordanous's argument is not as general as it seems: the "opinion surveys" she cites are based on asking the question "how creative is this?" to mixed expert and non-expert raters (Jordanous, 2012b). It does not follow that the same question asked in a structured way to experts (as in the CAT), or broken down into components (as in SPECS), will not provide suitable results. However, in any case involving human judges, care must be taken to avoid any problems caused by confused, inattentive, biased or inexperienced judges.

2.8.3 Meta-evaluation

Some researchers have turned to the question of meta-evaluation—that is, if there is a systematic way to judge the merits of evaluation techniques.

An early study in meta-evaluation was Pearce *et al.*'s (2002) for music generation. Meta-analyzing papers in that field, they discovered a "methodological malaise." Most researchers neither clearly specify a purpose nor choose an appropriate evaluation method.

According to Pearce *et al.* (2002), different purposes necessitate different kinds of evaluation. The purpose of a creative system might be to create art: that is, the developer wishes to express themselves using a computational system. Pearce *et al.* argue that, while there is nothing wrong with this, it is art and not science. It should not be published in a scientific journal unless the artist makes a scientific or technological advance in the course of their work. The art itself should be evaluated through Press: critical acclaim, popular appeal, placement in curated exhibits, *etc.* Indeed, as we noted in Section 2.6.7, these methods are used by many.

Another purpose might be the creation of a general-purpose creative system, either autonomous or with human collaboration. This is an engineering task, and should be subject to normal engineering process, including requirements analysis, specification, and testing. Researchers often fail to specify their engineering goals—in particular, to list practical scenarios in which their system would be useful, and to state the conditions under which they will deem the system successful.

The two other purposes cited by Pearce *et al.* (2002) are scientific: investigating a theory about an artistic matter (for example, a theory of musical style) or investigating a theory about the cognitive processes of human artists. These should be investigated using the scientific method: clearly stating a hypothesis, using methods derived directly from theory while minimizing confounding factors (for example, one should not go into the system and make *ad hoc* manual tweaks so that it sounds better), and systematically attempting to disprove the hypothesis. Subjective statements by the researcher, which the researcher does not make an attempt to falsify, should be avoided (Pearce et al., 2002).

Pérez y Pérez (2018) draws a similar distinction between two approaches in computational creativity: the engineering-mathematical approach in which systems are designed to create a creative product in the

most efficient or convincing manner, and the cognitive-social approach in which systems are designed to simulate the cognitive and social mechanisms of human creativity in the most faithful way. Rather than two binary camps, Pérez y Pérez posits these as a continuum, but he notes that applying evaluation methods without attention to the system's intended place on the continuum—for example, using engineering evaluations to evaluate a cognitive-social system—can lead to evaluations that miss the point. Pérez y Pérez also proposes a third branch of the continuum, for creative systems that are purely artistic in intent.

Jordanous (2014) proposes a set of five standards for evaluation methodologies. Good evaluations should be correct, in the sense of accurately and comprehensively portraying a system's creativity. Jordanous does not believe in ground truth about creativity, but correct feedback should be appropriate and realistic for the system being evaluated. Good evaluations should be useful for understanding and improving the system. They should faithfully capture creativity (as opposed to some other trait which could be conflated with creativity). They should be usable and easily applied. Finally, good evaluations should generalize across various types of creative system (note that this contradicts Baer's insistence on domain specificity). Jordanous states that a single methodology that works for every system probably does not exist; instead, we can use the five standards to talk about the strengths and weaknesses of a methodology and its suitability for a particular purpose. An exercise for the reader might be to choose some evaluation techniques discussed in this paper, and perform one's own informal analysis of their merits based on these five standards.

2.9 Conclusion: Best practices for the assessment of creativity in computational systems

We have now seen the major theories from the four perspectives of what creativity is; the major ideas from the four perspectives about how to evaluate creativity; some counter-perspectives proposing that creativity is not one thing, or perhaps should not be evaluated at all; and some additional pitfalls that occur when evaluating computational creativity in practice. A reader of this thesis is now equipped to think about how to evaluate their (or another researcher's) creative computational system in an evidence-based way.

We now summarize our suggestions for best practice based on each of the four perspectives, and overall.

2.9.1 Person

Evaluating from the Person perspective is useful if you want your system to be viewed as a creative agent because of inherent cognitive traits that resemble a human's. Existing tests which measure these traits in humans should be used to test the computational system. Outside of this narrow set of cases, the lack of close resemblance between a computer's architecture and a human's brain means that claims about a computer system should likely be based on another perspective.

2.9.2 Process

Evaluating from the Process perspective is useful if you want to model human creativity, or if you want to argue that your system is creative because of the kinds of tasks it does and how it does them. Process

evaluation is often not quantitative. Instead, it either places the system in a category or qualitatively describes the system's creative strengths and weaknesses.

The Process perspective is most useful when it is taken into account at all stages of system design, as Process theories have a lot to say about how creative work should be done. Even systems that one does not intend to evaluate from the Process perspective can benefit from these process theories. Researchers building a system should think about how the system explores its conceptual space and what the parameters of that space are. They should consider building the system's workings based on a generation-evaluation loop, in which the evaluation stage in the loop leads the system's work to incrementally improve, or on a more complex looping process. They can also consider building a system to move from inspiration to planning to full creation as working artists do, rather than trying to generate a full artifact all at once. Finally, researchers should consider the issue of autonomy. While extreme levels of autonomy may not be appropriate for all systems, it is good to be clear about which decisions are made by the computer and which by humans, both at the planning stage and when communicating with audiences and judges.

2.9.3 Product

Evaluating from the Product perspective is useful if the goal of your system is to produce something that is useful or pleasing to humans. Product evaluation is the category with the largest proliferation of specific criteria. Product criteria are also the easiest to encode in the system itself, if the system is meant to evaluate its own work (in a generation-evaluation loop or otherwise). The challenge for the researcher in a Product evaluation is to choose appropriate criteria for the system's design, domain, and goals.

The most general criteria, with the widest theoretical support, are novelty and value. These criteria, however, have subtleties and pitfalls. Researchers should carefully consider what kind of novelty their system's products should have. Novelty should be distinguished from randomness, and novel output should still be appropriate for its intended domain. Value should be defined in terms of the specific audience for whom the system's products should be valuable.

Other criteria besides novelty and value have been proposed, and Product evaluations also exist which lack criteria. The modified Turing test, which lacks criteria, is not appropriate unless the system is specifically meant to imitate existing styles of human work. The CAT is a very robust and valid criteria-free Product evaluation, but is more difficult than some other evaluations to perform, as we will see illustrated during our own attempts in Section 4.2.

2.9.4 Press

Evaluating from the Press perspective is useful if your system is intended to make an impact on society, influence human opinion, and be recognized for its work. A researcher should first decide what effect the system is intended to have on its audience, then attempt to produce this effect and assess their success.

Certain well-tested Press evaluations already exist for certain specific types of creative work. Interactive systems should be evaluated according to the standards of interaction design, especially in the domain of interactive gallery artwork, which is very well-studied. Similarly, co-creative systems should be evaluated according to existing standards for creativity support tools. Finally, if a system is meant to create finished work on its own, an excellent Press evaluation method is to submit the work to the same gatekeepers who judge human artifacts and see what happens.

2.9.5 Best practices regardless of perspective

Care should always be taken to avoid unfalsifiable claims. Systems should never be evaluated solely through rhetorical argument by the researcher. Goals, including the conditions under which the system would be considered successful or unsuccessful, should be clearly stated before the evaluation begins. Any important terms used, such as the criteria used in a Product evaluation, should be clearly defined.

Evaluation techniques should be based on existing evidence as much as possible, whether that evidence is experimental or theoretical. The techniques chosen must be appropriate to the given domain, as many aspects of creativity and its evaluation are domain-specific. Ideally, an evaluation technique should be tested for its reliability and validity in each domain for which it is used. However, for most domains in computational creativity, techniques that have been tested thoroughly in this manner do not yet exist. At minimum, the researcher should know if their system is art in itself, an engineering product, or a scientific experiment, and should choose a type of evaluation which is appropriate to the given mode of inquiry.

Finally, since most evaluation across perspectives involves some form of human judgment, attention should be paid to the qualifications of the humans who are judging. Expert judges should be used whenever possible, since they have better interrater reliability than non-experts, and (at least in domains, such as "high art" and science, which are typically overseen by experts) better validity. Not all judges, even if they are experts in the given domain, will be comfortable with the task of judging a computer's creativity. It should not be assumed that judges will be overall biased against computers, but they may struggle to know how to appropriately apply the concept of creativity to a machine, and should be guided accordingly.

2.9.6 Deviations from best practice

It is extremely common for researchers to deviate from best practices: for instance, to survey nonexpert judges, perform modified Turing tests, or make rhetorical arguments. One common reason for these deviations is time and effort. Many theoretically sound methodologies, such as the CAT, are too complicated to always be practical. Even the use of expert judges requires time, effort, and expense, as experts willing to perform such evaluations can be difficult to find. In an experimental field—particularly when making computational art that does not exactly resemble human art—it may be difficult even to identify an expert. If sheer difficulty causes the disconnect between theory and practice, than new ways of making theory simpler to implement are desperately needed.

Another interpretation is that the goals of practical researchers diverge from those of theorists. For example, if one's goal is not to construct a "really" creative system but to convince the public that one's system is creative, then rhetorical argumentation and the use of non-expert raters may be very appropriate. More seriously, it is appropriate to survey non-experts about one's system if the system's goal is to satisfy non-experts. Similarly, *ad hoc* surveys can be defensible on the grounds of domain specificity (Baer, 2012). Apart from the CAT, most existing evaluation methodologies are not domain specific. And as we have seen in Section 2.8.1, many *ad hoc* surveys used by computational creativity researchers contain domain specific criteria. It may be that researchers avoid or modify general theories of creativity because they already understand that their task is domain specific.

If this is the case, then what is needed is a proliferation of more rigorous methodologies appropriate to specific domains. If a researcher needs criteria specific to the needs of music, or mathematics, or Internet memes, then their need must be matched by systematic attention to the problem of which criteria represent

the specific needs of that domain. Such criteria can, one hopes, be based in existing scholarship pertinent to the domain in question, and solidified through experiments.

In the next chapter, we will begin our study of the domain of poetry by surveying the computationally creative work that has been done in this specific domain.

Chapter 3

Related Work in Computational Poetry

3.1 A taxonomy of generative poetry techniques

3.1.1 Introduction

Poetry is not the most studied domain in computational creativity today (Loughran and ONeill, 2017) but computer-generated (or "generative") poetry has a long history of study among various communities—from artists exploring the effects of algorithms on language, to Internet hobbyists, to computer scientists. Computer-generated poetry can be an intriguing curiosity, a satire, a meeting of AI and performance art, a serious artistic attempt to explore the possibilities of random and arbitrary text, an attempt to entertain and impress computer users, or an exploration towards concrete advances in computational natural language generation. This multiplicity of purposes leads to a wide variety of different authors in different communities, with different goals and interests, who do not necessarily communicate with communities other than their own. However, after examining many examples of computer-generated poetry, we find that the techniques used to generate such poetry can actually be boiled down into a few simple categories with well-defined relationships.

We define these categories as Mere Generation, Human Enhancement, and Computer Enhancement. In mere generation, a computer produces text based on a random or arbitrary algorithm. Most systems use some form of mere generation as a baseline. However, in the remaining two categories, either the results of mere generation are modified and enhanced, or some meaningful enhancement is built into the mere generation model's parameters. This occurs either through interaction with a human (Human Enhancement), or through the use of optimization techniques and/or knowledge bases by the computer (Computer Enhancement). The results of mere generation can appear nonsensical, though this is not always a bad thing from an artistic perspective. By bringing in knowledge about words and the world, and by setting artistic goals, both human and computer enhancement drive generative poetry towards coherence and artistic style.

In the rest of this section, after discussing other categorizations and our view of poetry, we illustrate with examples the techniques used in our three categories. We explain why scholars have moved away from Mere Generation, and argue that Computer Enhancement, while pursued primarily by scientists rather than artists, has the potential to solve *artistic* dilemmas in generative poetry. We then briefly bring up the related field of generative music, to show that our taxonomy can be generalized to other creative tasks besides the generation of poetry.

Prior versions of this section were published in BRIDGES 2016 (Lamb et al., 2016b) and in the Journal of Mathematics and the Arts (Lamb et al., 2017b).

3.1.1.1 The goals of poetry

As computer scientists, we should clarify what we mean when we talk about artistic perspective, artistic style, or artistic goals. Human poets write to satisfy a variety of complex goals, the full spectrum of which are beyond the scope of this chapter. According to the Princeton Encyclopedia of Poetry and Poetics (Abrams, 1974), a few of the most common poetic goals include:

- Describing or imitating reality (the Mimetic Theory)
- Having a specific effect, such as educating or inducing an emotion, on readers (the Pragmatic Theory)
- Communicating the poet's emotions (the Expressive Theory)
- Creating a work of art that exists for its own sake (the Objective Theory)

These goals are very broad and each can be conceptualized and implemented in a variety of ways. Most ways of achieving most of the goals will require, as a necessary but not sufficient aspect, that a poem is intelligible to an educated human reader. By understanding what is in the poem, the reader can more easily recognize the description of reality, understand the emotions of the author, experience the poem's intended effect, or appreciate the poem's inherent value. For the purposes of this chapter, when we speak of mainstream artistic goals, we are referring to this kind of poem in which the reader is meant to understand something meaningful.

It is, of course, possible to write a poem which is valuable without being understood. Indeed, this was the explicit goal of several 20th century poetic movements, including the Dadaist movement, which used intentionally nonsensical text in order to express rebellion against language's rules and traditions (Balakian, 1974). Intuitively, computer-generated poetry might seem to be more compatible with these rebellious movements than with mainstream poetry. There is nothing wrong with rebellious poetic movements, but part of our argument is that computers using AI techniques have the potential to generate content that is meaningful to humans and thus enter the mainstream. We will return to this mainstream/rebellious distinction throughout this section.

Digital poetics encompasses a wider range of techniques than those described here. Computer systems allow a variety of new techniques to human poets including hypertext poetry, kinetic poetry, chatbots as art, interactive fiction (Douglass, 2014), multimedia poetry, and even poetry that presents itself as a game (Funkhouser, 2012). However, for the purposes of this thesis, we are interested only in poems representable by a static text file, in which the computer has a meaningful role in determining what the text will be. This is our working definition of 'generative poetry'. We will focus primarily, though not exclusively, on generative poetry in the English language.

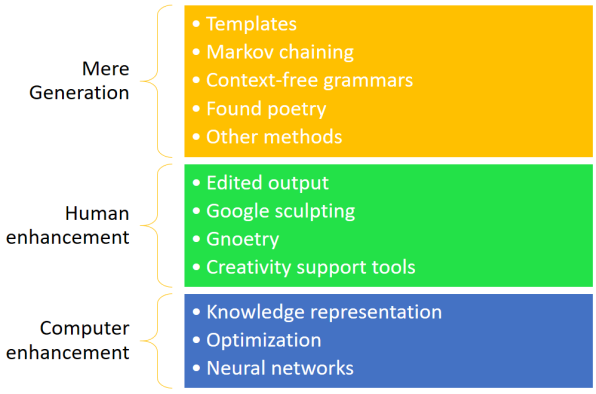

Figure 3.1: A diagram illustrating our three-part taxonomy.

3.1.1.2 Prior work

Other taxonomies of generative poetry exist besides ours. Roque (2011) classifies poems according to the goals of their creators, while Funkhouser (2007) uses the categories of permutational, combinatorial, and template-based generation. Gervás (2002) divides what we would describe as the Computer Enhancement category into four subcategories based on the type of artificial intelligence techniques used. Oliveira (2017b) compares and contrasts Computer Enhancement poetry systems on a number of axes, including their language (e.g. English or Finnish), the poetic features that are emphasized, the method by which they select their content, the AI techniques used, the degree to which human-authored text is exploited or reused, and the evaluation techniques used. However, given this large number of axes and limited space, Oliveira's survey does not have much room for deep analysis.

These taxonomies are useful. However, our taxonomy serves needs that others do not. It includes and takes seriously generative poetry from a variety of sources, whether scientific, hobbyist, or artistic, and focuses not on technical descriptions of processes but on the uses to which the processes are put: to generate text, to optimize text, or to add knowledge of the real world into the generation process.

The closest existing taxonomy to ours comes from Ventura's recent paper (Ventura, 2016), where he classifies computationally creative systems by the sophistication of the computational techniques used. Ventura's taxonomy is abstract and broad, and aimed at describing the degree to which a system is creative. Ours is focused specifically on poetry, including forms of poetry not made with computational creativity in mind. However, Ventura's taxonomy is a good counterpart to ours, and we will refer to it throughout this section.

3.1.2 Mere Generation

As Ventura points out, 'mere generation' is a term with a long and loaded history in computational creativity (Ventura, 2016). Many researchers claim that they want to transcend mere generation, but the term is often used without a precise definition. For the purposes of this chapter, when we refer to mere generation, we refer to systems in which *the computational element* of the system is either random

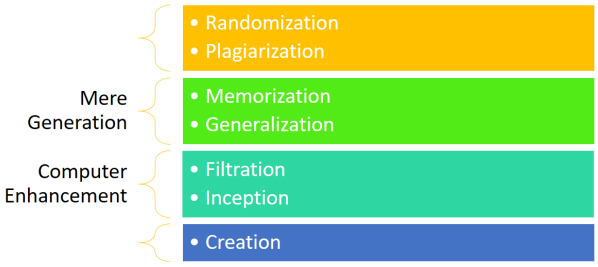

Figure 3.2: A diagram illustrating Ventura's taxonomy and its relation to ours. Outer labels ("Mere Generation" and "Computer Enhancement") are ours, not Ventura's, and are meant to illustrate areas of overlap between the two taxonomies, not to imply that Ventura would necessarily use the terms in this manner. Note that Ventura's taxonomy, unlike ours, explicitly proceeds in a hierarchy from the least creative (top) to the most creative (bottom) methods. It includes some categories which we have not observed in actual generative poetry systems, but does not include anything corresponding to Human Enhancement.

or arbitrary. Humans may have painstakingly handcrafted the system to increase the likelihood that the random calculation results in a pleasing output, and the randomness may be taken from a non-uniform distribution; but the *computational process* used by the system is random or arbitrary. From a Process perspective, this means that the systems are not especially interesting; they do not use algorithms that represent an advance in artificial intelligence, and they do not model the cognitive steps theoretically associated with creativity. Indeed, this is a major reason why researchers state that mere generation is to be avoided. It is quite difficult to argue from a Process perspective that a mere generation system, as opposed to its human programmer, is creative.

However, from a Product perspective, there is no a priori reason to assume that the poetry produced by mere generation will be inferior to the poetry produced in another category. Likewise, from a Press perspective, there is no reason to assume that the interaction between a mere generation system and the public is inferior to any other interaction. In fact, it can be argued from a Press perspective that the concept of mere generation does not even make sense, because creative systems involve a multiplicity of interesting interactions, not merely a single generation step (edde addad, 2018). If we wish to claim that there is anything superior about non-mere generation from Product or Press perspectives, we will have to look at the poems and their critical reception empirically.

Our definition of mere generation corresponds to the first few categories in Ventura's model: Randomization (in which elements of the creative product are chosen completely at random), Plagiarization (in which existing examples of the creative genre under consideration are copied), Memorization, and Generalization. In Memorization and Generalization, the system attempts to capture and replicate certain properties of a training set of human-made work, without duplicating the human-made work entirely, but no knowledge-based or goal-directed reasoning is performed. Note that in our research we have not come across any poetry systems simplistic enough to correspond to the Randomization or Plagiarization categories.

By using the term 'mere generation', we do not mean to imply that there is necessarily anything 'mere' or trivial about either the poems or the human artistic process that goes into making them. In some domains of computational creativity, such as musical theatre (Colton et al., 2016), simply producing a

result is a significant achievement. Even in the relatively simple domain of poetry, a poem which is Mere Generation from a computational standpoint may have more thought and artistic skill put into it by the poet than many poems from other categories. We use the term not to belittle but to distinguish these systems from more algorithmically complex ones.

In the rest of this subsection we will list methods used by mere generation systems. Note that not every system using one of these methods is a mere generation system. Systems that *straightforwardly* apply the methods below, without further processing and enhancement, are mere generation systems. This includes all the specific systems that we use below as examples. As we will describe in the following subsections, most Human Enhancement and Computer Enhancement systems begin with a mere generation method and modify or build on top of it.

3.1.2.1 Methods of Mere Generation

3.1.2.1.1 Templates. Template-based generation, also called slot-filling, has been common since the first generative poetry program, Theo Lutz's 'Stochastic Texts' (Lutz, 1959). Template generation is one of the simplest means of constructing a poem. The basic steps are as follows:

1. Create lists of words or phrases in different categories, e.g. nouns or verbs (though the categories need not be based on a part of speech).
2. Create one or more line templates with slots into which a word from a given list can be inserted.
3. Randomly select a word from the appropriate list to fill each slot.

This process corresponds to Ventura's category of Generalization. Essentially, the template models poetry as a sequence of types of words in a particular order, and randomly creates new work to fit that model.

'Stochastic Texts' uses templates and word lists based on lines from Kafka's 'The Castle', resulting in lines like the following (Lutz, 1959):

> NOT EVERY LOOK IS NEAR. NO VILLAGE IS LATE.
> A CASTLE IS FREE AND EVERY FARMER IS FAR.
> EVERY STRANGER IS FAR. A DAY IS LATE.
> EVERY HOUSE IS DARK. AN EYE IS DEEP.
> NOT EVERY CASTLE IS OLD. EVERY DAY IS OLD.

Template poetry can draw its word lists from existing art in this way, or from dictionaries representing the whole of the language, or the lists can be handcrafted by the programmer. Slots in a template need not be filled by a single word; the word lists can just as easily contain phrases or other structures.

A major issue in template poetry is repetitiveness. Often, running a template program several times will produce a repetitive effect in which the template's structure becomes obvious, as in this example (from 'The House', by Alison Knowles and James Tenney (Knowles and Tenney, 1968)):

A HOUSE OF STEEL
IN A COLD, WINDY CLIMATE
USING ELECTRICITY
INHABITED BY NEGROES WEARING ALL COLORS
A HOUSE OF SAND
IN SOUTHERN FRANCE
USING ELECTRICITY
INHABITED BY VEGETARIANS
A HOUSE OF PLASTIC
IN A PLACE WITH BOTH HEAVY RAIN AND BRIGHT SUN
USING CANDLES
INHABITED BY COLLECTORS OF ALL TYPES

Such repetitiveness may, as in the given example, be intentional. However, most poets want output that looks fresh each time the program is run. One way of achieving this is with templates that change over time. John Morris, for example, uses shifting templates to create haiku without obvious syntactic repetition (Morris, 1973):

Frogling, listen, waters
Insatiable, listen,
The still, scarecrow dusk.

Listen: I dreamed, was slain.
Up, battles! Echo these dusk
Battles! Glittering...

3.1.2.1.2 Markov chaining. A Markov chain is a statistical model applied to data in a sequence. Based on the last N entries, an Nth-order Markov chain calculates the probability distribution for the next entry. The sequence of N entries used to make the prediction are referred to as an *n-gram*; a critical feature of Markov chaining, which keeps computational costs low, is that no entries before the n-gram need to be considered. Based on the entries in the n-gram, the Markov chain calculates the probability of different entries appearing next. It then probabilistically generates the next entry, which produces a new n-gram (including the newly generated entry, but not the first entry in the old n-gram). Further entries can be generated in the same way indefinitely.

For poetry generation, the entries in the n-gram can be characters or words. A probability model is generated either from a broad corpus or a specific work. Based on a few user-generated starting entries, the system repeatedly samples from the model to create the next entry. Using words as entries ensures that no novel or partial words appear in the output. Using characters results in many non-words and neologisms, which can be an intended effect, as in this example from Dissociated Press by R.W. Gosper (Roque, 2011):

book her sist be chin up seen a good deal uneasilent for coursation
dropped, and the
litter on,

The Queen was
siliarly with them, the Footmance.
Would guess, an't
grom one foot to thistle, to keep and reachinah'll be she could not ever who had not atte-book hastily. Convers began to trings; into thing on wast the door, and
the.

Markov chaining is also an example of Ventura's Generalization stage. In this case, instead of a specific ordering of types of words, the system's model of poetry is the set of probability distributions forming the Markov chain, which is generated by copying the probability distributions of an existing text.

Markov chain poetry is related to Dadaist 'cut-ups', invented by Brion Gysin, in which a text is cut into N-character blocks and rearranged (Gysin et al., 1978). While Markov chains preserve many features of the input text, they fail to replicate grammar. Even with word-based models, output tends to meander incoherently from one topic to another. This wandering feature can be used to artistic effect. For example, a corpus built from two constrasting sources can produce language that switches strikingly from one style to another, as in the case of Charles Stross's Lovecraft.pl (Stross, 2013), which merges a Markov chain derived from the King James Bible with one derived from the horror fiction of H.P. Lovecraft.

3.1.2.1.3 Context-free grammars. Poets who want something more grammatically correct sometimes turn to context-free grammars, a class of models from computational linguistics. A context-free grammar constructs sentences by recursively applying generation rules. It consists of terminal symbols (the actual words or phrases that occur in a finished poem) and non-terminal symbols. Each non-terminal symbol corresponds to a set of terminal and/or non-terminal symbols with which it can be replaced. A non-terminal symbol can also be replaced by a group of symbols. During generation, the system begins with a single non-terminal symbol. It then runs replacement operations repeatedly. At each step, every non-terminal symbol is replaced with an entry in its corresponding set of symbols. The process concludes when no non-terminal symbols remain.

Compared to Markov models, context-free grammars are a better representation of the recursive structure of human language. For example, the non-terminal symbol [NOUN PHRASE] can be replaced with a noun, or with other structures that take the grammatical place of a noun. These structures can become arbitrarily complex ("the boat that the man in the green suit rowed across the river yesterday") and still be grammatically correct.

By adding Markov chain-like random choice of rules to a context-free grammar, one can construct a stochastic context-free grammar which constructs the most likely sentences based on some input corpus while remaining syntactically correct. Jim Carpenter's Electronic Text Composition uses such a probabilistic grammar, resulting in language which is semantically odd, yet more syntactically coherent than the output of a Markov chain (Carpenter, 2004):

The important statement, like one advance act.
A spit goes eastern, showering.
it perches on the branching foam
The statement.

3.1.2.1.4 Found poetry. Another mere generation method is to skip the generation process and, instead, use a computer to harvest text written by humans. While most generation techniques make some use of pre-existing human language—as a corpus for calculation of N-gram frequencies, for example—found poetry preserves entire human-written sentences without significant modification. The computer's role is to select text which meets some constraint and present it, perhaps juxtaposed with other texts, outside its original context. Examples include Ranjit Bhatnagar's 'Pentametron', which constructs sonnets by assembling pairs of rhyming, 10-syllable posts on Twitter (Bhatnagar, 2012), and the New York Times Haiku project, which mines phrases with a 5-7-5 syllabic structure from that newspaper (Harris, 2013):

> Surely that shower
> couldn't have been going since
> yesterday morning.

It is not clear exactly where found poetry would fit in Ventura's classification. Arguably, since the found poetry systems described here contain encoded knowledge about poetry (for instance, the number of syllables required), they too would belong in the Generalization category. However, if a found poetry system used actual poetry of the appropriate type as source material, it would quickly devolve into simply reproducing this poetry (Plagiarism from Ventura's classification) or, perhaps, vaguely reproducing it with some errors (which would be Memorization).

3.1.2.1.5 Miscellaneous methods. Some poetry is generated using other mere generation methods. Poetry can be constructed by, for example, selecting words which contain certain combinations of letters and arranging them on the screen (Funkhouser, 2012). Or the words in a short phrase can be permuted, as in Brion Gysin's 'I Am That I Am' (Funkhouser, 2007). Edde addad's poem 'Disappointment and Self-Delusion' is constructed around mistakes made by the speech recognition software Dragon NaturallySpeaking (Roque, 2011). We will not explore these methods in detail here. Since the linguistic processes done by the computer are random or arbitrary, they are still classed as mere generation.

3.1.2.2 The problem with Mere Generation

While the methods described above may seem endlessly flexible, there is a limit to what they can do. The frequency of words and the rules of grammar can be modeled, but semantic coherence is more difficult: the poems are valid sentences or language fragments, but do not really mean anything individually or as a whole. This is an obstacle to any poet who wants their poetry to meet a mainstream poetic goal and be understood by readers.

For this reason, some pioneers and experts in generative poetry have grown discouraged with the form. Charles Hartman, for example, after years of creating generative poetry, concluded that even the most syntactically correct programs 'did only a little to drive the random words toward sense (Funkhouser, 2007)." While poetry generated using templates can make sense, it is difficult for a template to retain interest and meaning after repeated use exposes its pre-set structure. Chris Funkhouser wrote in 2012 (Funkhouser, 2012):

> Researching the topic for nearly twenty years, I have encountered dozens of poetry generators. Many have issued convincing poetry, but even the best of them fatigue the reader with blatant slotted structures and repetition.

Note that both of these poetry experts phrase their complaints in ways which imply that a defect in Process is responsible for a defect in Product. A process which doesn't take the semantics or pragmatics of language into account leads to a product which is less valuable because it doesn't make sense; a process which relies too heavily on template and repetition leads to a product which is seen as fatiguing.

A lack of coherence is not necessarily a problem for everyone. As mentioned, Dadaists and other poets are interested in generating nonsensical language (Balakian, 1974; Funkhouser, 2007). Similarly, Oulipians and conceptual poets (Goldsmith, 2009) are interested in inventing new poetry techniques, not in the content or quality of the poetry itself. This is only to mention a few of the rebellious schools of poetry for which a poem failing to make sense is not a problem. However, poetry that becomes mechanically repetitive will eventually fail to shock anyone. For this reason, many new media poets have turned their focus away from the generation of poetic text as such, and towards other uses of computers, such as multimedia poetry (Funkhouser, 2012).

Others look for ways to improve on mere generation. If defects in process lead to defects in product, then from both perspectives, a possible solution is to find new processes which transcend mere generation's linguistic limits. One such method is for a human to edit the output; the other, more ambitious but still in its infancy, is to improve generation through artificial intelligence.

3.1.3 Human Enhancement

The most obvious way for a human to enhance computer-generated poetry is to edit the poetry generator's output. Some poets argue that this invalidates the generator's usefulness (Carpenter, 2007) (and it definitely poses problems for testing any scientific hypothesis involving the generator, as Pearce *et al.* describe (Pearce et al., 2002)), but among other poets it is an established practice. John Cage, for instance, removed unwanted words from the output of his algorithms (Funkhouser, 2007). Computational text generation is seen by many poets as a 'jumping-off point' (Carpenter, 2007) from which they acquire raw material. While almost any poetry generation system involves a human programmer selecting the best examples of output for publication, we define the Human Enhancement category as poetry in which post-algorithmic human involvement goes beyond selection and into actually modifying the output text. (Human-computer co-creativity is not mentioned in Ventura's taxonomy, as Ventura is concerned only with judging the sophistication of the computer.)

Human enhancement is a common technique for poetic groups outside the mainstream. One such group is the Flarf movement. Flarf revolves around intentionally inappropriate, lowbrow language harvested from the Internet. One Flarf technique, 'Google sculpting', consists of searching the Internet for particular terms, taking phrases from the results (a form of Found Poetry)—and then recombining and modifying these phrases however the poet sees fit. The result is a distinctive, over-the-top poetics—as in this example, which was published in the prestigious magazine Poetry (Gordon, 2009):

> Oddly enough, there is a
> 'Unicorn Pleasure Ring' in existence.
> Research reveals that Hitler lifted
> the infamous swastika from a unicorn
> emerging from a colorful rainbow.

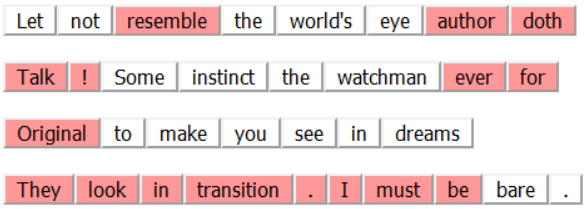

Figure 3.3: A screenshot of Gnoetry in action. Words in white have been selected by the human user as words to keep, while pink words will be replaced at the next generation step.

Even more interesting is Gnoetry, an application for interactive text generation which allows decisions to cycle between the computer and a human user. The computer generates poetry based on n-grams from a user-provided corpus. The user can click on individual generated words, deciding which words and phrases are worth keeping and which should be generated again. The computer then generates new phrases to replace those the user did not find satisfactory. This repeats as many times as the user would like. An example of what Gnoetry looks like during the generation process is shown in Figure 3.3.

Gnoetry turns generation into a dialogue between the human and computer. It is a dialogue that plays to the strengths of both participants: the computer can endlessly and tirelessly recombine words, while the human can use their superior judgment to select the most promising combinations. An online community, blog, and several chapbooks exist showcasing the work of various poets using Gnoetry. A favored technique, as with Markov chain poetry, is to mix together the styles of several contrasting works (eRoGK7 et al., 2011).

The notion of human and computer cooperation also appears in the development of creativity support tools. Kantosalo *et al.*, for example, design a system to help elementary school students write poetry, in which the computer suggests possible words in a magnetic poetry format (Kantosalo et al., 2014). The goal here, rather than showcasing the poetic skill of a computer, is to scaffold human learning of poetic skills by having a computer assist in some poetic decisions. A program like Gnoetry can also be used to test a human's poetic skill by requiring them to work within the computer's constraints.

3.1.4 Computer Enhancement

The other way to enhance computational poetry is to add more advanced concepts from computer science: not only generating words, but making sophisticated attempts to optimize the output. This set of methods comes not from the humanities but from scientists in the discipline of computational creativity. While a variety of AI techniques can be brought to bear on these problems, there are two main purposes to which these techniques are put. One is optimization of the system's poetic output on some metric; the other is connection of the generation apparatus to underlying knowledge about the world.

Nearly all computational poetry systems we analyzed begin with one of the mere generation techniques discussed above. Template generation is most common (Colton et al., 2012; Netzer et al., 2009; Toivanen et al., 2014, 2013, 2012), but the McGonagall system (Manurung et al., 2012, 2000; Manurung, 1999) uses a technique similar to context-free grammar. Barbieri et al.'s system uses an improvement on Markov chains (Barbieri et al., 2012), while DopeLearning (Malmi et al., 2015) is a found poetry system. What

distinguishes all these systems from mere generation systems is that something—optimization, a knowledge base, or both—is *added* to the basic technique. This addition can be incorporated into the mere generation step as guidance, or can be added after that step is complete.

Ventura's classification refers to the use of optimization metrics as Filtration, and the use of a knowledge base as Inception. Ventura considers Inception to be a higher level of creative sophistication than Filtration (Ventura, 2016). Ventura also provides an additional category, Creation, in which a computer's work is informed by direct perceptual processing. To the best of our knowledge, no poetry system has yet been constructed in this way. A few neural networks exist, which we will discuss in Section 3.2.3, which generate poems based on an image. This is partway towards Creation; but we have not seen any poems built by systems which directly perceive and comment on the world around them.

3.1.4.1 Data mining and knowledge representation

Merely generated poetry tends towards nonsense; whether desired by the programmer or not, this is a result of the computer's lack of real-world knowledge. By representing semantic facts, an artificially intelligent system can attempt to overcome this limitation. Just as Gnoetry puts human enhancement inside the generation process rather than adding it on after generation, a knowledge base can be put inside the generation process to guide its range of output. Creative humans, too, require knowledge about the world and their domain as groundwork for apparently spontaneous creative insights ((Beardsley, 1965; Kaufman and Baer, 2012; Rhodes, 1961; Sadler-Smith, 2015)).

One way of representing commonsense knowledge is to encode it in propositional logic. One experiment in this vein is Ruli Manurung's McGonagall system (Manurung, 1999). McGonagall uses the AI technique of chart generation which, given an input meaning, lexicon, and set of grammar rules, can exhaustively generate all grammatically correct representations of the input meaning. McGonagall generates poetry in a metrical form by encoding metrical constraints as grammar rules. When the input meaning is directly encoded by humans, this process results in metrically correct and very logically consistent poetry (Manurung, 1999):

> The cat is the cat which is dead.
> The bread which is gone is the bread.
> The cat which consumed
> the bread is the cat
> which gobbled the bread which is gone.

However, propositional logic alone does not solve the sense/nonsense problem. McGonagall's poetry only makes sense when the exact intended meaning of the poem is directly encoded by a human. When the system is allowed to construct its own propositions, it immediately lapses back into nonsense (Manurung et al., 2012):

> They play. An expense is a waist.
> A lion, he dwells in a dish.
> He dwells in a skin.
> A sensitive child,
> he dwells in a child with a fish.

It is not enough for a computer to be able to represent facts; to be anything other than nonsense, these facts must have some relation to real-world experience. Real-world propositions could be added to a system like McGonagall through semantic parsing of a corpus of human-written texts. Some systems have made good strides expanding their bank of propositions by taking advantage of existing semantic knowledge bases such as ConceptNet and WordNet (Ramakrishnan A and Devi, 2010; Soo et al., 2015; Agirrezabal et al., 2013; Gabriel, 2016; Misztal and Indurkhya, 2014; Oliveira and Alves, 2016). But this is most commonly seen in non-English languages.

What is more common in English is parsing of word associations. By calculating the co-occurrence of different words in a source text through methods like tf-idf or Latent Semantic Analysis, computers can gain a sense of which words are and aren't related to a given topic. This can guide the computer towards the construction of lines and sentences that relate to each other in sensible ways. It is important to note that word associations in this sense are different from the N-grams used to create a Markov chain. Markov chaining only measures words that appear immediately after each other. Word association does not take into account word order, but instead looks at what meaningful words are used in a given chapter, article, or paragraph which are different from the words in other units of text. In this way, word association mining gives a sense of what words are topically associated with each other or have similar meanings.

Toivanen *et al.*create poetry using word associations mined from Finnish Wikipedia (Toivanen et al., 2012) and news stories (Toivanen et al., 2014). Netzer *et al.* (Netzer et al., 2009) use a list of word associations from psychological testing, which they claim produces more intuitive, humanlike semantic connections than word associations from an encyclopedia. Combining these associations with syntactic templates results in plausible haiku (Netzer et al., 2009):

> cherry tree
> poisonous flowers lie
> blooming

Veale and Yao (Veale and Hao, 2011) search Google N-Grams for similes, which (as well as being a poetic device in their own right) contain implicit information about the properties of things in the real world. Their system then uses this implicit information to create new similes of its own. The Full-FACE system (Colton et al., 2012) modifies these similes to create full poems. Its poetry, while repetitive, is full of comparisons that make sense to a human (Colton et al., 2012):

> the wild relentless attack of a snake
> a relentless attack, like a glacier
> the high-level function of eye sockets
>
> a relentless attack, like a machine
> the low-level role of eye sockets
> a relentless attack, like the tick of a machine

3.1.4.2 Optimization.

The other mode of enhancement that computer science has to offer is optimization. Given some formal definition of the desired traits of a poem, a computer system can begin to, in some sense, think critically—testing different possibilities and choosing the ones which best fit its requirements. While this is a very

elementary form of critical thought, it is an important step towards true creativity on the part of computers: being able to understand one's own aesthetic and create work to match. In particular, some form of optimization is inherent in any creative system that uses a generation-evaluation loop.

Optimization techniques applied to generative poetry include stochastic hill-climbing search (Manurung et al., 2000), generate-and-test (Netzer et al., 2009; Colton et al., 2012), genetic algorithms (Gervás, 2013b; Manurung et al., 2012), constraint satisfaction (Toivanen et al., 2013), case-based reasoning (Gervás, 2001), dynamic programming (Yan et al., 2013), statistical machine translation (He et al., 2012), and recurrent neural networks (Loller-Andersen and Gambäck, 2018; Xu et al., 2018; Yang et al., 2018), the last of which we will look at in more detail in a later section. Each of these methods starts with a mere generation technique and a goal. The system evaluates how close the generated text is to meeting its goal, and then re-generates or modifies its text repeatedly until it meets the goal as closely as possible. Any mere generation technique can be combined with most optimization techniques. A few optimizations are designed for a specific mere generation technique, such as elementary Markov constraints (Pachet and Roy, 2011) and constrained Markov processes (Barbieri et al., 2012), both of which allow a programmer to impose structural requirements onto a Markov model.

What goals do these poetry systems work towards? Many concentrate on basics such as meter, rhyme, and grammaticality (Manurung et al., 2000; Netzer et al., 2009; Toivanen et al., 2013) or fitting generated text to a rhythm (Gonçalo Oliveira, 2015). However, a more exciting possibility is setting goals for the subject matter, emotions, or artistic style of a poem—traits which can be measured through natural language processing techniques. It's important to note that if a computer is judging the subject matter of its generated text, then optimization has been combined with knowledge representation: without some form of implicit or explicit knowledge representation, it would be impossible to process a text's semantic traits. A great deal of Computer Enhanced poetry uses both techniques, and thus would be considered Inception under Ventura's model.

As an example, ASPERA (Gervás, 2001) includes mood as one of its constraints. DopeLearning (Malmi et al., 2015) generates rap lyrics according to a variety of textual measures, including the maximization of internal and multisyllabic rhymes, and uses a neural net to represent semantic content. The Full-FACE system (Colton et al., 2012) chooses between possible goals, including relevance to the poem's topic, emotion, and some stylistic measures such as 'flamboyance' (the number of unusual words). Yan *et al.*(2013) maximize the relevance, importance, and coherence of individual words in each poem, while He *et al.*(2012) optimize coherence based on mutual information.

A particularly sophisticated optimization technique is that of Gervás's WASP system, which creates Spanish-language poetry through an evolutionary approach involving many modules which create, evaluate, and edit candidate poetic texts (Gervás, 2013b, 2016). "Babblers" are the system's mere generation layer, which create baseline text using an N-gram model. "Reviser" modules can make specific changes to the text aimed at correcting the problems detected by a "judge". This is the closest to a human poet's revision process that we have encountered in the field of generative poetry. The WASP system's optimization goals include poem length, verse length, rhyme, stress pattern, similarity to source (poems that plagiarize the source material are rejected), and fitness for a specific poetic form. Some of WASP's poems are successful enough to have been published in a book about generative poetry (Gervás, 2016).

3.1.4.3 Neural networks

A final Computer Enhancement method, recurrent neural networks, is complex enough to deserve a discussion of its own. Neural networks are a machine learning technique in which the computer, by mimicking the activities of human neurons, learns its own high-level encoding of input data. Like a verbal equivalent of Google's Deep Dream (Mordvintsev et al., 2015), neural networks trained on poetry can generate a high level, general encoding of the patterns in the poetry they have seen, and use that encoding to generate new work of their own. A key property of neural networks, however, is that their encoding is implicit. Information is contained in the strengths of thousands of connections between neuron-like nodes, rather than existing in a form which can be easily communicated to humans. Because neural networks are "black boxes" in this sense, it takes some thought to discern where in our taxonomy a neural network-based poetry system would belong.

Neural networks have been used to generate poetry in several languages, but are most prominent in the generation of Chinese classical poetry. A common form of neural network in poetry generation is a recurrent neural net or RNN, in which the connections between neurons can form cycles (Ghazvininejad et al., 2016; Goodwin, 2016; Zhang and Lapata, 2014); it is also common to use a paired encoder-decoder model (Ghazvininejad et al., 2016; Wang et al., 2016b; Yi et al., 2016; Xu et al., 2018; Yang et al., 2018). To this framework some researchers add features such as long short-term memory (Wang et al., 2016a; Goodwin, 2016) and an attention model (Wang et al., 2016a).

To constitute Computer Enhancement in our taxonomy, a neural network would need to either contain implicit knowledge representation or to be optimized on some metric. Most poetry neural networks are optimized during their training phase: the weights of connections in the network are repeatedly adjusted in order to produce output that most closely meets some metric chosen by the programmer. The most common such metric is cross-entropy, which is a statistical measure of the similarity of character distributions between the output and a source corpus (Wang et al., 2016a; Yan, 2016; Zhang and Lapata, 2014). The decoding portion of the network, which translates from connection patterns back into human language, can contain additional constraints for goals like rhyme, topicality, and tone pattern (Ghazvininejad et al., 2016; Zhang and Lapata, 2014).

The argument regarding knowledge representation is slightly less obvious, but we would argue that if a neural network contains global information about word associations, then it contains knowledge representation, just as poetry systems do which acquire this information in another way. In practice, many systems use pre-existing word association corpora to process user-defined keywords before feeding them as input into the neural network (Ghazvininejad et al., 2016; Wang et al., 2016b; Zhang and Lapata, 2014).

However, it is possible to imagine a naive application of a neural network which would be Mere Generation. Imagine a neural network which is trained to produce the next character of a sequence based on the last N characters. It uses an n-gram representation for input which is similar to that of a Markov model, without any global information, and instead of having its connection weights adjusted to optimize its output, it is allowed to discover the patterns in these characters on its own. Such a neural network would contain neither optimization nor knowledge representation. It would be essentially just a very complicated Markov model, and one would expect its output to be similar to that of a Markov model. Indeed, simplistic applications of recurrent neural networks to poetry in English can strongly resemble the output of a Markov model (Karpathy, 2015):

> O, if you were a feeble sight, the courtesy of your law,
> Your sight and several breath, will wear the gods

With his heads, and my hands are wonder'd at the deeds,
So drop upon your lordship's head, and your opinion
Shall be against your honour.

The Chinese systems which do incorporate Computer Enhancement produce strikingly convincing output. At least one Chinese system (Wang et al., 2016b) produced results that non-expert Chinese speakers could not distinguish from classical Chinese poems. Another (Wang et al., 2016a) produced poems that human experts rated as more "poetic" (meeting the constraints of the chosen poetic form) than human poetry, although it did not meet human standards for fluency or meaningfulness. An expert in Chinese poetry whom we queried agreed that the best published Chinese poems looked similar to something a human could write (Lingentfelter, 2017).

The neural network aspect of these systems is often combined with other mechanisms to ensure that they meet all their constraints. In addition to the word association tools already mentioned, Zhang and Lapata's system (Zhang and Lapata, 2014) adds statistical machine translation to improve the poems' coherence. Also notable is the use of something resembling planning mechanisms in many systems. Several systems generate an outline of their poem, either using keywords (Wang et al., 2016b; Xu et al., 2018; Loller-Andersen and Gambäck, 2018), separate layers (Zhang and Lapata, 2014), or a separate neural network (Yan, 2016), before feeding this outline into the central RNN to generate actual lines. This is one of the few areas in which we have seen computers follow the path of a human artist from idea to structure to finished work (see Section 2.4.3).

Few systems in languages other than Chinese have replicated these results. An exception is the recent work of Goodwin, who trained an LSTM-RNN (recurrent neural network with long short term memory) on a corpus of contemporary English poetry (Goodwin, 2016):

And still I saw the Brooklyn stairs
with the shit, the ground, the golden haze
Of the frozen woods where the boat stood.
When I thought of shame and silence,
I was a broken skull;
I was the world which I called it...

Goodwin reports that his results with the LSTM-RNN were uneven, and overall semantic coherence is still an issue; the poems are still arguably not 'about' anything. But in our opinion, the stylistic rendering is good enough to give the impression of images and moods that a human reader can understand, and of a writer with an ear for rhythm and phrasing at a level no other English poetry system we are aware of has approached. Goodwin's work has also received a favorable reception from human poets not involved in generative poetry (Goodwin, 2016), and he has worked on other generative language since, such as a Beat poetry project (McDowell, 2017).

3.1.4.4 Combined computer and human enhancement

It is possible to combine both computer and human enhancement. In a combined enhancement system, both the workings of an AI system and the judgment of a human are involved. One example of this type of

system is Barbieri et al's (2012), which generates lyrics using a constrained Markov process, but allows a human to perform the final step in the generation by choosing each verse from a set of 5 computer-generated candidates.

The science of computational creativity is in its infancy, and these are small steps compared to the complex goal-setting process of skilled human poets. Nonetheless, further refinements in goal specification and knowledge representation could produce truly interesting generative poetry.

3.1.5 Separation of generative poetry communities

Researchers working on Computer Enhanced poetry, and artists working on Human Enhanced and Merely Generative poetry, often have little communication with each other. Computer Enhanced poetry, with the exception of recent work on recurrent neural networks, is typically created by members of the computational creativity research community. These researchers are interested specifically in the computational implementation of creative decisions, and thus, Merely Generative poetry is of little interest to them—in fact, for some researchers, mere generation is anathema to their stated goals (Ventura, 2016). Previous work by such researchers categorizing generative poetry, such as Gervas' taxonomy (Gervás, 2002), make no mention of work outside the Computer Enhancement category at all. While this is logical for researchers with a technical focus, a lack of awareness of generative poetry created by poets and artists outside their field prevents computer science researchers from making compelling arguments about what their work contributes to the field artistically.

Some cultural critics and researchers in the humanities do a better job at presenting a broad perspective. Roque's work (Roque, 2011) places computational creativity projects in context with similar projects from outside that field. However, other important and detailed reviews, such as Funkhouser's (2012), make no mention of projects created by computer scientists. As Cook and Colton (2018a) explore, this lack of connection between compatible generative creativity communities is typical across all computational creativity domains—and even between computational creativity and other subdomains of computer science, such as the AI and machine learning communities.

Greater communication between computer scientists and poets would benefit both sides. There are many ways to gauge a computational creativity project's success, but producing credibly artistically successful work is one of the goals of most such projects. Cultural critics are ideally placed to provide feedback and critical analysis of computationally creative work on these grounds, provided there is communication between them and computer scientists and awareness of each other's existence. (Success with cultural critics is important from the Press perspective, and cultural critics also have informed comments to make from the Product perspective, making communication with such critics important for researchers who are interested in these perspectives.)

Meanwhile, as explored above, computational creativity researchers potentially have something to give back to the humanities - by widening and deepening the scope of what generative poetry systems can achieve.

One recently emerging community which does bring both sides of the generative poetry community together is Google's Artists and Machine Intelligence program, including an informal conference which was initiated in 2016 (McDowell, 2016). This program is explicitly designed to bring together computer scientists, artists, neuroscientists, and psychologists to work together on generative art, including poetry. Institutions such as the University of Helsinki also have ongoing poetic collaborations between computer

scientists and working artists (University of Helsinki Computer Science Department, 2016). The #CreativeAI community and National Novel Generating Month (in which a 50,000-word computer-generated literary work is prepared in a month, often using poetic techniques) are examples of other adjacent communities into which researchers from computational creativity are making inroads (Cook and Colton, 2018a).

With such multidisciplinary efforts becoming more mainstream and widespread, there is a great deal of reason to hope that researchers on both sides of the art / science divide will continue to come together to make ever more interesting generative poems.

3.1.6 Generalization and comparison with music

Our taxonomy is more interesting if its principles can be generalized to other domains. An obvious example would be other domains of digital poetry, such as those described by Douglass (2014) and Funkhouser (2012). Certainly many such poems combine the kind of generation that is of interest to us with hypermedia techniques, which can themselves be seen as a form of either human or computer enhancement, applied to the poem's mode of presentation rather than to its words. Nick Montfort and Stephanie Strickland's 'Sea and Spar Between', for example, uses a handcrafted grammar (a mere generation technique) to combine phrases from Emily Dickinson and Herman Melville's work, and displays these phrases using a combinatorial framework which could only be accomplished using the calculating power of a computer (Montfort and Strickland, 2010).

However, in our specific research program we are more concerned with whether our taxonomy is generalizable to other domains of computational creativity. To investigate this question, we will look very briefly at the field of generative music. Our taxonomy—Mere Generation, Human Enhancement, and Computer Enhancement—does apply to music, with only small modifications.

3.1.6.1 Mere Generation.

Like poetry, music can be composed using a Markov model (Collins and Laney, 2017; Eigenfeldt, 2015; McDonald, 2017; Percival et al., 2015; Kalonaris, 2018). In fact, Markov models are thought to represent musical style particularly well, at least over short sequences (Pachet and Roy, 2011). Other ways of generating music include cellular automata (Zareei et al., 2015), a random trajectory in a directed graph (Scirea et al., 2015), or even producing music from a visual image as if it were a spectrogram (Heep and Kapur, 2015). What these techniques have in common is that, as with mere generation in poetry, the creative choices made by the computer are either random or arbitrary. The program simply produces notes according to its rules.

3.1.6.2 Human Enhancement.

The Human Enhancement category in music is most noticeable in jazz improvisation. Computer Improvisation systems are built to play alongside human improvisers. They need to process human musical input in real time and respond to that input with novel sequences of notes (Bown, 2015; Kalonaris, 2018) or with an appropriate synthesis of previously recorded notes and chords (Pachet et al., 2013). A more avant-garde alternative is the creation of digital instruments. An instrument's responses to human input

can be non-obvious or shifting, which results in unpredictable interactions between the musician and the instrument (Zappi and McPherson, 2015). Like Gnoetry, improvisation systems and digital instruments can create new works through interaction which neither the computer nor the human could have created on their own.

3.1.6.3 Computer Enhancement.

As with poetry, the output of mere generation can be fit to hard and soft optimization constraints using techniques such as answer set programming (Opolka et al., 2015), genetic algorithms (Scirea et al., 2015), principal component analysis (McDonald, 2017), self-organizing maps (Kalonaris, 2018), constraint satisfaction (Videira et al., 2017), or elementary Markov constraints, which were in fact specifically designed with music in mind (Pachet and Roy, 2011; Papadopoulos et al., 2016). Such constraints can be rooted in music theory (Lattner et al., 2018; Opolka et al., 2015; Scirea et al., 2015), or a specific goal, such as making cover songs incorporate features of the original (Percival et al., 2015) or imposing a melodic contour chosen by the user (Pachet and Roy, 2011). Music can also be generated using neural nets, which range from mere-generation-like results that produce a noisy copy of the input, to some of the most sophisticated existing results (Lattner et al., 2018; McDonald, 2017; Sturm and Ben-Tal, 2017). It is harder to make a case for knowledge representation in music generation, due to the non-representational nature of music, but it is true that systems can use machine learning to absorb information about musical structure and style, just as some poetry systems use machine learning to mine word associations (Kalonaris, 2018); or the above-mentioned systems which constrain output based on music theory could be seen as encoding music theory as a form of knowledge. We also see music systems in which human and computer enhancement are combined. Systems can, for example, improvise with human partners and use machine learning to optimize that improvisation (Bown, 2015).

Note that, if a system "hears" in real time by processing audio generated by humans or otherwise, and takes this hearing into account when producing its own output—as all co-creative improvisation systems do—then it is possible to argue that the system meets the requirements of Creation, the highest level in Ventura's (2016) hierarchy.

3.1.7 Conclusion

In our taxonomy, there are three areas of work in generative poetry: a mere generation technique, human extensions to the technique, and computer extensions to the technique. We believe that when critics such as Funkhouser (2012) declare that generative art has reached a plateau, it is because they are looking only at mere generation and not at the more powerful computational techniques from outside the field which can enhance it. Artistic optimization and knowledge representation techniques have not yet reached their full potential, but they have the power to push generative poetry forward towards the kinds of sense and style that are currently lacking. Far from being a played-out form, generative poetry is just getting started.

3.2 State of the art in Computer Enhanced poetry

In Section 3.1 we focused on illustrating our taxonomy, giving enough examples to properly define and instantiate each category. In this section, we look deeper into current trends in the Computer Enhancement category of poetry. Our focus here is primarily on academic work by computer scientists, with a bias towards more recent work. We describe work in different types of optimization, knowledge use, and neural network-based poetry. We then focus in more closely on a category that is especially relevant to the work we will do in Chapter 5: Twitter and social media-based poetry, including non-Computer Enhancement examples of these forms.

It is important to note that not all interesting poetry is easily categorizeable as Mere Generation, Human Enhancement, or Computer Enhancement, because we do not always have access to the relevant information about how the poetry is created. Thus, famous computer systems such as RACTER (Chamberlain, 1984) and the Cybernetic Poet (Kurzweil, 2001) are hard to discuss because their inner workings are proprietary. We focus on systems whose authors have published a discussion of their process and architecture.

3.2.1 Optimization / Filtration

As mentioned previously, the two main areas of work in Computer Enhancement are optimization and the use of knowledge bases. As optimization occupies a lower level of Ventura's (2016) taxonomy, we will address it first. Methods of optimization differ widely across the field, but can be split for our purposes into three broad categories. One group of systems uses constraint satisfaction in order to methodically generate a poem that best fits a set of specifications for the desired output. Another uses genetic algorithms and other forms of stochastic search to explore a broader range of possible poems with a more loosely defined goal. Finally, there are systems that make specific edits to their output in order to ameliorate specific problems, often under the auspices of one of the constraint satisfaction-based specifications from the first group. We will explore each of these in turn.

3.2.1.1 Constraint Satisfaction

In constraint satisfaction, the essential idea is that candidate poems, or parts of poems (such as lines, or candidate words to fill a template), are created using a mere generation technique before being assessed by how they meet certain computationally calculable constraints. The poem is then selected, or constructed out of the appropriate parts, which best fits the group of constraints.

Rashel and Manurung's Pemuisi system (2014b) generates poetry in Indonesian in this manner. They use template-filling as their mere generation technique and a constraint solver taking into account the number of lines, number of words, number of keywords, number of syllables, and rhyme. Rashel and Manurung evaluate Pemuisi using a modified Turing test combined with a questionnaire on the poems' structure, diction, grammar, unity, message/theme, and expressiveness. However, they do not evaluate their results for statistical significance.

Toivanen *et al.* (2013) similarly generate English poetry using template-filling and a constraint solver; in this case, their constraint solving technique is Answer Set Programming. Their constraints include poem length, number of lines, number of words per line, rhyme, a limit to the number of times a particular

word can be repeated, and some grammatical constraints. They do not formally evaluate their system. An example is included below:

> Music swells, accent practises, traditionalism hears!
> Her devote narrations bent in her chord:
> - And then, vivaciously directly a universe
> she disappears!
> An anthem in the seasons of record!

Barbieri *et al.*(Barbieri et al., 2012) use a Constrained Markov Process in English, which fits a Markov chain-like process to constraints. They use rhyme, meter, and semantic relatedness to a title as their constraints (we will discuss semantic relatedness further in the next section). They also use lyrics from specific human songwriters as the base text for the Markov chain, so as to produce text in the style of an existing artist. There is a Human Enhancement element to their system, in that the system produces several candidate lines and a human user chooses between them. Barbieri *et al.* evaluate their system by asking human judges to rate the poems on syntactic correctness and semantic relatedness to the title; the poems generated by a Constrained Markov Process (such as the one below) outperform pure Markov poems as well as poems generated through pure constraint solving.

> There is a note in his eyes
> He backs the beat of the key
> Down the song in my eyes
> You back the beat of the sea

McGovern and Scott's system EloquentRobot (2016) generates lines from an English Markov model and selects those that meet the system's constraints, which it learns based on studying the structure of poems in a corpus of haiku and limericks. These constraints include appropriate number of syllables, rhyming or not rhyming with previous lines as appropriate for its position, and ending with a word from an appropriate part of speech. The Markov model is trained either on a general corpus or on specific corpora, such as political speeches or lists of insults. EloquentRobot is evaluated based on how well it matches the forms of the types of poems it has learned, based on the results of a naive Bayes classifier - a form of automated typicality judgment. Its output can be classified as either haiku or limericks with 100% accuracy. Of course, it is relatively easy to distinguish a poem with five lines from a poem with three.

> forget why you entered a room
> scratches on your arm and assume
> May you receive nude
> you be so pursued
> when your parents enter the room

Singh *et al.*'s MABLE system (2017) converts the English stories generated by the MEXICA narrative system into lyrics for ballads. Odd-numbered lines are taken directly from MEXICA, while even-numbered lines are chosen from candidates generated by a Markov model. The system chooses lines that match the meter and rhymed ending of the previous line, and then selects from those lines the ones whose sentiments

best match the intended sentiments of the story. A final module changes words where necessary to ensure that the ballad is told in third person. MABLE is evaluated by a group of human judges who rate its output, along with the output of other systems and ballads by humans, on plot coherence, emotional engagement, and overall rating. MABLE is rated as more coherent and better overall than the other two systems in the study (Full-FACE and a rap lyrics generator), but does not approach any of the ratings of the human lyrics.

> The priest was born under grace of the great god
> And evolving from the shadows lifted
> The lady was an inhabitant of the great city
> But just remember there's a sign of intensity

Tobing and Manurung (Tobing and Manurung, 2015) use chart generation—a variation on context-free grammar—as their mere generation technique, writing in English, and attempt to fit the grammar's output to constraints using dynamic programming. Their constraints are purely metrical. Even with these simple constraints, their system is costly to run. Replacing full chart generation with a greedy algorithm reduced runtime, but also reduced quality (in the author's opinions, at least; they do not formally evaluate their system). Examples with and without the greedy algorithm are reproduced below.

> 1.
> Ask in french surface
> Call her years, check her
> Think were toy tennis
> Skid in chase, land her
>
> 2.
> Is she
> Court were full even
> She were take couple
> She were a woman

According to Oliveira (2017b), Tobing and Manurung's system illustrates a characteristic problem with constraint solving: with many constraints or many possible solutions to the constraints, these systems become computationally cumbersome. However, this depends on the method of constraint solving, and other systems using constraints have not necessarily encountered this difficulty.

The InkWell system (Gabriel, 2016) uses a more complex constraint satisfaction system to write English haiku. A user inputs seed words, and InkWell adds more seed words based on a randomly selected section from a large corpus of (non-poetry) human source texts. A vector-like structure is constructed from all of the seed words, and is combined with a vector representing the important words from an existing haiku, which is used as a template. (The specifications in the template, such as "noun-animal", are more semantically well-defined than typical templates in computer-generated poetry.) InkWell then assigns random weights to 32 haiku-writing constraints, including sophisticated constraints such as "personality traits", which might be represented using links in WordNet or techniques such as bag-of-words machine learning to infer the characteristics of different writing styles. InkWell can also be given custom constraints,

such as a request to use unusual synonyms or to use n-grams that are like or unlike those of a particular human writer. It tries many words and phrases to fill the template before selecting the best ones based on this combination of constraints.

InkWell was evaluated by bringing eighteen of its poems to a professional writers' workshop without disclosing that they are computer-generated. The participants in the workshop took the poems seriously and praised many of them, though they criticized others (as is typical at a serious workshop!). Although some writers in the workshop were aware that the author had experience using computers to generate text, they stated a belief that the poems were probably written by the human author, and were surprised when the truth was revealed. Moreover, contrary to some criticisms of Turing-style tests, the writers praised not only InkWell's ability to write a "good" poem, but also its ability to surprise them.

> awake in the dark
> the edge of the water can
> spread in your presence

3.2.1.2 Genetic and Stochastic Algorithms

There can be many reasons to move from constraint satisfaction to genetic and stochastic algorithms. As Oliveira (2017b) mentions, some researchers find exhaustive search too slow and the constraint space too large; a genetic algorithm can allow exploration of the most promising paths in a space without being exhaustive. Other researchers might find that a genetic algorithm allows goals to be specified more broadly than the goals of a constraint solver.

Rahman and Manurung (2011) use the SPEA2 evolutionary search algorithm to evolve poetic representations of a semantic target in English, using a variation on context-free grammars as their mere generation technique. Their evolutionary algorithm chooses the best poems of each generation using a multiobjective function including grammaticality, meter, and similarity to the target semantics. They evaluate their system by showing that its output meets its own multiobjective function better than the output of prior experiments in the same vein.

ePoGeeS, a rare example of Computer Enhanced poetry that is not published as computer science research, uses both Human Enhancement and this type of search algorithm to generate English poems (Roque, 2011). ePoGeeS uses class-based n-grams, a variation on Markov models: instead of predicting which character word will come after the current one, a class-based n-gram predicts the next part of speech. Thus, class-based n-grams occupy a middle ground between word-based Markov models and context-free grammar; they take grammar into account somewhat, without fully modeling language's recursive structure. The user controls the text from which the n-grams and words are taken, as well as the constraints on the form of each poem, such as rhyme, number of lines in a stanza, and number of words in a line. The user also sets variables to determine the desired *phonemic* properties of the poem: for example, emphasizing a certain letter or sound. As Roque points out, this is consistent with the "sound poetry" created by Dadaists and Futurists. The phonemic properties are the part of the poem to which Computer Enhancement is applied. ePoGeeS can either use a random search, in which many lines are randomly generated and the one with the best phonemic properties is selected, or a stochastic beam search in which several best lines are kept at each generation, and the next generation is created by changing a word in each line. The following example is generated by ePoGeeS optimizing for back phonemes (i.e. sounds created in the back of the mouth), and using Shakespeare sonnets as a base:

For thou thine eyes to whom all my
song: and water for myself mine eye more aye
thy worth to whom thy worst all forwards
for her treasure thou truly write good allow.

Kirke and Miranda (2013) use a different method to stochastically develop English poems. Rather than selecting and recombining the best poems at each generational step, their MASTER system instead simulates a society of artificial poets who are influenced by each other. Each simulated poet has an emotional state, which is initialized by the programmer but can change as a result of the output of other poets. The only explicit goal of each poet is that it use words which reflect its current emotional state. Kirke and Miranda do not formally evaluate MASTER, although they state that their intent is to imitate modernist poets such as Kurtz Schwitters who use repetitive, non-humanlike syntax:

quiet book comet and fornicate quiet
tourist ignite live quiet quiet book comet and wine ejaculate
and boring welfare fire with fornicate
quiet book comet and rape boring fatigued sadness it quiet
tourist ignite live quiet quiet book comet and wine ejaculate and hysterical rage
collaborations fornicate quiet tourist
ignite live quiet
quiet book comet and wine ejaculate and boring
welfare fire with hysterical explosion sensations
explosion explosion provoked explosion
explosion prizes quiet quiet

Misztal and Indurkhya (Misztal and Indurkhya, 2014) generate English poems based on an unusual method. A user-provided source text is evaluated for its emotional state and most pertinent phrases. The poetry system generates a pool of words by searching for hypernyms, antonyms, and other words related to the selected user-provided phrases (a simple form of knowledge representation). The words are then extended into lines using a complex form of template generation, with different modules containing instructions for creating different sorts of phrases out of the topical words and the pool of words with specific relations to those words. The lines are constrained for number of syllables, grammatical form, and tense. A control module then rates all the lines and puts together a poem balancing two optimization constraints: use of the lines with the highest scores, and use of lines that were generated in a variety of ways.

I knew the undisrupted end
I was like the various end
As deep as a transformation
O end the left extremity
Objective undisrupted end
I hated the choleric end
O end the dead extremity

Misztal and Indurkhya evaluate their system using the cumulative interpretation of the FACE model (Colton et al., 2011), stating that it satisfies all of the model's requirements except perhaps framing information. They also subjectively evaluate their system using Manurung *et al.*'s criteria of Grammaticality, Meaningfulness, and Poeticness (Manurung et al., 2012), concluding that the system could improve with the use of more stylistic constraints and better awareness of the words' context.

3.2.1.3 Targeted edits

Constraint satisfaction approaches try different options before picking the best-performing ones, and evolutionary approaches combine aspects of the best-performing ones to see if they can improve more. Some systems, a step further, make targeted edits which seek to ameliorate a specific problem, much like a human poet who looks at a poem draft critically and fiddles with the parts that aren't working. Indeed, Gervás (2013a) specifically promotes this type of edit, referring to mere optimization as "opportunistic".

Díaz-Agudo *et al.* (2002), with the Spanish COLIBRI system, have the goal of personalizing a poem for an occasion. Their mere generation method is template-filling, substituting words belonging to the appropriate parts of speech with words provided by the user. They use a case-based reasoning package called CBROnto to ensure that the poem still fits rhyme and metrical constraints, and to edit further if it does not. COLIBRI can identify several types of problem that could cause a poem not to meet its constraints, and has a specific strategy for attempting to fix each one. However, COLIBRI does not reason about the semantics of its chosen words. Díaz-Agudo *et al.* do not formally evaluate COLIBRI.

Gervás's WASP system (Gervás, 2013a,b, 2016) builds on the editing process of COLIBRI. As mentioned in Section 3.1.4.2, WASP contains independent modules for Spanish text generation ("babblers", which use a Markov model), evaluation ("judges"), and editing ("revisers"). Various versions of WASP can revise for rhyme, stress pattern, excessive similarity to the source text, sentence length, verse length, plausibility of sentence ending (e.g. not ending on a word like "and" or "to"), unacceptable foreign words (i.e. penalizing for words that cannot be easily fit into a Spanish metrical pattern), and control over repetition. Revisers can attempt to remove these problems by various strategies such as replacing individual words, changing the position of line breaks, or adding a new sentence. However, Gervás states that revisions made by randomly replacing a word are more likely to harm than help.

Gervás (2016) also expresses a desire to optimize for topicality, but so far, a way of employing judges and revisers for this purpose has not yet been found. Instead, WASP is constrained for topicality by constraining the source text.

In the 2013 version of WASP, the result is a verse several sentences long into which appropriate line breaks are inserted by a "poet" module. In the 2016 version, the results are single sentences, usually a line or two long, which are then put together with other lines into full poems by another module. Only the 2013 version of WASP has been seriously evaluated: in the case of that version, it was given a de facto Press evaluation by the fact that some of its poems were printed in a book about computer-generated poetry (Gervás, 2013a).

3.2.2 Knowledge representation / inception

There are both strong and weak uses of knowledge representation in poetry, as the word "knowledge" can refer to several things. Not every use of a knowledge representation algorithm counts as knowledge

representation for our taxonomy's purpose. For instance, Díaz-Agudo *et al.* (Díaz-Agudo et al., 2002) use a case-based knowledge representation ontology to construct their poems, but the knowledge is solely about rhyme, meter, and parts of speech. For our purposes, a poem belongs to the knowledge representation category if it is able to gather, represent, or reason about *semantic* information.

This information can be gathered in several ways. A human can hand-code a semantic representation of a specific poem. An existing, general-purpose knowledge base can be used, such as WordNet or ConceptNet. Semantic information can be mined using implicit word associations in a general-purpose corpus such as the Wikipedia. Or some form of bespoke knowledge collection can take place. We examine examples of each of these in turn.

3.2.2.1 Hand-coded knowledge representation

As discussed in 3.1.4.1, Manurung's (2000; 2012) McGonagall system creates English poems out of small hand-coded semantic propositions. Often the purpose of these representations is to try to reproduce a specific existing poem (Rahman and Manurung, 2011; Manurung et al., 2012). However, when McGonagall is allowed to generate its own propositions, it creates lines that have nothing to do with the human meanings of words. Therefore, McGonagall serves as a motivating example for the rest of the systems in this section. The use of data mining or of a large general-purpose knowledge base is intended as a means for the system to obtain new propositions without divorcing itself entirely from the meanings of words.

3.2.2.2 General-purpose knowledge bases

General-purpose knowledge bases are those that explicitly encode semantic knowledge in a machine-readable form. The difference between these and the hand-coded knowledge bases mentioned above is a matter of scale. Most knowledge bases such as WordNet are hand-coded by humans, but are meant for a general purpose and to apply to every sufficiently common word in the language, and are then released to the public or to other researchers; the researchers using these knowledge bases to create poetry are not the same researchers who created them. However, the potential utility of such knowledge bases for poetry is clear.

PoeTryMe (Oliveira, 2012) uses CARTÃO, a Portuguese knowledge base containing information about types of relations—such as hypernymy, part-of, causation, and purpose—between different words. PoeTryMe searches for uses of each of these relations in existing Portuguese poetry, and then uses the syntactic templates extracted from these poems to define rules about how to use each type of relation in a line. With a context-free grammar as the mere generation technique, and a group of seed words whose CARTÃO relations are meant to be used in the poem, PoeTryMe generates lines and poems based on these rules, and scores them according to their number of syllables and rhyme. Oliveira does not formally evaluate PoeTryMe, but notes that it often fails to meet its own rhyme and metrical requirements.

Agirrezabal *et al.* (2013) create poetry though template-filling in Basque. The templates are taken from existing poems. They try several ways of filling the templates, and evaluate their results with a modified Turing test. The poems getting the best results are changed as little as possible from the originals: only nouns are changed, and only to their antonyms or hypernyms, based on the use of Basque WordNet. Misztal and Indurkhya's previously mentioned system (2014) similarly uses English WordNet to constrain the choice of words for template filling.

Das and Gambäck's (2014) system co-creatively generates lines in Bengali to follow a line input by a user. The system uses a support vector machine to predict the syllable sequence pattern of the next line, and candidate output words matching the pattern are selected from a syllable-marked word list. The system is constrained to only use words that are directly related to at least one of the input words in the ConceptNet knowledge base. Since the equivalent of ConceptNet does not exist in Bengali, the words are first automatically translated into English, then looked up in English ConceptNet, then translated back. Das and Gambäck's system is evaluated through active testing by three expert and five non-expert evaluators, who found that the system did reasonably well at creating good rhyme patterns, but had more mixed results in grammar and meaning.

3.2.2.3 General-purpose data mining

A researcher might, for many reasons, choose to mine semantic data from a large corpus instead of using a hand-coded knowledge base. One might be interested in topics (such as the news) or types of relations that are not present in such a knowledge base, or one might be interested in data mining for its own sake, as a process that gives a little bit more autonomy to the poetry-writing system. Data mining from a large, general-purpose corpus enables the system to determine its own semantic representations which still have some connection to the human meanings and connotations of each word.

Wong and Chun (Wong et al., 2008) present an English found poetry system that uses light knowledge representation. Their system begins with a condensed list of keywords based on the 500 most common words in haiku, and uses a blog search engine to compile a list of candidate sentence fragments containing those words. The knowledge representation aspect comes into play when choosing sentence fragments to compile into poems. Using a simple vector space model, the system chooses the most semantically related fragments. This model is based on search engine results rather than, as is more common with vector space models, the full contents of a large corpus such as the Wikipedia. Wong and Chun do not evaluate their results except by measuring the cosine of the angle between its lines in the vector space.

> The snowy mountains
> Search field
> Of the honeymoon night

Toivanen *et al.* (Toivanen et al., 2012) use knowledge representation to constrain the choice of words for template-filling. Rather than a formal knowledge base, Toivanen *et al.* use a background graph mined from the Finnish Wikipedia. A single word is given by the user as a topic, and additional words are chosen based on their log-likelihood of co-occuring in the Wikipedia with the topic word. This system is evaluated by a group of non-expert judges who evaluate the poem on several criteria: whether it is a poem or not, as well as its typicality, understandability, quality of language, evocation of mental imagery, evocation of emotions, and how much the judge likes it. Human poems were rated significantly higher than the system's poems on every criterion, but the differences were not large and the ranges of scores often overlapped. The biggest difference between the human- and computer-generated poems was in their understandability.

A later version of Toivanen *et al.*'s system, P.O.Eticus, is modified to create poems on specific topics from the English news (Toivanen et al., 2014). In this version, topics are generated not based on words, but on news articles. The log-likelihood with which words are chosen to fill the templates is calculated

using a model that gives higher values for words that co-occur frequently in the chosen news article but not in the Wikipedia background corpus. Toivanen *et al.* do not formally evaluate P.O.Eticus, but instead make the interesting choice of providing 18 uncurated poems at the end of their paper so that individual readers can "decide for themselves".

Tobing and Manurung (2015) and Rashel and Manurung (2014b), like Toivanen *et al.*, use specific news articles in English and Indonesian as a source of topical semantic information. Barbieri *et al.* (2012), discussed above, use links in the Wikipedia (rather than the full text of Wikipedia articles) to determine semantic relatedness. Ramakrishnan *et al.* (2010) use a hybrid approach, referencing a hand-coded ontology for Tamil nouns and verbs but mining an unnamed text corpus for adjectives. Netzer *et al.* (2009) take a similar hybrid approach in English. A topic keyword is input by the user, and is clustered with words that are associated with it in a hand-coded psychological word association database. A template is then selected, and phrases that fit the template and contain one of the words from the topic cluster are retrieved from Google N-Grams. Thus, different knowledge representation approaches are taken at different stages of creation.

Droog-Hayes and Wiggins' IDyOT system (2015) applies semantic data mining to an English Markov-based generator. A vector space model is used to identify clusters of related words from a large general-purpose corpus. The Markov model's weights are adjusted so that it is biased towards words that are close to the previous word in the vector space model. IDyOT is evaluated by displaying pairs of haiku to human judges—one generated by the pure Markov model, and the other generated using the vector-space-biased weights—and asking which haiku is more meaningful. The poems generated using vector-space-biased weights are, on average, significantly more meaningful.

The Poet's Little Helper system (Astigarraga et al., 2017) uses Latent Semantic Analysis, a vector space model applied to a general-purpose corpus, and combines sentences from the corpus (that is, using found poetry as its mere generation technique) based on their closeness in the vector space as well as rhyme and metrical constraints. Poet's Little Helper can be applied to a corpus in any language, so long as rhyme and syllabification information for that language is provided, but is initially used in Basque. It also does exploratory analysis of a corpus to inform the user how easy or difficult it will be to use the corpus to create poems. Astigarraga *et al.* do not formally evaluate Poet's Little Helper, but they mention that it does not meet their personal expectations.

McGregor *et al.*'s system (2016) uses a model that can dynamically generate topical subspaces of an existing vector space. The underlying large vector space model is based on the English Wikipedia, while the subspaces are generated based on input keywords. A second model represents the phonological properties of poetry by counting the co-occurrences of sounds in English sonnets. A third is a class-based n-gram model which represents the frequency with which parts of speech follow each other in an English sonnet, and a fourth represents sentiment by using a corpus of sentiments related to topics in telephone conversations. The system chooses topics based on the fourth model, and uses the first model to generate a set of conceptually related keywords for each one. The second model is then used to generate a poem template, and words that are as close as possible to the words in the topical subspace are chosen to fill it. Finally, each word in the poem is given a score based on its semantic appropriateness to the fourth model and its phonological appropriateness to the second one. The least appropriate words are selectively removed and replaced with more appropriate words, until the poem converges on a maximal score. McGregor *et al.*'s system therefore combines a mined semantic model with the targeted editing-based optimization of COLIBRI and WASP. McGregor *et al.* then evaluate the system using a survey in which participants rate the poems on creativity, meaningfulness, and quality. Three different experimental groups are used in

which each group gets a different type of framing information. Regardless of group, the system's overall scores on all three criteria are relatively negative (all below 3.5, and mostly below 3, on a Likert scale ranging from one to seven). However, standard deviations are high. McGregor *et al.* seems to be an example of a system with an interesting and relatively sophisticated Process, but which still has, at best, mixed results in terms of Product:

> and wondered but talked me shifty Sinatra
> like hang says in current or that four man
> because this full gets really there makes both
> another golden way though your man

Gervás's SPAR system (2017) also uses a large vector space model, constructed from a corpus of English adventure novels from the public domain. Based on a user provided seed word, it builds a set of keywords related to the seed word in the vector space. It then searches the original corpus for phrases of a suitable length that contain both a keyword and a potential rhyming word. Using these phrases and a Markov model, it then constructs candidate phrases that connect rhyming and topical words, and joins them together in a way that satisfies its rhyme scheme. Like WASP, SPAR is evaluated through its Press-based cultural success, with the poems being exhibited in a Spanish poetry festival. Gervás also evaluates SPAR through *ad hoc* automatic measurements meant to capture thematic cohesion and enjambment, two features upon which Gervás is especially focused with this system. SPAR outperforms samples from Full-FACE (Colton et al., 2012), Toivanen *et al.*'s system (2012), and PoeTryMe on both metrics, but does not approach the scores of human-authored samples. Stereotrope (Veale, 2013a) is also tested, and outperforms both SPAR and human poems on thematic cohesion, but its score for enjambment is zero, since each line in a Stereotrope poem is always a complete sentence.

The Poem Machine (Hämäläinen, 2018) generates poems in Finnish using semantic bigram data from the Finnish Internet Parsebank. The Poem Machine's goal is to bring standard practices from other domains of natural language generation into computational poetry: the typical four-stage NLG pipeline consists of content determination, sentence planning, surface generation, and morphology and formatting. During content determination and sentence planning, the Poem Machine selects words which have a metaphorical relation to each other: this is done by choosing words from certain parts of speech that appear together more than a minimum number of times, but less than a maximum number of times, in the data set. The remaining steps are done using a syntactic tool that the author developed himself for generating morphologically correct sentences in Finnish. This is similar to template generation, but somewhat more flexible, and allows words to be lemmatized and transformed into the appropriate case based on Finnish grammatical rules. The Poem Machine is evaluated by the same method as P.O.Eticus (Toivanen et al., 2014) and compared to a prior approach in which fragments of existing Finnish poems are combined. The non-expert judges agreed unanimously that The Poem Machine's outputs are poems. Additionally, The Poem Machine outperforms the prior approach on typicality and on how good the language is, but does worse than the prior approach on understandability.

3.2.2.4 Bespoke data mining

In addition to representing semantic relations explicitly or mining them from a general-purpose corpus, some researchers also take pains to mine specific types of data that are especially relevant to their poetic goals.

The full-FACE system (Colton et al., 2012), discussed in Section 3.1.4.1, uses a corpus of English simile data gathered by the researchers. Similes mined from Google N-grams are processed into propositional logic statements about semantic properties of things, which are then combined into new similes from which the poetry system draws to fill its templates. Colton *et al.* do not formally evaluate their system, although it is emphasized that the system is designed to fill the requirements of the FACE model.

Veale's Stereotrope system (2013a) also focuses on simile and metaphor. Stereotrope searches Google N-grams for the properties and actions that are stereotypically associated with concepts in English, such as "cops eat donuts". It encodes these properties propositionally and then searches for pairs of concepts that have similar or contrasting stereotypical properties. If two concepts are associated with similar properties, then one can be used as a metaphor for the other. Similarly, contrasting pairs of properties can be used to highlight the tension or contradiction between aspects of a concept. The Stereotrope system chooses a central metaphor for its poems, and then fills a series of line templates with words describing the conceptual properties underlying the metaphor. For example, the following poem revolves around a comparison between "marriage" and "prison":

The legalized regime of this marriage
My marriage is an emotional prison
Barred visitors do marriages allow
The most unitary collective scarcely organizes so much
Intimidate me with the official regulation of your prison
Let your sexual degradation charm me
Did ever an offender go to a more oppressive prison?
You confine me as securely as any locked prison cell
Does any prison punish more harshly than this marriage?
You punish me with your harsh security
The most isolated prisons inflict the most difficult hardships
O Marriage, you disgust me with your undesirable security

Veale evaluates the Stereotrope system by evaluating the pointwise mutual information of its N-grams, rather than the poems themselves.

Data mining directly from poetry, rather than from a corpus of general knowledge, is also possible. Yan *et al.*'s (2013) iPoet begins with a set of user-specified keywords in Chinese. iPoet retrieves a set of poems from a database which contain these keywords, and ranks them on their relevance and importance based on term frequency. Latent Dirichlet allocation is used to cluster the terms from each poem. iPoet then assigns a cluster to each line of the poem to be generated, and uses a generative summarization algorithm to compress the meaning of each of these clusters into a single meaningful line. The lines are constrained for number of characters, rhyme, and tonal pattern. Yan *et al.* evaluate iPoet using ROUGE, an evaluation methodology for summarization and translation; they also survey quasi-expert humans about the generated poems' fluency, rhyme, coherence, and meaning. iPoet's poems outperform four other sets of generated poems, including randomly generated poems, with which it is surveyed.

Oliveira and Alves (Oliveira and Alves, 2016) expand PoeTryMe to work with an English tool called TextStorm, which creates customized concept maps by parsing a written document. These customized maps are then used, instead of the general-purpose maps of CARTÃO, to generate a poem describing the

information in whatever document is specified by the user. PoeTryMe is evaluated using the cumulative interpretation of the FACE model; Oliveira and Alves argue that it meets all four of the FACE model's criteria. However, they mention that the contents of the concept maps are not as good as they could be, as TextStorm does not perform anaphora resolution or named entity recognition.

> why ask my tower? that old sight will swear
> a name of weight; line little meter heir
> thus the great people of almighty year
> and elysées, and street shall disappear

3.2.3 Neural networks

We now turn to the topic of poetry generated using neural networks. As discussed in Section 3.1.4.3, neural networks are a popular area of research which are difficult to place into this taxonomy, due to their "black box" nature. However, in practice, most poetry neural networks contain both knowledge representation and optimization.

Zhang and Lapata (Zhang and Lapata, 2014) generate classical Chinese quatrains based on a user-supplied keyword. Their system uses the ShiXueHanYing phrase taxonomy, a specialized word clustering created for Chinese human poets, to associate the supplied keyword with other words and phrases. It creates all possible lines combining these phrases that satisfy their tone and rhyme constraints, and then uses a recurrent neural network trained on other Chinese poems to select the phrase that is most likely to appear. The second, third, and fourth lines are similarly selected to be most likely based on the first line. They evaluate their system automatically using the machine translation evaluation metric BLEU, and manually by asking a group of experts in Chinese poetry to rate the poems' fluency, coherence, meaningfulness, and poeticness as well as ranking the generated poems relative to each other. The experts are also given human-written poems and poems generated by other Chinese systems. Zhang and Lapata's system significantly outperforms all the systems in the experiment except the human ones.

Goodwin (Goodwin, 2016), as mentioned in Section 3.1.4.3, uses an LSTM-RNN to generate English poetry. Unlike many of the other neural networks discussed in this section, Goodwin's system does not use a multi-part architecture, instead simply generating characters and using the long-term memory property of an LSTM-RNN to avoid the incoherence of Markovian character generation. Goodwin's system has been used on modern poetry as excerpted in Section 3.1.4.3, but also on 19th-century poetry as well as image captions, dictionary definitions, and other forms of text.

Wang *et al.* (Wang et al., 2016b) generate Chinese Song iambics, a form of poetry with lines of variable length but strict rhythm. They use an attention-based model, in which an encoder neural net and a decoder neural net convert lines of poetry back and forth into a hidden, semantic representation. The current status of the decoder (i.e. the previous character), plus the status of all the hidden variables, are used to predict the next character. A relevance factor, for which the attention-based model is named, also helps direct the model to the most relevant hidden variable at a given time. The encoding involves the use of a vector-based semantic model to reduce data sparsity, and the decoder is constrained to only select characters which fit the rhyme and tone constraints. Wang *et al.* test their system with BLEU, and by asking human experts to rate poems' fluency, poeticness, and meaningfulness. It outperforms two other models, including Zhang and Lapata's (2014), on all three metrics, and even outperforms human poems on poeticness, but human poems are significantly more fluent and meaningful.

Yi *et al.* (2016), like Zhang and Lapata, generate Chinese quatrains based on an input word. Like Wang *et al.*, they use an encoder-decoder model with an attention mechanism. Three separate models of this nature are used: one to write a first line based on the input word, one to write the second line based on the first line, and one to write the subsequent lines based on the two lines preceding them. They test their system by using BLEU and by asking human experts to rate the poems' fluency, coherence, meaningfulness, poeticness, and a general impression. Compared against other systems and human poems, the system outperforms the other systems on all five metrics, but not the humans.

Ghazvininejad *et al.* (2016) use a recurrent neural network to generate poetry in English. Their system, Hafez, also begins with a user-supplied keyword. It uses this keyword and a word2vec semantic model trained on Wikipedia to generate a list of topically related words. Next, it sorts the topically related words into end rhyme classes. For each rhyming pair of lines in the poem, it chooses the rhyming pair of words that are most closely semantically related to the keyword. Hafez then builds a large finite-state acceptor containing all possible English word sequences that end with the chosen words, obeying the constraints of meter and ending with a comma or period. The number of possible paths through this acceptor is fantastically large, so one version of Hafez uses its recurrent neural network at this step, training it on the word sequences of a large corpus of English songs and using it to guide a beam search that selects the most likely sequences. The likelihood is further modified with penalties and bonuses to discourage repeated words and encourage topical ones. A different version also uses an encoder-decoder model more similar to the ones used by Yi *et al.* (2016) and Wang *et al.* (2016b), which is trained to generate English song lyrics based on their rhyming words. Ghazvininiejad *et al.* use human judges to compare one version of the system to another. This testing confirms that Hafez generates better poetry when the bonuses to encourage topical words are in place, and that the encoder-decoder model performs better than the beam-search-only model.

> Civil War
>
> Creating new entire revolution,
> An endless nation on eternal war,
> United as a peaceful resolution,
> Or not exist together any more.

Schlegel *et al.*'s (2018) G-Rap system contains multiple long-short term memory RNNs, each of which produces the next line of an improvised English rap lyric based on user input. At each step, the user indicates which lyric they prefer, which can be used as feedback to the RNNs as well as providing a measure over time of which network is performing best. Schlegel *et al.* do not formally evaluate G-Rap, instead viewing it as an example of how different RNN structures built for the same task can be compared against each other.

> For the love of money dollar bills
> We just getting started dont panic
> When I wake up in the morning
> I can see the sun come up

Xu *et al.* (2018) apply the attention-based encoder-decoder networks which have been successful in previous poetry systems to the novel task of generating Chinese poetry from an image. Both visual features

and linguistic keywords are extracted from the raw visual data of the image using a convolutional neural network before being fed into the encoder-decoder. In addition to an attention model, the encoder-decoder also maintains a latent representation of a topic which is derived from the group of keywords and held constant throughout the generation process. The poem is then generated character-by-character, based on the previous character, the attention model, and the topic.

How to train this neural network for its novel task is a pertinent question. Xu *et al.* construct a large database of image-poem pairs by using an existing visual keyword extractor on a large database of images from the Internet and matching them automatically to the keywords of poems generated by Zhang and Lapata's (2014) system. After being trained on these image-poem pairs, Xu *et al.*'s system is judged by asking humans to judge the poems' poeticness, fluency, coherence, meaning, and consistency with the inspiring image. The system is judged against two partial versions of itself—one without the visual features, and one without the keywords—and three other poetry systems, including Zhang and Lapata's (2014) and Wang *et al.*'s (2016a). Image-poem consistency is also evaluated automatically by computing the recall rate of key concepts from each image. The full version of the system outperforms the other systems surveyed on all metrics.

Similarly, Loller-Anderson and Gambäck (2018) train a neural network to generate visual image-inspired poems in English. Their approach uses an existing convolutional neural network to identify objects in an image. ConceptNet is used to find lists of words related to the inspiring objects, which are then searched for rhyming pairs. The system then uses a technique similar to that of Hafez, in which a tree representation of many possible lines ending with the chosen rhyming pairs is generated, and the path through the tree is chosen which has the lowest weights based on connections between words in an LSTM trained on song lyrics. Loller-Anderson and Gambäck recruit non-expert judges to rate the system's grammaticality, poeticness, and meaningfulness, as well as performing a modified Turing test. Loller-Anderson and Gambäck do not evaluate these results for statistical significance or compare the results for grammaticality, poeticness, and meaningfulness to the results for a control group. They state that their results are inconsistent but sometimes good, and that poeticness is rated higher than the other two criteria.

> The sun is in my big raincoats
> I dont know what to do scapegoats
> Im raining and it looks like rain
> Theres so much for me to abstain

Yang *et al.* (2018) generate Chinese quatrains using a pair of encoder-decoder systems. Based on a user-provided query which can be anything from a word or sentence to a full document, the first encoder-decoder generates a series of four latent variables representing keywords for each line of the poem. The second translates from this keyword-based outline to a full poem. Yang *et al.*'s second encoder-decoder is trained not only "horizontally" (i.e. on the lines that come before and after a given line) but also based on "vertical" slices (i.e. a set of four characters taken from the nth character in each line) encoded in augmented word2vec, so as to encourage parallelism and rhythm between lines. This is a technique which is easier to do in Chinese quatrains than in some other languages and forms, since each line in a Chinese quatrain contains the same number of characters and each character has a discrete semantic meaning. Yang *et al.*'s system is used by comparing it to Wang *et al.*'s (2016b) system and to partial versions of itself on a number of automated metrics, including BLEU and some other neural network-specific metrics, and a measure of Yang *et al.*'s devising which captures the rhythmic and tonal rules of Chinese quatrains.

Yang *et al.* also survey human quasi-experts about their poems, asking about readability, consistency, aesthetic feeling, evocation of emotions, and an overall score. Their poems average above 3 (on a Likert scale of 1 to 5, with 5 being better) on all metrics, and 73% of individual poems have scores above 3. However, the human portion of the evaluation does not contain a control group to compare the poems against, nor does it otherwise test the significance of its results.

3.2.3.1 Twitter Poetry

We now turn to an area that more directly inspires our current research: poetry made from or with Twitter.

Many inspiring examples in this area are mere generation. For example, Pentametron (Bhatnagar, 2012) is a found poetry system that posts pairs of rhyming English tweets from Twitter in iambic pentameter. Aside from measuring the rhyme and meter of each tweet, no optimization or use of knowledge is performed.

> Cats fucking love lasagna. Holy shit.
> I cannot stand a fucking hypocrite :(

The Longest Poem in the World (Gheorghe, 2013) performs an even simpler calculation, only posting pairs of English tweets with the same end rhyme:

> i turn 21 in 7 days
> Trying to get through the phases, going through a phase

Wood's (2013) Tweet Haiku bot searches Twitter for English tweets that have the appropriate number and grouping of syllables to be reformatted as a haiku:

> All we have in this
> house is dark chocolate. This is
> indeed a problem.

Poetweet (b arco cultural centre, 2013) generates poems based on the tweets of a specific Twitter account, thus allowing Twitter users to create a personalized poem out of their own activity. It captures fragments of tweets which have the correct rhyme and meter for a poetry form that the user chooses:

> That Goes Like This" from Spamalot.
> Is good because I am LE TIRED.
> I think about this problem a lot.
> Of course - I'd be honoured!

Twitter poetry also appears on occasion in the academic world. Mobtwit (Hartlová and Nack, 2013) uses tweets selected based on their location, with the intent of writing poems that summarize current happenings in a specific city. The poems are also optimized for a positive, negative, or mixed sentiment; a machine learning algorithm classifies tweets by sentiment in this way, training itself by measuring which words are associated with happy or sad emoticons.

Charnley *et al.* (2014) build a Twitter poetry generator in English as an illustration of their FloWr framework for computationally creative software design. The software selects an uncommon adjective with a negative emotional valence from a pre-built dictionary, and then searches Twitter for tweets containing this adjective. Keyphrases are extracted from the tweets. Tweets with inappropriate content, such as specific people's names, unpronounceable words, or strong profanity, are removed, as are duplicate tweets and tweets that do not contain a personal pronoun. A set of rhyme and meter-matching modules then groups the tweets into pairs in which all tweets end with the same sound and have the same number of syllables. These pairs are then brought together into a form specified by a template, optimizing for the tweets with the most negative emotions. Enjambment is also added to the lines. This generator is not formally evaluated, as it is intended merely as an illustration of the FloWr framework's functionality.

> I hate the basement level of buildings.
> You always lose reception and its always quiet and eerie.
> This doesn't quite capture the eerie pink glow of this morning.
> Is pop culture satanic?
> In a spiritual (not religious) sense?
> I dont really know.
> But man, there are some eerie parallels.
> It's concerning.
> I find it very eerie when someone is tinkering with your teeth and telling jokes.
> Or is that just me?

Oliveira (2017a) builds a Portuguese Twitter poetry generator, O Poeta Artificial, with the PoeTryMe architecture. Like prior implementations of PoeTryMe, it uses a context-free grammar filled with words that are related, in a given semantic knowledge base, to a user-supplied seed word. O Poeta Artificial selects seed words by reading the most recent Portuguese tweets about a trending news item, extracting the content words from each tweet, and using the most commonly appearing of these words as its seeds, sometimes supplementing with related words from the Wikipedia. An earlier version of O Poeta Artificial experienced issues because the most relevant words to a trending news item were often slang or hashtags which did not appear in its knowledge base. The later version addresses these issues through several strategies. Content words' frequencies, in the selection process, are divided by their frequencies in a large general-purpose Portuguese corpus, using a tf/idf-like strategy to emphasize the most topical words. In addition to the context-free grammar, the later version of O Poeta Artificial also makes use of text fragments cut directly from trending tweets, sometimes with paraphrase techniques applied, and of templates designed to highlight the trend and some of its extracted semantic relations. O Poeta Artificial is not formally evaluated.

3.2.4 Conclusion

We have now seen in-depth examples of poetry systems illustrating what can be done in different parts of our taxonomy, including systems that specifically inspire the work we will do in the next sections. Although this is not a full list of all poetry generation systems that have been published academically in the past several years, it is enough to be a representative sample.

We can infer a few things from our look at this sample. First, poetry systems have been reasonably successful in many corners of the taxonomy; there is no one true way, at this stage in the state of the art, to

create a system. However, within a category, some approaches may appear to be more fruitful than others. For instance, in neural network poetry, there is a strong trend towards attention-based encoder-decoder systems as the most effective means of generation.

We have also looked at the different ways that existing poetry systems have been evaluated by their authors. We can see that, despite the emphasis in recent years on the need to formally evaluate creative systems, many researchers still do not formally evaluate their poetry systems. Others use methods whose effectiveness is questionable, such as modified Turing tests or surveys of non-experts, or ad-hoc tests whose validity is in question. However, many poetry researchers do use reasonably good evaluation methods. In particular, it is notable that the Chinese neural network poetry community has developed a somewhat consistent standard of evaluation, in which both the system being developed and relevant previous systems (and, often, human poetry) are evaluated both on an automatic machine learning metric and on an expert survey with fluctuating but somewhat consistent criteria. As these are the closest to standardized evaluations existing so far in computer-generated poetry, it is also notable that each system tested on them seems to consistently outperform the systems that came before. This may be an indication of consistent progress in this corner of the field, or perhaps of careful selection of systems to compare one's work against, or an unwillingness to publish when there are negative results.

When systems are not evaluated formally, one might intuitively suspect the researchers of trying to fool readers into thinking that their system is effective without evidence. But in practice, in systems that are not evaluated, there is a range of informal opinions by the researchers, including researchers who see their systems as underperforming, and researchers who see evaluation as irrelevant because their poetry system is a proof of concept, or a demonstration of one particular poetry-related ability, rather than a serious attempt at making finished poetry that will pass a Product evaluation. Although this section is not a formal survey or meta-analysis, perhaps these mixed examples can begin to shed light on some of the reasons why researchers, in practice, do not always follow the evaluation guidelines that are set out in theory.

Chapter 4

Our experiments in poetry evaluation

We have now surveyed enough work in computational creativity, and in poetry specifically, to have an idea of where some of the low-hanging fruit lies in this research field. In Chapter 5 we will discuss our actual attempts to build a poetry system, but in the current chapter, we attempt to make inroads on a more abstract concern: how to develop an evidence-based, domain-specific method of evaluating computational poetry. In particular, we will explore Product-based evaluations of poems.

A few specific research questions are addressed in this chapter. To what extent can non-expert humans reliably evaluate poems at all? Are existing domain-general Product metrics appropriate for poetry? How do expert humans evaluate poems, and can we extrapolate from this to develop better ideas for evaluating computational poetry?

To investigate these questions, we perform two exploratory studies. First, we study what happens when non-expert humans use computational creativity metrics to evaluate human poems. Second, we study what happens when quasi-expert humans are asked to evaluate computer-generated poems, without being restricted to a specific set of metrics, and we derive our own tentative poetry evaluation metric from their comments.

It should be noted that, rather than cleanly completing Chapter 2, then Chapter 3, then this chapter, then Chapter 5, our research has actually meandered back and forth between the three areas of research. Some studies in this and the following chapter, especially in early sections of these chapters, was done before we had finished our thorough survey of related work in computational creativity. We will point out, where necessary, aspects of our experimental research that could have been improved if we had already known at the time all the theoretical background that we presented in Chapters 2 and 3.

4.1 Human competence in evaluating poetry

4.1.1 Introduction

In our first study, we are interested in what happens when Product-based computational creativity metrics are applied to human-generated poetry. Intuitively, this serves as a type of preliminary validity test for

these metrics. Assuming that there is some sort of ground truth about human-generated poetry - that is, that we can select some human-generated poems that we are quite sure are creative, and some others that we are quite sure are not as creative - then a metric that purports to measure creativity, by measuring the attributes of a creative product, should be able to identify the more creative poems. We test this idea by having non-expert human judges rate more- and less-creative human poems on four sets of criteria: novelty and value, typicality and quality (Ritchie, 2001), the Creative Tripod (Colton, 2008), and the IDEA model (Colton et al., 2011).

For our ground truth, we selected the published work of professional poets in the prestigious magazine "Poetry" as our more creative group. As our less creative group, we selected poems that were posted by beginners in an online forum for poetry critique forum and that received negative comments from the forum's moderators. For a third, "medium" group, we selected poems that were published in more obscure magazines that pay authors a token amount. In other words, we made our selections by assuming that the Press evaluation of poetry (its commercial success and reception by editors/critics) is a reasonable stand-in for the poetry's creativity on a Product level.

To our surprise, we find that non-expert human judges are very bad at judging human poetry. In fact, on most sets of criteria, they significantly prefer the less-creative poetry. Thus, in poetry, it may be even more important to rely on experts than in other domains. We suspect that novices have more difficulty understanding the more creative poems. We tentatively support this suspicion by repeating the study with poems intended for children, for which the reverse preference effect goes away, although the non-experts still could not significantly distinguish more-creative children's poems from less-creative ones. We also find that the specific metric used (value, novelty, appreciation, skill, *etc.*) does not seem to make much difference to this result; most of the metrics are moderately or highly correlated with each other.

A possible objection to this method of inquiry is the argument, discussed in Section 2.7.2, that the products of computational creativity should not be measured by human standards. Therefore, the reverse is also true, and one should not expect the standards for computational creativity to apply to humans. We remain unconvinced by this argument for two reasons. First, if computational creativity does not have to be anything like human creativity, then we are left without a definition of creativity by which to evaluate our work at all—or, perhaps, an alternate definition that has no demonstrable link to what we mean by the word "creative" in any other context. Second, we have chosen methods of evaluation that do *not* require computational and human products to be exactly the same. Some methods of evaluation, such as the modified Turing test, do have a problem with encouraging close pastiche over other forms of creativity. But our four chosen sets of criteria are all relatively broad. If we define creativity as novelty and value, for instance, then it is possible that a creative human might be novel and valuable in one way, and a creative computer might be novel and valuable in another. Thus, our method does not require human and computational creativity to be exactly the same: it only requires that they be similar enough that the same basic evaluative concepts should apply to both.

As far as we know, other validity testing of Ritchie's criteria, the Creative Tripod, and the IDEA model has not yet been done. That is, while each of these models appears plausible, it has not been empirically and falsifiably tested whether or not they work to separate more creative products from less creative products in practice. While our experiment is not a full validity test, we consider it an important preliminary step in that direction.

Our research suggests that the implementation of popular creativity metrics is less straightforward than intuition would suggest, and that using non-expert judges can not only produce less robust results, but can produce actively misleading ones. In a research climate where many poetry evaluations are still

done by non-experts or groups of judges undifferentiated by expertise (Toivanen et al., 2012; Rashel and Manurung, 2014b; Barbieri et al., 2012; Singh et al., 2017; Hayes and Wiggins, 2015; McGregor et al., 2016; Loller-Andersen and Gambäck, 2018), or by the researchers themselves, who are not necessarily poetry experts (Tobing and Manurung, 2015; Misztal and Indurkhya, 2014; Oliveira, 2012; Astigarraga et al., 2017; Colton et al., 2012; Oliveira and Alves, 2016), these results are quite important.

A prior version of this section appeared at the Sixth International Conference on Computational Creativity (Lamb et al., 2015a).

Ritchie's model	**Novelty and value**
• This resembles other poems I have read. *(Typicality)* • This is a high quality poem. *(Quality)* • I don't think this is a very good poem. *(Quality, reverse coded)* • This is not a poem. *(Typicality, reverse coded)*	• This is a high quality poem. *(Value)* • This poem is not like other poems I have seen before. *(Novelty)* • I don't think this is a very good poem. *(Value, reverse coded)* • This poem is clichéd. *(Novelty, reverse coded)*
Colton's Creative Tripod	**IDEA model**
• The author of this poem seems to have no trouble writing poetry. *(Skill)* • The author of this poem is imaginative. *(Imagination)* • The author of this poem understands how poetry works. *(Appreciation)* • The author of this poem isn't very good at writing poetry. *(Skill, reverse coded)* • The author of this poem isn't bringing anything new or different into the poem. *(Imagination, reverse coded)* • The author of this poem doesn't really know anything about poetry. *(Appreciation, reverse coded)*	• I like this poem. *(Wellbeing)* • I am willing to spend time trying to understand this poem. *(Cognitive Effort)* • This poem makes me unhappy. *(Wellbeing, reverse coded)* • This poem is not worth bothering with. *(Cognitive Effort, reverse coded)*

Table 4.1: The questions used in our study for each of the four evaluation metrics tested.

4.1.2 Experiment I

4.1.2.1 Method

We tested 4 common metrics for creativity evaluation: novelty and value, Ritchie's typicality and quality model (Ritchie, 2001), the Creative Tripod (Colton, 2008), and the IDEA model (Colton et al., 2011). These metrics are easy to test on human poetry since they focus on the creative product and not on the process of its creation. Since none of these metrics have been put into a standardized questionnaire form, we constructed our own agreement scale for each, with responses ranging from 4 (strongly agree) to -4 (strongly disagree). The questions used are shown in Table 4.1.

Focus on the shapes. *Cirrus,* a curl,
stratus, a layer, *cumulus,* a heap.

Humilis, a small cloud,
cumulus humilis, a fine day to fly.

Incus, the anvil, stay grounded.
Nimbus, rain, be careful,

don't take off near *nimbostratus*,
a shapeless layer

of rain, hail, ice, or snow.
Ice weighs on the blades of your propeller,

weighs on the entering edge of your wings.
Read a cloud,

decode it,
a dense, chilly mass

can shift, flood with light.
Watch for clouds closing under you,

the sky opens in a breath,
shuts in a heartbeat.

Figure 4.1: A sample poem from the Good dataset: "Flying Lesson" by Dolores Hayden.)

Magazine	URL
Abyss & Apex	http://www.abyssapexzine.com/
Amethyst Arsenic	http://www.amethystarsenic.com/
Astropoetica	http://www.astropoetica.com/
GlassFire	http://www.peglegpublishing.com/glassfire.htm
Goblin Fruit	http://www.goblinfruit.net/
Ideomancer	https://www.ideomancer.com/
Neon	http://neonmagazine.co.uk/
Punchnels	http://www.punchnels.com/
Raleigh Review	https://www.raleighreview.org/
Rawboned	http://rawboned.org/
Silver Blade	https://www.silverblade.net/
Strong Verse	http://www.strongverse.org/
Through the Gate	http://throughthegate.net/
Vine Leaves	http://www.vineleavesliteraryjournal.com/
Writing Tomorrow	http://writingtomorrow.com/

Table 4.2: Full list of the 15 magazines used in our Medium dataset. Note that these magazines were accessed in 2014; as of the completion of this thesis in 2018, many are now defunct, and some of these URLs no longer function. Given the short lives of most minor literary markets, this is not unexpected.

I'm convinced stars dream of being planets
that they would trade each roiling flare
to taste the heft and feel the quake of granite;

to have their scalding plasma winds, each rife
with tortured particles, replaced by winds supporting
birds and bats and bugs above a scenery of life.

What might their fortunes be if stars could roll the cosmic dice?
Most likely, they'd metamorphosize to silent, pockmarked rocks
or coalesce as spheres of jellied gas or frothy ice.

The universe respects no guarantees,
which may explain the mystery of why
the stars continue burning in their vacuum seas,

resembling people in that way, safe
in what they are and what they have,
if not content, galled by millennia of constant chafe.

I wonder if our tortured Earth, within its roiling heart
of molten sunfire, ever dreams it could
escape us for eternity and be reborn a star?

Figure 4.2: A sample poem from the Medium dataset: "Stars Dream" by Craig W. Steele.)

To die seems quite an adventure
One i will wait to see
Though death i do not desire
Nor dead i wish to be

The folds of hell seem far too deep
The gates of heavn too vast
Though part of me wishes to see
The place i'll rest at last

Figure 4.3: A sample poem from the Bad dataset: "On Dieing" by AsILiveAndBreathe.)

4.1.2.2 Data

We used three hand-collected data sets of contemporary poetry written by humans. Each set contained 30 short poems in English of between 5 and 20 lines, inclusive. We stuck to contemporary poetry so as to avoid different eras of poetry becoming a confounding factor, and so as to minimize the probability that a study participant had already read the poems before, and we stuck to short poems in order to make it easier for participants to rate a large number of poems. In no case did more than one poem by a single author appear either in the same data set or across data sets.

The **Good** data set was composed of poems from Poetry Magazine[1] between November 2013 and April 2014. Poetry Magazine is a very long-established, professional magazine which can reasonably be considered to contain the work of the most critically acclaimed mainstream literary poets working in English today. In other words, a poem published in Poetry magazine can be considered successful from a Press perspective. All poems meeting the length and non-duplication requirements and appearing in the magazine during this time window were selected, with the exception of a few which were discarded due to complex visual formatting and two which were discarded due to experimenter discomfort with content. The remaining 30 poems comprised the Good data set. An example is given in Figure 4.1.

The **Medium** data set was composed of 2 poems each from 15 lesser-known online magazines. Some of these were magazines devoted exclusively to poetry while others were a combination of poetry and prose; some were devoted to a particular genre or subject matter while others were not. Each magazine pays a token amount (between US $5 and $10) for a poem of the required length. (See Table 4.2 for a full list.) For each magazine, the most recent 2 poems meeting length and author uniqueness requirements were chosen for the data set, with a single exception in which one of these poems was discarded for thematic unsuitability as with the Good data set and the third-most-recent poem chosen as a replacement. This added up to a Medium data set of 30 poems. An example is given in Figure 4.2.

The **Bad** data set was composed of poems by unskilled amateur poets. We chose these poems by going to the Newbie Stretching Room at the Poetry Free-For-All [2], an online poetry critique forum. This section is for newcomers who have not posted poetry on the forum before; both experienced moderators and other newcomers can comment on the poems. We chose only poems meeting the length and author

[1] `https://www.poetryfoundation.org/poetrymagazine`
[2] `http://www.everypoet.org/pffa/`

uniqueness requirements from this section, and discarded any which had received any positive feedback or compliments from a moderator. Most of the chosen poems received generic comments from moderators instructing the author to read some introductory articles on how to improve their poetry; a few had more specific comments from moderators about why their poem was bad. (Example feedback from moderators on Bad poems includes: "This lacks any sort of imagery that will allow me to settle in and get a picture. Therefore, it is left as a mere statement of fact without any musicality, and nothing novel to chew on." Or, from a more acerbic moderator, "This is dreadfully bad beginner's doggerel that fails for many, many, many reasons." We did not factor any comments from novices into our analysis.) Selecting the most recently posted poems which fit these requirements resulted in a Bad data set of 30 poems. An example is given in Figure 4.3.

Finally, we collected a **Test** data set containing 6 texts which were the same length as the chosen poems, but were obviously not poems. 3 of these were snippets from business news, and 3 from sports news.

A note should be made about the two Good poems (and one Medium poem) that were excluded due to content. These were poems containing highly graphic sexual and/or violent themes. Excluding poems from a study due to content always raises a potential validity issue. For instance, the handling of potentially objectionable themes could itself be an important difference between the three poetic groups. We believe our exclusion of certain poems is justified for two reasons. First, the number of poems excluded is quite small. Second, requiring study participants to view offensive content is a potential ethical issue, and one we have had to handle carefully throughout all of the studies in this thesis; according to our office of research ethics, even mild profanity has required a warning in the study introduction and a provision for participants to leave the study if they become uncomfortable. Because of the small number of poems excluded, their exclusion mitigates this ethical issue but is unlikely to strongly affect the study's results.

4.1.2.3 Collection

We recruited study participants on Crowdflower, a crowdsourced microtasking website. In order to minimize cultural and linguistic difference as a confounding factor, participants were limited to those living in the United States.

Each participant was given six poems at a time selected from any or all of the data sets, and asked to rate each poem based on one of the metrics. Each participant gave ratings using only one metric at a time. We collected enough responses to amass 20 responses on each metric for each poem - with 96 individual poems, six poems per participant, and four metrics, this resulted in a total of 1280 participants.

Results for each poem in the data set were then run through a single-factor ANOVA with a Bonferroni correction for multiple hypotheses. Since there were three metrics being tested with two factors each, and one metric being tested with three factors, we corrected for nine hypotheses. The null hypothesis was that, for all metrics, participants' responses to Good, Bad, and Medium poems would be drawn from an identical distribution. The alternative hypothesis was that, for at least one part of at least one metric, poems from one data set would be rated more highly than others.

4.1.2.4 Results

Results for this experiment were the opposite of what we expected. For most criteria in most metrics, participants rated Bad poems significantly more highly than Medium ones, and Medium significantly more

Model	Criterion	Good	Medium	Bad	F
Ritchie's model	Typicality	0.20	0.41	1.23	10.6**
	(SD)	1.09	1.10	1.06	
	Quality	0.23	0.67	1.40	10.2**
	(SD)	1.09	1.08	1.03	
IDEA model	Wellbeing	0.78	1.14	1.54	13.9**
	(SD)	1.05	1.04	1.03	
	Cognitive Effort	0.60	0.94	1.46	14.9**
	(SD)	1.07	1.06	1.04	
Creative Tripod	Imagination	0.75	1.16	1.07	2.3
	(SD)	1.04	1.00	0.99	
	Appreciation	0.67	1.11	1.68	8.3**
	(SD)	1.06	1.00	0.92	
	Skill	0.44	0.84	1.40	7.4*
	(SD)	1.06	1.02	0.98	
Novelty and Value	Novelty	0.96	0.80	0.49	9.8**
	(SD)	1.02	1.00	1.04	
	Value	0.17	0.44	0.72	3.6
	(SD)	1.04	1.03	1.02	

Table 4.3: Average ratings, standard deviations, and F scores prior to Bonferroni correction for poem categories according to each metric. Each component is scored between -4 and +4. Significant results following Bonferroni correction are marked with a *, or ** if highly significant.

highly than Good. The exception was Novelty, in which Good poems were rated more highly than Medium and Medium more highly than Bad. Imagination (Colton's tripod) and Value showed Bad poems rated more highly than Medium and Medium more highly than Good, but not significantly. Exact F-values for each of these criteria are shown in Table 4.3. Post hoc tests were performed using Tukey's method. Tukey's method showed that Medium poems were not significantly different from either Good or Bad poems on any of the criteria; only Good and Bad poems were significantly different from each other.

This was a highly surprising result since it is not attributable to mere rater incompetence or failure to pay attention. Incompetent raters who failed to pay attention might easily give the same scores to all poems. Our raters, however, had significantly different reactions to the different groups, indicating that they can indeed differentiate between these groups—but that their preferences are largely the opposite of what we had imagined.

A BB gun.
A model plane.
A basketball.
A 'lectric train.
A bicycle.
A cowboy hat.
A comic book.
A baseball bat.
A deck of cards.
A science kit.
A racing car.
A catcher's mitt.
So thats my list
of everything
that Santa Claus
forgot to bring.

Figure 4.4: A sample poem from the C-Good dataset: "December 26" by Ken Nesbitt.

4.1.3 Experiment II

One potential explanation for why participants preferred Bad poems to Good ones was that the Bad poems were more accessible. Poems from a prestigious literary journal may be difficult for non-experts to understand due to heavy allusiveness, odd figures of speech, and other poetic conventions. This hypothesis is consistent with much of the work discussed in Sections 2.5.5 and 2.8.2, in which experts' evaluations of creative work are compared to non-experts' evaluations. On average, experts in many areas prefer more novel products, products that are more emotionally negative or challenging, and products that are more difficult to understand. Jacobs' (2015) theory of foregrounding, in particular, suggests that the use of turns of phrase which are difficult to understand is one of the key aesthetic features of creative literature. To an expert, these turns of phrase produce surprise followed by understanding, but to a non-expert, if

Dollar went to dimes closet looking for change
Dollar fell apart when he got so disarranged
Larry met a dragon that was out of fire
He taught the dragon how to fly on a wire
Doughnut could never find its way home
Coffee took it in so it would not be alone

Bob the candy man was a real man of power
No matter how you looked at him he was sour
An old man whispered, no one heard what he said
Ruby woke up and found herself in bed

Figure 4.5: A sample poem from the C-Bad dataset: "Ruby's Dream" by B4i8islept.

the use of language is too novel, the understanding may not arrive. Thus, non-experts in poetry and other critically acclaimed literature could actively dislike such work due to their inability to understand it.

To test the inaccessibility hypothesis, we ran a second experiment focusing only on children's poems: poems written with children as the indended audience. The idea is that poems written for children, even when they are critically acclaimed and very creative, should be easy to understand. Thus, if our reverse preference results are due to non-experts' inability to understand creative poems for adults, then they should not appear when the same non-experts are tested on poems for children. Our experiment was identical to Experiment I, but with new data.

The **C-Good** data set was composed of children's poems found in the Children's Poetry section of the Poetry Foundation website [3] in November 2014. The same selection constraints were used as with the first data set: poems were between 5 and 20 lines in length and no poet's work was used more than once. We also excluded poems written by poets born prior to the 20th century. We collected a total of 10 C-Good poems, by critically acclaimed children's authors such as Kenn Nesbitt and Shel Silverstein. An example is given in Figure 4.4

The **C-Bad** data set was composed of poems posted in the "Poems for Children" section of the publicly accessible Family Friend Poems forum [4] by amateur poets between September and November 2014, meeting the length and author uniqueness criteria. 10 such poems were selected. As there is no expectation of detailed critique or analysis of poems posted at Family Friend Poems, we did not filter poems by the critique that was given as we did with the Bad adult poems. In fact, most responses posted to these poems were brief and complimentary (e.g. "Brilliant. Loved it 10"), even when the poems made large and obvious mistakes with meter and rhyme. An example is given in Figure 4.5.

These poems were randomized and evaluated in the same way as the poems from Experiment I, on the criteria from the same four metrics. Since there are only two data sets in Experiment II, a t-test was performed on every criterion to detect differences in how the children's poems were rated.

[3] `https://www.poetryfoundation.org/poems`
[4] `http://forums.familyfriendpoems.com/`

Model	Criterion	C-Bad	C-Good	t
Ritchie's model	Typicality	1.25	1.46	0.48
	(s.dev)	1.03	1.03	
	Quality	0.08	0.77	0.13
	(s.dev)	1.04	1.04	
IDEA model	Wellbeing	1.30	1.61	0.35
	(s.dev)	1.00	1.00	
	Cognitive Effort	0.11	0.63	0.23
	(s.dev)	1.07	1.10	
Creative Tripod	Imagination	0.65	0.80	0.74
	(SD)			
	Appreciation	1.12	1.50	0.42
	(SD)			
	Skill	0.84	1.20	0.40
	(SD)	1.01	1.07	
Novelty and Value	Novelty	0.32	0.24	0.74
	(SD)	1.06	1.05	
	Value	0.11	0.34	0.62
	(SD)	1.03	1.04	

Table 4.4: Average ratings, standard deviations, and t scores for children's poem categories according to each metric. There were no significant differences found after Bonferroni correction

4.1.3.1 Results

The children's poem results were the opposite of the adult poem results. Participants rated C-Good poems more highly than C-Bad poems on almost every criterion. However, none of these results were statistically significant. Using the highly accessible children's poems seemingly removed any preference for bad poems, but did not introduce a significant preference for good poems.

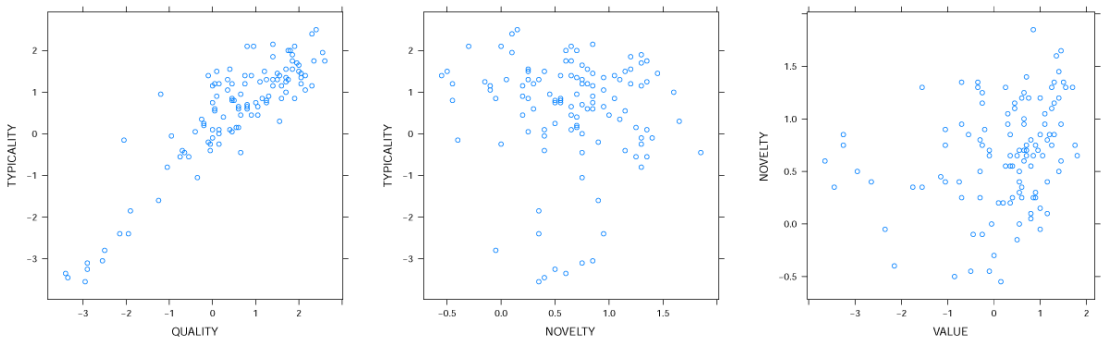

Figure 4.6: Sample scatterplots showing relationships between Novelty, Typicality, and Quality for poems in all of the data sets from both experiments.

4.1.3.2 Correlations within and between metrics

One goal in the development of creativity evaluation models is to tease out different components of the creative process. We investigated this by combining the data from Experiments 1 and 2. We generated scatterplots and correlation coefficients to examine the relationships between different criteria. With the exception of novelty and value, all criteria within the same model were strongly correlated with each other ($0.82 < r(88) < 0.99$), and scatterplots showed approximately linear relationships. Even when comparing criteria from different models, relatively high levels of correlation remained ($0.65 < r(88) < 0.99$; if both Novelty and Wellbeing are removed, then the lowest correlation coefficient jumps to 0.74). Scatterplots comparing Novelty to other criteria did not reveal anything interesting; it seems to have no significant positive, negative, or non-linear relationship with any other criterion. Example scatterplots are given in Figure 4.6.

These results suggest that, although novelty and value are statistically different from each other, the different components in the other models studied here might not be. It is possible to hypothesize, from results like this, that the different components are not *conceptually* different: that is to say, rather than typicality and quality (for example) being two separate desirable traits, perhaps they are different ways of describing essentially the same thing. Or perhaps participants choose which poem they prefer, and reason *post hoc* that the poem they prefer must be more typical and have higher quality. However, since our participants are non-experts who have difficulty understanding creative poetry, it would be premature to generalize these results to experts. Perhaps for experts, there is a more refined understanding of the different qualities of a creative poem, and the correlation between different criteria will be less.

4.1.4 Discussion

Our goal was to illustrate differences in effectiveness between different metrics, but we ended up finding something quite different. One potential purpose of using an evaluation metric, rather than directly asking "How creative is this?", is to give raters an objective framework to use. However, none of the metrics tested were objective enough to prevent two basic problems. First, the inaccessibility of Good poems produced a strong preference against these poems among non-experts, even though it would seem intuitively obvious that a poet would prefer to learn to produce this type of poem than a Bad poem. The use of children's poems removed this effect, but non-experts were still not discriminating enough to produce a statistically significant preference for Press-successful poems on any metric. Rather than arguing definitively for one metric over another, our work highlights the challenge in using humans to rate creativity through any metric.

4.1.4.1 On Novelty, Typicality, Quality and Value

Novelty could intuitively be seen as the opposite of Typicality, but this intuition is not supported by our research. Indeed, the correlation between Novelty and Typicality is nearly zero ($R = -0.05$). Poems with high Typicality may have high or low Novelty, and vice versa. Ritchie (2007) suggests that a computational system's first struggle may be with Typicality—that is, with learning to produce outputs with the basic characteristics of the target class—with Quality becoming important once that initial goal has been reached. This may be true for a computer, but our results suggest that it is probably not true for a human.

The relationship between Typicality and Quality is linear, without any sign of a threshold at which the importance of Quality changes.

Good poems are rated as more novel than Medium poems, and Medium poems more novel than Bad. Taken at face value, this would suggest that Novelty might be a better metric than others for measuring creativity. Certainly, the importance of novelty is underscored by research we discuss in Section 2.5.1.5. However, the effect for novelty reverses itself when applied to children's poems; so novelty as measured by non-expert raters is not a reliable indicator of more creative poetry across the board. When combined with our other results, it is more parsimonious to suggest that non-experts rate certain poems as more novel, not necessarily because they are more creative, but because they are more difficult to understand.

4.1.4.2 On accessibility and the target audience

It is perhaps not surprising that critically acclaimed contemporary poems would be offputting to an ordinary reader. The poems in Poetry Magazine are so complex that the magazine comes with its own explanatory Discussion Guide. Poems allude heavily to other poems and works and imply or illustrate things instead of stating them outright; some raise difficult philosophical questions such as "who is creating what, as well as who is inside the work and who is outside" (Poetry Foundation, 2014). Without a comprehensive eduction in contemporary English poetry, a reader may have difficulty understanding a poem. Our results suggest that this offputting effect may be stronger than any difference between skilled and unskilled poetry as such. To a general audience, skilled but inaccessible poems are less appealing than the poems of unskilled amateurs. Yet to an expert in poetry, it would be absurd to say that these poems are therefore of lower quality.

It is not news that non-experts and experts evaluate creative work differently (c.f. Section 2.8.2). Yet the strength of the effect here—not just negating but reversing expected trends—is surprising. It is all the more vital, in light of these results, to identify whether the judges used in computational poetry evaluations are experts, and what sort of expertise they possess.

Not every computational poetry system will necessarily have the goal of producing the sort of poetry that appears in Poetry Magazine. Some computational poetry systems are made with such lofty goals in mind, but others are built specifically to entertain or amuse non-experts. Rather than attempting to produce works which are somehow objectively valued, or to identify a single most desirable expert population, it may be more useful to first identify a target audience, according to the researcher's goal, and then select evaluators and evaluation methods which work for that audience. However, if the target audience is non-experts and the goal of a creative system is to entertain the general population, a Product survey of non-experts may still be undesirable due to the poor interrater reliability of non-experts. For a system of this nature, researchers may instead wish to use a Press evaluation—or, if available, an expert rater who specializes in writing accessible poetry for a broad audience.

In the meantime, without an identifiable target audience, it may be very dangerous to talk about a criterion like quality, value, or even skill in computational creativity as though it is only one thing. The quality of popular appeal and the quality of appeal to experts may be diametrically opposed, and there may be other audiences with still other views of quality. Until such an audience is chosen and the choice justified, the notion of quality, without the notion of quality *to whom*, is operationally meaningless.

4.2 Poetry criteria derived from consensual assessment

4.2.1 Introduction

Our second line of inquiry into evaluation is to study the judgments of poetry experts more closely. After all, if non-experts are not up to the task of judging poetry accurately, then it is worth looking at domain experts and at how they evaluate computer-generated poetry. Additionally, in Chapter 2, we discussed the importance of domain-specific criteria for creativity evaluation. Some criteria, such as novelty and value, may be domain-general, but it is still useful to ask empirically what makes a poem novel or valuable in the eyes of experts.

Some domain-specific criteria for computer-generated poetry already exist, such as Manurung *et al.*'s (2012) Meaningfulness, Grammaticality, and Poeticness, or the popular Chinese criteria (adapted from Manurung's) of Meaningfulness, Fluency, Poeticness, and Coherence (Zhang and Lapata, 2014; Wang et al., 2016b; Yi et al., 2016). However, these criteria are not based in empirical study of how poetry experts respond to computer-generated poems, but rather in computer scientists' ideas about poems. Therefore, a set of criteria derived directly from poetry experts' methods of evaluation would be a new contribution to the field.

Our study consists of two parts, performed by the same participants in the same evening. First, we use the Consensual Assessment Technique (Amabile, 1983a) to gather a group of poetry experts' evaluations of a diverse group of computer-generated poems and assess their interrater reliability. Second, we ask the experts for freeform written responses explaining some of their judgments. By analyzing these responses, we identify four major desiderata for digital poetry: Reaction, Meaning, Novelty, and Craft. Evaluating digital poetry may be more complicated than typical uses of the CAT (e.g. children's collages) due to the heterogeneity and experimental nature of digital poetry, and some of our statistical results reflect this. However, our four desiderata, combined with process-based evaluation, can be used towards a more standardized, evidence-based evaluation of a complex field.

A prior version of this section appeared as a poster at the Seventh International Conference on Computational Creativity (Lamb et al., 2016a), with the full paper appearing in the proceedings.

4.2.2 Background

The Consensual Assessment Technique (Amabile, 1983a) has been discussed already in Section 2.5.4, but for ease of reading, we summarize it again here. The CAT is a means of judging Product creativity without the use of criteria, by asking experts to rate a group of creative artifacts in the same domain and seeing if they agree. Baer and McKool (Baer and McKool, 2009) summarize current best practices:

- Judges must possess expertise in the domain being judged; novice judges have poor interrater reliability. What constitutes expertise is a matter for debate, and can vary depending on medium. Skilled novices can have decent reliability (Kaufman et al., 2009), but some theoretical experts, such as psychologists, do not.
- Judges make their judgment independently, without consulting other judges.

- Judges review the artifacts blindly, without knowing framing information such as the author's identity.
- Judges are not told how to define creativity or asked to explain their ratings.
- Judges rate artifacts on a numerical scale with at least 3 points.
- Judges use the full scale. The most creative artifacts in the group should be at the top of the scale, and the least creative should be at the bottom.
- The number of judges varies from 2 to 40, with an average just over 10.
- Interrater reliability should be measured with Cronbach's coefficient alpha, the Spearman-Brown prediction formula, or the intraclass correlation method.
- An interrater reliability of 0.7 or higher is considered good. Expert judges generally achieve interrater reliabilities between 0.7 and 0.9.
- While the CAT was designed for a homogeneous group of subjects, it works in practice even when the artifacts were made under different conditions.

The CAT has become a gold standard for assessing human creativity (Baer and McKool, 2009). It has good reliability as long as the judges possess sufficient expertise. Obtaining experts is the major bottleneck in performing the CAT. However, the CAT is reliable even with relatively few (Baer and McKool, 2009).

To our knowledge, CAT-related methods have not been used previously to assess computational poetry.

4.2.3 Method

For our study, we recruited graduate students from the University of Waterloo's Experimental Digital Media program (XDM) as our judges. XDM includes digital poetry among a variety of other avant-garde, multimedia art forms, and students in the program produce such work. We judged XDM students likely to understand both the demands of poetry as a genre and the challenges of generating poetry with a computer. They are experienced with writing and critical analysis as well as with the use of technology for art. They are not as experienced with these as a professional digital media artist, but for the purposes of our study, they struck a good balance between level of expertise and ease of access to them as research subjects.

Judges were given a set of 30 poems, in randomized order. The poems are listed in Table 4.5. They evaluated each poem on a scale from 1 to 5, with 1 being "least creative" and 5 being "most creative". They produced these ratings without group discussion. Judges were told that some poems were written by computers, and others by humans, using computers; they were not told which of the poems were which.

The 30 poems, although not labeled as such, came from three different groups. Ten, group A, were poems in which we judged the authors were trying to create a relatively autonomous creative system; all but one of these poems were taken from published papers in the field of computational creativity. Another ten, group B, were poems in which the human exerted tighter artistic control (for example, by handcrafting templates), and the computer's role was relatively limited. The final ten poems, group C, were poems generated using specific source material which remained recognizable in the final product—either "found"

Group	Title	Author	Score	SD	Response
B	"Notes on the Voyage of Owl and Girl"	J.R. Carpenter	4.4	0.69	6
B	Excerpt from "Definitions II - Adjectives"	Allison Parrish	4.3	0.70	18
C	"Conditionals"	Allison Parrish	4.0	1.04	5
B	"trans.mission [a.dialogue]"	J.R. Carpenter	3.9	1.02	4
A	Untitled	Unnamed system (Toivanen et al., 2013)	3.6	1.09	0
B	"Walks From City Bus Routes"	J.R. Carpenter	3.6	1.18	11
B	Excerpt from "]]"	Eric Goddard-Scovel and Gnoetry	3.4	1.18	0
C	"St. Louis Blues 2011"	Christopher Funkhouser	3.3	1.36	10
A	Untitled	P.O. Eticus (Toivanen et al., 2014)	3.1	0.64	0
A	"Angry poem about the end"	Unnamed system (Misztal and Indurkhya, 2014)	3.1	0.87	0
B	"Good Sleep"	George Trialonis and Gnoetry	3.1	1.36	10
A	"Voicing an Autobot"	Allison Parrish	2.9	1.43	4
B	"The Ephemerides"	Allison Parrish	2.9	0.56	0
A	Untitled	IDyOt (Hayes and Wiggins, 2015)	2.9	1.16	3
C	"HaikU"	Nanette Wylde	2.8	0.75	5
C	Excerpt from "Dark Side of the Wall"	Bob Bonsall	2.8	0.66	0
B	Excerpt from "Exit Ducky?"	Christopher Funkhouser, James Bonnici, and Sonny Rae Tempest	2.7	1.03	9
C	"Spine Sonnets"	Jody Zellen	2.6	0.79	1
C	"Ezra Pound Sign"	Mark Sample	2.6	1.58	13
A	"Blue overalls"	Full-FACE (Colton et al., 2012)	2.6	1.32	5
C	"Regis Clones (Couplets from ZZT-OOP)"	Allison Parrish	2.6	1.40	6
A	"quiet"	MASTER (Kirke and Miranda, 2013)	2.4	1.32	11
B	Excerpt from "Permutant"	Zach Whalen	2.2	0.96	12
A	Untitled limerick	Unnamed system (Rahman and Manurung, 2011)	2.2	1.36	3
A	"The legalized regime of this marriage"	Stereotrope (Veale, 2013a)	2.1	1.09	7
C	"Times Haiku"	Jacob Harris	2.1	1.03	10
C	Untitled	Mobtwit (Hartlová and Nack, 2013)	2.1	1.36	15
A	Untitled	Unnamed system (Tobing and Manurung, 2015)	1.9	0.83	4
B	"Rapbot"	Darius Kazemi	1.9	0.99	0
C	"The Longest Poem in the World"	Andrei Gheorghe	1.7	0.67	7

Table 4.5: The 30 poems used in our experiment, ranked from highest to lowest average rating. The "Response" column lists how many lines of explanation, in total, were given for judges' ratings of the poem in part 2 of the study.

poetry or modifications to a well-known human poem. All poems were published between 2010 and 2015. We presented the poems in plain text and without their titles. When the published work was a generator producing arbitrarily many poems (for example, a Twitter bot), we provided a single generated poem. In cases where the generated poem was excessively long, a 1-page excerpt was provided. While excerpting may bias judge responses to long poems, this is relatively unimportant to our analysis.

Once all 30 poems had been rated, we began the qualitative portion of the study. Each judge was asked to go back over the poems and, for at least 3 poems, write an explanation for their judgment. In order to adhere to CAT best practices, we did not present this request until the judges had made all 30 quantitative ratings and did not permit them to change their quantitative answers once the qualitative portion began; thus, the request for written explanations was prevented from affecting the results of the CAT itself.

We obtained a total of 7 judges, which is within normal bounds for the CAT (Baer and McKool, 2009). Participation took 1 hour.

We analyzed our data in three steps. First, we calculated the intraclass correlation between the seven judges. Second, we used the Kruskal-Wallis test to see if there was a difference between ratings of poems from groups A, B, and C. Third, we used open coding to determine what major factors were used by our judges in their qualitative evaluations.

4.2.4 Results

4.2.4.1 Interrater reliability

Unfortunately, the intraclass correlation between our judges (a statistic that can range from -1.0 to 1.0) was only 0.18—far below the 0.7 to 0.9 standard for CAT results. A bootstrapped sample of 10,000 permuted versions of the data showed that interrater agreement hovered around zero, with a standard deviation of 0.04, meaning that if each judge gave their ratings by chance, there would be much less correlation between ratings than what our results show. The agreement between our judges, therefore, is statistically significant at $p < 0.01$, but the actual amount of the agreement is small. Therefore, it is not a strong enough agreement to be used for the usual applications of the CAT, such as judging admissions to academic programs in creative fields.

Looking at the data, some poems were rated very highly by nearly all judges, while others were rated very poorly, but a large mass of poems in the middle had inconsistent or inconclusive results. When the data is reduced to those poems with the highest and lowest average scores, intraclass correlation becomes very high. With the seven best and seven worst poems—nearly half of our original data set—the intraclass correlation is 0.73, and narrowing the number of poems raises that statistic still higher. It is intuitive that poems with the highest and lowest ratings would have relatively good agreement, while poems about which judges disagreed would have average scores closer to the middle. However, when we looked at only the seven best and seven worst in each of our 10,000 bootstrap samples, the mean and standard deviation for intraclass correlation did not rise. Therefore, the agreement on the best and worst poems is not merely a statistical artifact; our judges really were able to agree on these ends of the spectrum, but simply disagreed about the rest of it.

It appears that our judges agree when selecting the best and worst poems in a group, but cannot reliably rank the group of poems as a whole. Calculating Kendall's tau between pairs of judges did not

result in any useful clusterings of the judges into factions.

4.2.4.2 Kruskal-Wallis test

While a disproportionate number of the best-ranked poems were from Group B, a Kruskal-Wallis test showed that this difference was not significant ($\chi_2^2 = 3.8$, $0.25 > p > 0.1$). We used the Kruskal-Wallis test because we were interested in the poems' ranks rather than their numerical scores. Both relatively creative (from a Product perspective) and relatively uncreative poems existed in groups A, B, and C, and no group did systematically better than others.

The individual poems, their group membership, and their average scores are shown in Table 4.5.

Grouping	**Category**	**% of comments**	**% Positive**	**% Negative**
Reaction	Feeling	12%	68%	31%
Reaction	Comparison	10%	44%	44%
Reaction	Base/Other	12%	57%	29%
Meaning	Message	14%	60%	40%
Meaning	Coherence	7%	67%	33%
Meaning	Content	5%	89%	0%
Craft	Technique	12%	67%	29%
Craft	Imagery	7%	67%	33%
Craft	Form	4%	50%	50%
Craft	Skill	3%	20%	80%
Novelty	Novelty	15%	42%	58%

Table 4.6: Categories derived from our qualitative data.

4.2.4.3 Open coding of qualitative data

We performed open coding by giving each line of written response a content label and a valence (positive, negative, or neutral), then clustering the content labels into categories. Each category gave insight into the implicit values used by our judges. Overall there were 179 coded lines distributed over 63 explained judgments—an average of 2.8 lines per judgment and 9 judgments per reader. Table 4.6 shows the proportions of lines of each type and their valences.

One coder performed the initial clustering, while a second repeated the labeling to validate the first coder's responses. Our two coders agreed on roughly half of lines as to the exact categorization, and for two thirds of lines agreed as to the general category. For the remaining one third, in half of cases, the reviewers agreed that a line could easily be coded as having both reported categories, such as cliché imagery, which has to do with both novelty and imagery. Of the remaining ones, the coders did not initially agree, but were quickly convinced of one or the other categorization.

Below we explain the meanings of each of our labels.

Reaction. 34% of lines described, structured, or contextualized the judge's affective reaction to the poem.

Feeling. Statements about the emotions evoked in the judge by the poem.

- "This was just super fun to read."
- "Felt empty."

Comparison. Lines that compared the poem to something else, including existing poems or poetry movements.

- "I like how this echoes, say, Siri giving instructions."
- "This reminds me of bad, early 2000s my space poetry... Angsty teens spewing 'creativity' on the world wide web."

Base/Other. Base lines are statements that the poem is or is not creative, without immediate explanation. Often a judgment containing Base lines contained explanation in other lines, so a Base line can be thought of as a topic sentence, not necessarily an unsupported judgment. Lines coded "Other" similarly contain statements more to do with structuring a judgment than with the judgment itself, such as statements that the judge felt conflicted about the poem.

Meaning. 26% of lines described the meaning of the poem, the concepts involved, and their clarity.

Message. Statements about the idea that the judge believes the poet intended to communicate. Most judges made comments with negative valence, stating that a poem was not sufficiently meaningful. A major exception was judge 7, who left long positive comments closely interpreting the meaning of several poems.

- "It would be more creative/interesting if there were a distinct theme or repetition of some sort—some sort of message to the reader."
- "It begins with an opinion about a campy TV show and ends on gleeful nihilism. The real American Horror Story is the nuclear apocalypse, the end of the world effected by some hideous war games between two self-obsessed nations flexing their muscles at each other. (good twerking). It packs a lot into a very compressed collection of sentences, and also manages to serve as a brutal indictment of contemporary culture."

Coherence. Nearly all of the lines we coded as Coherence referenced a lack of coherence, or nonsense.

- "It felt too disjointed."
- "This one is interesting because it doesn't make any sense, yet I'm intrigued by it."

Content. Statements about the characters, objects, or events in the poem. Judges mentioned this aspect of meaning less frequently than more abstract ideas.

- "It's a complete narrative in just 3 very short lines."

Craft. 20% of lines described the way in which the poem's concept was executed.

Technique. Statements assessing specific literary techniques used in the poem. They include defamiliarization, enjambment, phrasing, repetition, rhyme, rhythm, vocabulary, and voice, as well as more general statements such as "playful use of language". Poor technique was coded negative.

- "I like this one because I feel like I can hear a distinct voice."
- "I found the creative intentions - caps, quotation marks, the fragmentive narrative, the asterisks - forced and not really used well."

Imagery. Statements commenting on the poem's use of imagery. Imagery is a specific type of content involving direct sensory descriptions, and is important in contemporary poetry (Kao and Jurafsky, 2012).

- "Good consistent imagery and figurative language."
- "The imagery isn't provocative."

Form. Only a few poems received comments on their form. Three poems in inventive forms, such as imaginary dictionary entries, were praised for these concepts. Another received a comment that it was too short. In addition, two haikus received negative comments for lacking subtle features of the traditional haiku.

- "If there were another stanza I'd like it more."

Skill. Statements assessing the poet's skill or cleverness.

- "Very rudimentary and woe is me."

Novelty. 15% of lines were statements about the poem's novelty. Positive valence lines stated that the poem was unusual, unique, or subversive. Negative valence lines stated that the poem—or aspects of the poem—were obvious, derivative, unoriginal, trite, clichéd, banal, failed to push boundaries, or did not sufficiently change their source text.

- "I don't think this is very creative because it doesn't push the boundary of poetry in any way."
- "This is creative because its unique. I've never seen a poem like this before."

Judges frequently disagreed on what traits a poem possessed, and on the valence assigned to those traits. A poem might be described as incoherent by one judge but interestingly disjointed by another, or banal by one judge but unexpected by another. An extreme example is a poem generated by Mobtwit (Hartlová and Nack, 2013). The poem was written by arranging tweets to generate emotional contrast. It was described as random and devoid of meaning by Judge 1 (who rated the poem as 1 out of 5), but Judge 7 rated the poem a 5 and gave a long exposition of its meaning (quoted above, under "message"). Restricting the sample to the seven highest and seven lowest rated poems did not remove these qualitative disagreements.

There were slightly more positive (93) than negative (79) lines overall. There was a slight positive correlation between quantitative score and number of positive comments, a similar slight negative correlation between quantitative score and number of negative comments, and no correlation at all between quantitative score and total comments ($r = 0.26$, -0.27, and 0.008 respectively).

4.2.5 Discussion

Since we did not achieve the usual inter-rater reliability standard of the CAT, our method is not a finished evaluation. It is possible that the CAT will not provide standardized computational poetry evaluation at all. However, the qualitative portions of our study illuminate how judges with some expertise evaluate computational art, which leads us to a better understanding of what criteria could go into such an evaluation in the future.

4.2.5.1 Judge selection

Why did our judges disagree about poems in the middle of the set? Should we have chosen a different set of judges? We believe that our judges' lack of interrater reliability speaks to something more complex than a simple lack of expertise.

The question of who, exactly, has sufficient expertise for the CAT is a difficult one. Kaufman *et al.* review prior work in the differences between expert CAT judges and novices (Kaufman et al., 2009). Novices lack the interrater reliability of experts and their judgments only moderately correlate with expert judges. However, in many cases, gifted novices (which Kaufman *et al.* describe as "quasi-experts") produce judgments that are more in line with those of experts than with the general population. Novices have fewer problems serving as judges when the art form in question is one that the general population encounters regularly: stories rather than poems, for instance. However, psychologists—even psychologists of creativity—are not experts; they perform as inconsistently as novices from the general population. The expertise necessary for the CAT seems to have more to do with experience in a specific creative field than with knowledge of the theoretics of creativity.

Pearce and Wiggins used both music researchers and music students as judges (Pearce and Wiggins, 2007). Why did their experiment achieve close to the recommended interrater reliability while ours did not? One answer is that Pearce and Wiggins' study was an evaluation of chorale melodies, which are simpler, less diverse, and defined by more well-established rules than computational poetry.

We argue that Experimental Digital Media students should be considered quasi-experts. Even more advanced than Kaufman *et al.*'s gifted novices, these students are more like experts-in-training, undergoing advanced education in how to produce and critique art in their field. However, the field of digital poetry is too new to be well-defined. It is also possible that the different poets in our study are performing different tasks that ought not to be grouped together. The CAT's more typical uses revolve around homogeneous products, such as the poetry or collages of elementary school students. Mature artists and researchers, in a new field where a variety of movements, motivations, and techniques are still under development, likely produce a more complex and contentious body of work, and in fact, artists from movements with very different goals (such as computational creativity researchers vs. poets who work in a humanities department) were combined together into this one assessment.

A good idea for future work might be to replicate the CAT with other groups of experts and quasi-experts, or with a more homogeneous group of digital poems. Poets who have been paid for their published work, or participants in events such as the E-Poetry Festival (Glazier, 2016), might be appropriate experts.

However, we strongly advise against the use of computational creativity researchers as expert judges unless they themselves are practicing artists in the field being studied. Computer researchers without such artistic experience are likely to have the same problem as psychologists judging human art. They may

thoroughly understand the theory, but they are unlikely to have an expert sense of the *artistic* aspects of their work. Moreover, because academic publishing depends heavily on theory and argumentation, and because the field of computational creativity is so new, computer researchers (including ourselves) are likely to be distracted from evaluations of specific products by our beliefs about where we would like the field to go.

4.2.5.2 Judge bias

Recalling the mixed results on the existence of bias against computational art, it is worth asking if our XDM students exhibited this bias. We did not attempt to empirically measure bias, but it is worth noting that, anecdotally, they did not appear to have one. At no point did a student express, in their freeform comments, that a poem was less creative because the author was not human. If anything the bias was in the other direction. As one judge put it after the experiment, "Sometimes I wanted to say that a poem was childish, but then I thought, 'What if a computer wrote it?' and I didn't want to hurt the computer's feelings."

Even where bias against computers exists, it is not relevant unless computer products are compared with the products of humans and the judges are somehow aware of which products are from which group. All the poems in our study had some involvement from both humans (who wrote a computer program) and computers (which put together words based on the program), and judges were not told what the computer's specific Process role was. It is easy to imagine studies where the role of the computer is more homogeneous, or even studies which compare outputs from different versions of a single program.

4.2.5.3 Desiderata for domain-specific poetry evaluation

Baer argues that creativity is an umbrella term for a variety of independent domain-specific skills (Baer, 1998). If this is the case, then evaluations of computational poetry would be expected to contain criteria that apply only to poetry, perhaps only to computational poetry. Studies like ours are a step towards developing these criteria.

Our study suggests a set of desiderata shared by most of our judges for poetry:

- **Reaction.** The poem should provoke feelings of enjoyment and/or interest from the reader.
- **Meaning.** The poem should intentionally convey a specific idea. Even if the poem is difficult to understand, its difficulty should enhance the underlying meaning. (For example, a Dadaist poem uses apparently meaningless text to illustrate ideas about how language and meaning work.)
- **Novelty.** The poem should be unusual or surprising in some way, and not merely repeat familiar tropes.
- **Craft.** The poem should make effective use of poetic techniques in service of the other three criteria. This can include form, imagery, auditory effects such as rhyme, psychological effects such as defamiliarization, visual effects such as enjambment, and verbal effects such as voice. Effective use of these techniques requires skill.

These desiderata are not straightforward. In particular, some of the literary techniques praised by our judges oppose each other. At least one poem received positive comments for its detailed imagery, while another received a positive comment for simplicity. Requiring detail and simplicity at the same time is a contradiction!

We suggest viewing literary techniques as a toolbox of strategies for poetic success. Some may be more appropriate to a particular goal than others. The question asked to a judge about craft should not be, "How many times are literary techniques used?" It should be something more like, "What techniques are used, and how effective are they?" It should be assumed that such questions can only be answered by expert or quasi-expert judges.

4.2.5.4 Relations between our desiderata and existing theories

Our desiderata have overlap with other evaluation theories, but are not identical to them. For example, Novelty and Value are frequently used to evaluate creativity. Our judges did emphasize Novelty, but Value either did not appear *per se* or was divided into many sub-criteria.

Van der Velde *et al.* (2015) use word association to define creativity criteria. Our Novelty criterion corresponds to their Original and Novelty/Innovation, while their Skill and Craftsmanship correspond to our Craft. Van der Velde *et al.*'s other criteria are Emotion and Intelligence.

Ritchie suggests Typicality and Quality as criteria (Ritchie, 2001). A hint of Typicality can be seen in Comparison judgments. It appears that for a positive typicality judgment, a poem must strike the judge as not merely typical of poetry, but typical of *good* poetry. The "Dadaist dictionary" was rated highly, but poems typical of "the scrawl of a high school senior" were not. Groundedness in relevant poetic movements led to a positive response, but so did poems seen as entirely novel. Typicality as Ritchie conceives it may be neither necessary nor sufficient for computational poetry.

The Creative Tripod (Colton, 2008) consists of Skill, Imagination, and Appreciation. While Skill as such was a minor category for us, everything under the Craft grouping presumably requires skill. Imagination was rarely mentioned, but it could be argued, as by Smith *et al.*, that Imagination is the underlying trait which allows for Novelty (Smith et al., 2014). This reading is supported by Van der Velde *et al.*, who group "Imagination" under Novelty/Innovation (van der Velde et al., 2015). Appreciation is difficult to read into any of the coded comments. However, the highly fluid definitions of traits in the Tripod make it difficult to definitively state if they are present or not.

Manurung *et al.*'s criteria of Meaningfulness, Poeticness, and Grammaticality (Manurung et al., 2012) overlap with our desiderata. Meaningfulness and Meaning are synonymous; Poeticness and Craft are similar concepts. However, Manurung *et al.* operationalize Poeticness as meter and rhyme, while judges in our study had a more expansive view of Craft. Grammaticality was not emphasized; some poems received positive comments despite being quite ungrammatical. However, Grammaticality is related (though not identical) to Coherence, and judges did note when poems were incoherent or did not make sense.

Our Reaction criterion does not appear in many existing models, since most models focus only on qualities imputed to the poem or poet. However, it bears some resemblance to the Wellbeing and Cognitive Effort criteria of the IDEA model (Colton et al., 2011), which could perhaps be used to break judge reactions down more finely. In some sense, our Reaction criterion is a sign of a Press judgment, at the scale of the individual person, being incorporated into a task that we assumed would be about Product.

This porousness between Press and the other perspectives is inevitable in any human judgment since humans are culturally situated agents who bring their emotions and complex full-brain reactions, to some degree, into any judgment they are asked to make.

4.2.5.5 The Product or the Process?

We believe that Product evaluation is an important part of creativity assessment. Nevertheless, it has a major drawback: it cannot differentiate between the creativity of the computer system and the creativity of the human who programmed it.

Some examples from our data set illustrate this problem. "Notes on the Voyage of Owl and Girl", our most highly rated poem, is a Mere Generation poem based on a tightly handcrafted template. The human author provides a narrative structure which does not alter, and the computer selects details (from a human-curated list) to fill it in. In its original form, "Owl and Girl" exists on a web page and is periodically re-generated before the viewer's eyes. "Owl and Girl" is interesting artistically, but its high ratings refer mostly to the creativity of the human author.

Conversely, "The legalized regime of this marriage", created by Stereotrope, is among the most poorly rated. Stereotrope is an experiment in computational linguistics from a leading member of the computational creativity community. The system mines existing text for similes, produces a common-sense knowledge base using these similes, and uses the knowledge base to generate similes and metaphors of its own. The new similes and metaphors are then used to fill in templates and construct a poem. Our judges disliked this poem, calling it obvious, clichéd, unskilled, and uncreative. However, Stereotrope is doing something more *computationally* interesting than "Owl and Girl".

We must ask what the goal is for a system like Stereotrope. Do we wish to construct a system whose use of simile and metaphor is artistically successful? Then Stereotrope—in its current form—fails. But if we wish to construct a system using *humanlike* simile and metaphor, then it is easy to argue that Stereotrope succeeds: its metaphors feel obvious *because* they are humanlike. Such a system might not be artistically creative from a Product perspective, but it might be a good model of the everyday creativity of non-artist humans expressing themselves. We will not know if a system has succeeded unless we know which of these goals it was aiming for. (Other goals than these are, of course, possible.)

This problem of evaluation recalls Pérez y Pérez's distinction between the engineering-mathematical approach and the cognitive-social approach. From an engineering-mathematical perspective, it makes no sense to say that a poetry system is successful if its poetry, on a Product level, does not succeed. From a cognitive-social perspective, a system writing "bad" poetry could still make important contributions towards modeling and understanding the cognition of an ordinary human poet—in fact, it might be significantly more useful than a system that creates novel and valuable poetry through an inhuman process.

Our own belief that ultimately, computational creativity must succeed on both fronts. To set a Product goal while ignoring Process is to invite pastiche and handcrafting at a level which makes it difficult to argue that the computer (as opposed to its programmer) is creative, no matter how excellent its work. But to set a Process goal while ignoring Product is to fail to take seriously the very medium in which the computer is working. A system which fails to take art seriously can still have value as a cognitive model, but that model will not represent the cognition of skilled human artists.

One could argue that a computer must first establish a humanlike process before refining that process to be more artistic. This is reasonable, but debatable. It is also possible that producing good art and

using a humanlike process are two tasks at which the computer can progress simultaneously. The learning process of an initially-uncreative computer may or may not look like the learning process of an initially-unskilled human, and setting a goal of behaving like an unskilled human may in some circumstances be counterproductive.

An interesting idea for future work would be to replicate the CAT study and present information about the specific tasks assigned to the computer, in a standardized form such as the diagrams suggested by Colton *et al.* (2014). CAT judges would then be asked how creative they believed *the computer* had been. An alternative would be to use one evaluation technique for Product, and another for Process.

Chapter 5

TwitSong: Developing a computational poetry system

After all of the prior theoretical work and investigations into the evaluation of creativity, we now turn our attention to our own computational poetry system, TwitSong. Although it makes the most sense conceptually to place this research after the other projects, in practice, all of them were going on at once and work on TwitSong began long before all of our theoretical framework was in place.

TwitSong has gone through three major incarnations, all different from each other. Our initial inspiration for the project was generation of song lyrics based on an arbitrary source text. Like others who have used the news as a starting point (Toivanen et al., 2014; Tobing and Manurung, 2015; Rashel and Manurung, 2014b), we were delighted by the design fiction of a system that could greet users by summarizing the morning's news in a light-hearted, perhaps mildly satirical poem or song. Moreover, as one of our supervisors has previously done work in the area of social media (Yang et al., 2016; Tan et al., 2015, 2016), we liked the idea of using Twitter as a source text to summarize people's varying *reactions* to the news, which are otherwise quite difficult to summarize, in poem form.

Several existing poets, both human and computer, served as inspiration for our work with TwitSong. The most notable is Ranjit Bhatnagar's Pentametron (Bhatnagar, 2012), which creates sonnets out of tweets that happen to be in iambic pentameter. We wondered if we could improve on Pentametron by adding a layer of optimization to pick the tweets that not only have correct rhyme and meter, but are also more poetic or appropriate in other ways. Another inspiration was the New York Times Haiku project (Harris, 2013), which demonstrates that found poetry based on the news is viably interesting to readers.

5.1 Generation one: Line selection proof of concept

A prior version of this section appeared at the 2015 BRIDGES conference (Lamb et al., 2015b).

5.1.1 Introduction

At first, our goal with TwitSong was to improve on Pentametron by adding optimization. The basic algorithm in mind was:

- Collect a large corpus of Twitter data on some topic
- Select the lines from the corpus that have the appropriate number of syllables
- Group these lines into rhyme sets
- *Rate* the lines in each rhyme set on a combination of poetic criteria, such as emotion or imagery
- Create a poem out of the highest rated lines.

However, we were concerned that, if this algorithm failed to make good poetry, we would not know how to interpret this result. Would it mean that the algorithm itself was unsuitable? Or that we had chosen the wrong poetic criteria? Or, perhaps, that we had simply operationalized our criteria ineffectively?

In order to discern the suitability of our algorithm and criteria, we therefore decided not to automate our line ratings at first. Instead, the first generation of TwitSong used crowdsourced human judgments to produce line ratings. Evaluation of this generation had promising results.

When we created the first generation of TwitSong, we had not yet derived our four-part Product model for evaluating poetry, which we discuss in Section 4.2. Thus, we somewhat arbitrarily chose three criteria:

- Topicality (the extent to which the line reflects the poem's topic)
- Sentiment Polarity (the intensity of positive or negative emotions in the line)
- Imagery (the extent to which the line contains concrete, sensory information)

These criteria are not completely arbitrary. Topicality is logically an important quality for a poem that describes a topic. Expressing a coherent emotion may also seem like a logical poetic goal, although some researchers (Hartlová and Nack, 2013) prefer to select different contrasting emotions. Imagery is anecdotally one of the skills beginning poets are taught to improve. Both Kao and Jurafsky (2012), studying the differences between professional and amateur contemporary poetry, and Simonton (1990), studying the differences between famous Shakespeare sonnets and less successful ones, both found that concrete sensory imagery, along with lexical complexity, is one of the most important traits that explain this difference. However, one goal of this study was to test whether each of these three criteria did, in fact, contribute to a more successful poem in the context of found Twitter poetry.

We also created a Combined metric which was calculated simply by adding the normalized scores for Topicality, Sentiment Polarity, and Imagery together. Since Sentiment Polarity can be either positive or negative, we choose based on the topic which score to add to the Combined metric for each data set. We err on the side of choosing positive sentiment, except for gloomy topics (such as climate change) for which a positive sentiment is inappropriate.

Rather than using our methods only on lines in iambic pentameter, we made use of all lines with 10-11 syllables. Our reason for this was that, when trying to select lines for strict iambic pentameter, too many tweets were removed and there was not enough data left over to perform the rest of the algorithm. The issue of meter is one to which we will return in Section 5.3.

5.1.2 Method

5.1.2.1 Specifics of the TwitSong system

The first generation of TwitSong uses the Twitter API to gather tweets from a specific time period and filters them based on a topic keyword or other regular expression. The topic keyword is chosen by the programmer. (Another option would be to filter based on hashtags. We chose not to use this method because we estimated based on our own Twitter experience that there could be a large number of tweets about a specific topic which did not use the hashtag for that topic.) Tweets that are not in English are removed, and excessive hashtags and other non-syntactic features of tweets are removed from them.

It might seem odd that we are both selecting a topic by keyword and optimizing for topicality, but this appears to be necessary. For example, one of our data sets concerns the topic of New Year's Day, 2014. All tweets in this data set were posted on December 31, 2013 or January 1, 2014 and all contained the string "2014". However, not all tweets containing the string "2014" are actually tweets about New Year's Eve celebrations. Also, sentences within a tweet, rather than the entire tweet, can be used, and excessive hashtags (including the hashtag #2014 and related ones) can be removed. Therefore, not every sentence processed by TwitSong contains the string "2014". This is a good thing, as the poems would otherwise be quite repetitive. While the New Year's data set contains many tweets about New Year's Eve celebrations, it also contains sentences about other topics, spam, and meaningless strings of numbers. This is typical for our Twitter data sets across a number of topics. Therefore, optimizing for topicality is still useful.

Syllable count and rhyme endings are calculated using Hirjee and Brown's (2010) RhymeAnalyzer, a set of very powerful rhyme and rhythm detectors which are built on top of the CMU Pronouncing Dictionary. We modify Hirjee and Brown's algorithms slightly for use with Twitter. These modifications include shortening drawn-out words (e.g. "hiiiiiii"), detecting compounds, and using a hand-written supplement to the dictionary for topical or Twitter-specific words such as slang, brand names, common misspellings, and proper names. For words that cannot be pronounced using one of these methods, TwitSong rejects the tweet instead of trying to sound it out, even though RhymeAnalyzer has the capability to sound out arbitrarily complex strings of letters; we found that, if we used this capability, we ended up with too many meaningless strings of letters or numbers being rhymed with each other.

Tweets with the right number of syllables are grouped into RhymeSets of two or more rhyming lines. Long tweets with too many syllables are separated into their component sentences before processing. A RhymeSet is based on a seed line, plus any other line that is sufficiently close to the seed line in rhyme and meter. RhymeAnalyzer allows for close but potentially inexact matches ("slant" rhymes) between lines in the set; we specify the level of allowable slant rhyme by hand.

The tweets in the RhymeSets are then scored by topicality, sentiment, and imagery. In this generation of TwitSong, the scoring is done via Crowdflower. To create the final poem, Twitsong retrieves RhymeSets with the desired number of syllables and selects the seven best pairs of rhyming lines (i.e. the top two lines from the top seven RhymeSets). If the second-best line in a RhymeSet ends with an identical word to the best line, then TwitSong goes further down the list until it finds a more acceptable rhyme. These pairs are then assembled into the form of a Shakespearean sonnet (three quatrains and a couplet, with the rhyme scheme ABAB CDCD EFEF GG) The ordering of rhyming pairs within the sonnet is arbitrary; for the purposes of this generation, we place the best rhymes last.

5.1.2.2 Line ratings on Crowdflower

Metric	Highest Rated		Lowest Rated	
Topicality	5	2 teams to go #Sochi2014	1	One day he gone say you crowding my space
	5	Way to go USA Men's Hockey team	1	73205
	5	Sochi Winter Olympics day six live	1	i have done SOOOO much work this afternoon!!!
Sentiment	5	I smile when you smile...I love when you care. :)	1	i hate how people judge me on my size.
	5	Love this sport #Olympics2014	1	WERE STUCK IN A SHITTY ANIME DEAN
	5	The Olympic free skating is so cool!!	1	hey fuck Anthony , everyone hates him
Imagery	5	Food. Food. Food. Food. Food. Food. Food. I love food	1	! ! !!!!!!! !!!! #2014 #sochi2014
	4.67	15 Pictures That Will Make Your Heart Stop	1	something George Costanza would think about.
	4.67	Sochi Olympic Park As Seen From Space	1	You mess one section up and you pay f
Combined	12.67	Love this sport #Olympics2014	4	hey fuck Anthony , everyone hates him
	12	The Olympic free skating is so cool!!	4	Nobody owes anyone anything
	12	Figure Skating judges give it a 9	4	but when I do it, I'm being a dick

Table 5.1: Examples of some of the highest and lowest-rated tweets for all three scoring metrics from the Olympics dataset. Theoretically possible Combined scores range from 3 to 15; other scores range from 1 to 5.

We use Crowdflower[1], a crowdsourced microtasking service, to gather human judgments. Each tweet for each topic is scored by three Crowdflower workers (the number that Crowdflower's documentation recommends for most tasks) on a five-point Likert scale on three metrics: sentiment polarity (very positive to very negative), topicality (very relevant to very irrelevant), and concreteness (very concrete to very abstract). We train the workers on the meanings of these terms, especially concreteness/abstractness, which is not a typical concept used by microtasking workers. The exact questions are as follows:

- Topicality: "How relevant is this tweet to the topic of [topic]?" (Very Irrelevant to Very Relevant)
- Sentiment: "How positive or negative are the sentiments in this tweet?" (Very Negative to Very Positive)
- Imagery: "How abstract or concrete is this tweet?" (Very Abstract to Very Concrete)

The 3 scores given to each tweet on each metric are then averaged. Table 5.1 shows examples of high and low-rated tweets on each of these metrics.

[1] http://www.crowdflower.com

5.1.2.3 Experimental groups of sonnets

To test the suitability of our TwitSong algorithm, we generated sonnets based on the processes above, based on four topical data sets. Our topics were New Year's Day 2014, the 2014 Winter Olympics, the 2014 Oscars, and climate change.

For each data set, we generated five different sonnets based on five different ways of rating the lines in the data set: Topicality only, Positive Sentiment only, Negative Sentiment only, Imagery only, and Combined. The result was a set of $4 \times 5 = 20$ computationally generated sonnets. To this data, we then added two control poems for each topic. The first control poem, intended to serve as a lower bound, was constructed by TwitSong without using ratings. When ratings are not available, all potential lines are implicitly rated 0. Thus, TwitSong assembles the first seven valid pairs of rhyming lines it encounters without any attention to their content or meaning. These control poems are meant to be similar to the output of Pentametron, in that only rhyme and meter and not content are considered; however, in practice, since the lines are taken from a data set defined by topic keywords, they are likely to still be somewhat more cohesive in content than Pentametron's poems.

The second control poem, intended to serve as an upper bound, was made by the author of this thesis, who is a published poet, and who manually chose appropriate lines from the RhymeSets that were available. This brought the total number of poems in the experiment to $4 \times 7 = 28$.

Excerpts of TwitSong's output, and of control poems, are given in Table 5.2.

Human	**Control**
In 2014 I'll talk less and listen more	PLEASE FOLLOW ME ? ILY GUYSS 25
live a little more and stress a little less.	Skies the limit #NewYears #2014 #BelAir
Never give up,Do it better than before	I LOVE YOU VERY MUCH MY ANGEL ¡3 5
Oh and try to lose some weight in the process	Oh hey, it's 2014. #Ireallydontcare
Combined	**Negative**
Hey Nashville...2014 is pretty awesome!	Nothing's changing except people I fuck with
Happy 2014 friends! Be safe out there!!	2014 already took Uncle Phil
Had a great New Years Eve at Magic Kingdom	2014 already startin off with death
We started off 2014 with a prayer	started my 2014 off vomiting brill.

Table 5.2: Excerpts from poetry used in our study. All poems are from the New Year's Day 2014 dataset. The Human poems were put together by a human from the tweets available, using TwitSong only to create sets of possible rhyming lines to choose from. The Control poems were put together by TwitSong through arbitrary selection of lines from these sets. Also shown are poems made by TwitSong using the Combined metric (the sum of the topicality, positive sentiment, and imagery scores), which performed well, and the negative sentiment metric, which performed very poorly. For space reasons, we include only a single stanza from each poem; the full poems are 14 lines long and in sonnet form.

5.1.2.4 Evaluation.

From the beginning, we knew that rigorous evaluation based on falsifiable hypotheses would be important for the development of a poetry system. However, when we developed and tested the first generation of

TwitSong, we had not yet performed our survey of evaluation techniques, nor our study of non-expert poetry evaluation from Section 4.1.

For the purpose of evaluation, we used a simple pairwise comparison metric. Human raters were recruited using Crowdflower and each rater was given four pairs of poems. For each pair of poem, the rater was instructed to indicate which poem they preferred, and to justify their choice in one sentence or less.

To ensure data quality, a filtering mechanism was then used. Of each rater's four pairs of poems, two were control pairs—comparisons of a human-constructed Twitter poem with a control poem (made arbitrarily by TwitSong without using any ratings) on the same topic. If a rater did not prefer the human-constructed poem to the control poem in both of their control pairs, they were then removed from the data. The intuition behind this filtering mechanism is that a human-constructed poem, in "ground truth", should be better than a poem of essentially random lines.

We checked the validity of the control pairs as a data cleaning mechanism by asking the two supervisors of the author of this thesis to blindly judge each possible control pair. Both distinguished human-constructed poems from control poems with perfect accuracy, and both preferred the human-constructed poem in each control pair. In theory, raters who were answering at random would still have a 25% chance of passing this test. In practice, our filtering mechanism removed about half of the data, which means that approximately two-thirds of the data left can be trusted to be non-random. That half of the raters did not perform to specifications is not overly surprising, given known issues with the reliability and performance of crowdsourced workers (see, for example, (Kittur et al., 2013)).

There are other conceivable reasons, besides answering at random, why a particular rater might prefer the arbitrary control poem to a human-constructed poem. However, for the puposes of this study, we were interested in constructing computational poems which share the traits that make human-constructed poems more meaningful and entertaining than arbitrary assortments of tweets. Thus, we were interested only in the opinions of raters who showed a clear preference for human-constructed poems.

Following this data cleaning, we were left with a set of pairwise comparisons in which each computationally generated type of poem (those chosen for topicality, imagery, positive and negative sentiment, the combined metric, and the control poems) appeared between 100 and 120 times. For each individual poem, we counted the number of times that the poem was selected in preference to the one next to it, and divided this by the total number of times that the poem appeared. We then ran a one-way ANOVA to test for significant differences between the six varieties of poem.

5.1.3 Results

5.1.3.1 Scoring

One concern for us was the potential accuracy or inaccuracy of line ratings from Crowdflower. We therefore ran these ratings through a few informal tests.

First, we informally looked at the tweets sorted from smallest to largest on each metric. The distribution of which tweets were scored as most topical, off-topic, happy, sad, or neutral accorded very closely with our own intuitions. For example, the tweets which consisted of arbitrary numbers were consistently rated as very off-topic, and tweets expressing joy or strong negative emotions were found at the appropriate ends

of the sentiment spectrum. The scores for abstract vs. concrete imagery also accorded somewhat with our intuitions, but we noticed some irregularities in the data. The data was very biased towards rating tweets as abstract. The average rating for a tweet was below 2 on a scale of 1 to 5 (1 being most abstract, 5 being most concrete), with more than half of the tweets rated as 1. Very few tweets were scored as being highly concrete. Table 5.1 shows examples of the highest and lowest-scored tweets on each metric, using the Sochi Olympics dataset.

We also tested each scoring metric by taking a subset of tweets and calculating the correlation between average Crowdflower rating and the manual rating of one of our authors. Again, crowdsourced workers gave results very similar to our own ratings for topicality and sentiment ($0.77 < r < 0.81$), but less so for imagery ($r = 0.37$).

5.1.3.2 Evaluation

Figure 5.1 shows the results of our pairwise evaluations for each poem. Human raters preferred the topical, positive sentiment, and combined poems to control poems. (Because the human poems had been used for data cleaning, we could not validly include the results for these poems.) Imagery-based poems were rated slightly better than controls. To our surprise, negative sentiment poems performed worst of all, being selected in only 27% of pairs on average (compared with 41% for control poems).

A one-way ANOVA demonstrated that these differences were statistically significant, $F(5, 18) = 5.79$, $p < 0.01$. Post hoc tests were performed using Tukey's method, showing that positive emotion, topicality, and combined poems performed significantly better than control poems; imagery and negative emotion were not significantly different than controls.

Although negative emotion in and of itself did not produce a significant effect, raters' written comments indicated that human raters often reacted very negatively to poems they saw as negative, depressing, angry, sarcastic, or crude. Comments like "Poem A is very negative and makes me angry reading it," and "poem A has too many offensive words" were common.

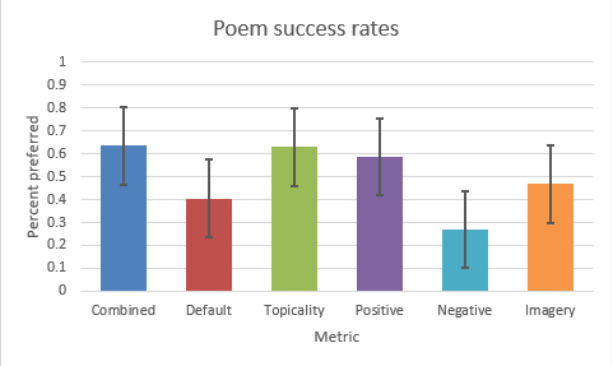

Figure 5.1: Success rates for types of computationally generated poems in pairwise comparisons with other poems. The height of a given bar represents the number of times a poem from that category was selected in preference to any other poem, divided by the number of times a poem from that category appeared in a comparison. Error bars represent 95% confidence intervals.

5.1.4 Discussion

Our work demonstrates quantitatively and with reasonable controls in place that selecting lines based on simple metrics like the ones we have chosen can significantly improve the appeal of the poem to a general audience, and that some metrics perform better than others. In other words, our proof of concept for TwitSong is successful, and we had every reason to believe that, with proper implementation, a version of TwitSong which algorithmically rated and selected its lines would also perform well.

Although it may seem obvious, it is notable that the Combined metric performs best. Colton *et al.* (2012) report anecdotally that combining more than one metric diluted the style of a poem. We did not experience this effect with TwitSong; instead, poems using a combined metric performed as well as poems using the highest performing single metric. One possible explanation for the difference between our results and Colton *et al.*'s is that TwitSong's form is already quite constrained in terms of style: a particular meter and rhyming scheme, and only phrases that have already been used by a human on Twitter, etc. The metrics that are applied in this case may be more general and less in conflict with each other than metrics that govern individual word choice. Alternatively, since Colton *et al.*'s assertion about style dilution was not tested or backed up with evidence, it may be a subjective judgment which is not replicable.

Poems selected by imagery did not perform significantly better than controls. We believe this is due to the poor performance of crowdsourced workers at correctly identifying tweets with concrete imagery. Workers on Crowdflower reported, to a much higher degree with this than the other tasks, that it was too difficult. (The imagery task was rated an average of 3.125 and 3.175 out of 5 on Instructions Clear and Ease of Job, respectively; compared to 4.03 and 3.98 for topicality and 4.06 and 4.26 for sentiment.) Given that many beginning poets struggle with concrete imagery, it is perhaps not surprising that people without a poetic background could not be quickly taught to identify it. Such results run counter to our original assumptions, that crowdsourced workers would do better at scoring tweets based on their meaning than a computer. While the workers were good at identifying topicality and sentiment, it is plausible that a specialized resource, such as the dictionary of primary process imagery used by Simonton (1990), might do better at identifying concrete imagery than most humans.

An alternate explanation for the poor performance of imagery might be the relative dearth of tweets with good imagery in them. For instance, in the Olympic data, only 96 out of 333 lines were rated more than 3 out of 5 by Crowdflower workers, and 61 rated more than 3 out of 5 when rated manually by the author. With few or no good imagery tweets to choose from, it might simply be more difficult to use imagery to put a good poem together.

The fact that we used non-experts as judges, given our results in Section 4.1, call our evaluation's validity into question. Certainly, if we had performed that experiment before testing the first generation of TwitSong, we would have attempted to recruit poetry experts. However, we have reason to believe that our results are not a complete reversal of ground truth in the way that the results in Section 4.1 were. First, TwitSong's poems are not allusive or obscure in the way that the Good poems from Poetry Magazine were; they are created out of communicative text written by ordinary people and are intended for a general audience. Second, the data cleaning that we performed ensured that we only considered the judgments of people who consistently preferred a human-written found poem to an arbitrary assortment of tweets. Even though these people may have little expertise as such, they at least know enough to recognize and prefer some of the shades of meaning and emotion that appear in human work. Thus, while we would not expect these results to be as robust as results produced by expert judges, we suspect that they are basically correct.

The extremely poor performance of negative sentiment poems surprised us. It does not take much expertise to know of poets, such as Sylvia Plath, who are admired for their eloquence in expressing negative feelings. However, there are many possible explanations for why raters in this task would strongly dislike negative sentiment poems. Describing negative sentiments in an engaging manner may be more difficult than describing positive sentiments engagingly, and TwitSong may not be up to the task. Strong negative emotions may not have been a good fit for the subject matter or the casual tone of the poems. Or the raters on Crowdflower, likely to be ordinary people without much poetic background, may have feelings about negative or depressing poetry which differ from those of literary scholars. Although we did not know it at the time, this result is entirely consistent with the literature about differences between experts and non-experts (Section 2.8.2); non-experts in an art form are more reliant on emotion in their judgments, and are less able to tolerate artwork that expresses or induces negative emotions. We hypothesize, although it remains to be tested, that poetry experts would not dislike the negative sentiment poems as strongly.

5.2 Generation two: Full automation

5.2.1 Introduction

After the success of the first generation of TwitSong, we proceeded to implement line ratings as fully computational rather than crowdsourced human judgments. There are several reasons to pursue this as the next step in TwitSong's development. First, it follows the principle—recall Colton's "meta-mountain" (Colton, 2012)—of steadily giving more creative responsibility to the system itself. A system that judges the best lines for itself, based on a set of programmed rules, has more autonomy than a system that assembles lines based on ratings that were directly assigned by humans. It is still far from the autonomy of a human poet, but it is a small, sensible step in that direction. Second, automated line judgment allows poems to be created much more quickly and cheaply, thereby allowing for a greater volume of poems, greater variety of topics, and more room for informal experimentation on the part of the programmers.

This incarnation of TwitSong was given the name TwitSonnet, because the name TwitSong was already being used on Twitter by a music app.

A prior version of this section appeared at the Eighth International Conference on Computational Creativity (Lamb et al., 2017a).

5.2.2 How TwitSonnet works

TwitSonnet's workings are very similar to TwitSong's, with a few changes made:

1. Collect a large corpus of Twitter data on some topic. Previously, we were piggybacking on another research group's use of the Twitter API. For TwitSonnet, we use the Tweet Archivist service [2] to easily retrieve a keyword-specific corpus of tweets from a specified time interval.
2. Select the lines from the corpus that have 10 or 11 syllables. Tweets with too many syllables can be split up not only into sentences, but based on other punctuation such as semicolons, colons, and commas if necessary.

[2] `https://www.tweetarchivist.com/`

3. Group these lines into rhyme sets. Although tweets are no longer necessarily in iambic pentameter, we still pay attention to syllable stresses when detecting rhymes (e.g. "bed" rhymes with "head", and "painted" rhymes with "acquainted", but "bed" does not rhyme with "painted", because the stresses do not line up even though the sounds are the same).

4. *Rate* the lines in each rhyme set. This is the stage at which we made the biggest changes, described below.

5. Create a poem out of the highest rated lines. Optionally, the selected lines can be re-ranked and placed in a meaningful order. For example, they could be ordered from the most abstract introductory statements (least imagery) to the strongest concluding image (most imagery). Otherwise, the tweets are ordered according to score, with the highest scoring couplet at the end.

TwitSonnet is a fully functional system which can create a sonnet out of any sufficiently large collection of tweets. From July through the end of October 2016, we posted several of TwitSonnet's poems per week at `http://twitsonnet.tumblr.com/`.

5.2.2.1 Choosing poetic criteria

Unlike the first generation of TwitSong, TwitSonnet was developed after we did our experiments in poetry evaluation (Chapter 4) and developed our set of four criteria: Reaction, Meaning, Novelty, and Craft. Therefore, we had the opportunity to conceptually adapt the earlier TwitSong criteria (Topicality, Sentiment, and Imagery) to these criteria.

For the most part, our TwitSong poetic criteria fit fairly well into the existing set. That is to say, they do not express everything about the four-criteria model, but they express things that are definitely a part of that model. Imagery is contained as a subcategory under Craft, and Topicality fits very logically into Meaning. Sentiment and Reaction are not exactly the same, but we can assume for the sake of argument that a person is more likely to have an emotional reaction to a poem that adequately expresses an emotion.

Other subcategories of Craft are also addressed by parts of TwitSong's process. Since tweets are selected to fit into a sonnet form and to rhyme with each other, these are at least two other matters covered under Craft. Novelty is a more difficult category to address, although we could imagine, for example, asking humans to select the tweets that are most different from others.

It is important to note that this rhetorical analysis of a match between TwitSong's design and our four-criteria model *does not constitute an evaluation of TwitSong.* To evaluate the claims we have just made about TwitSong, we would need to empirically test whether or not its line selection strategies actually improve its performance on the appropriate parts of the model. We mention it here, not as evidence that TwitSong is creative, but as evidence that we thought about our four-criterion model when designing the second generation implementation of TwitSonnet and re-evaluating its goals.

Since there is a reasonable correspondence between our existing TwitSong criteria and the four-criterion model, we decided to keep these criteria when implementing TwitSonnet. One advantage of this approach is that, by using the same criteria from the first-generation experiment, we know that these are criteria that will have a positive effect on the poem if implemented properly: readers prefer poems that had higher (user-identified) topicality, imagery, and positive sentiment.

5.2.2.2 Operationalizing criteria

For Sentiment, we changed from measuring sentiment on a positive/negative axis to expressing specific emotions from the set of 8 found in the NRC Hashtag Emotion Lexicon, which was created specifically for Twitter (Mohammad and Kiritchenko, 2015). The eight emotions in the NRC Hashtag Emotion Lexicon are anticipation, anger, disgust, fear, joy, sadness, surprise, and trust. We chose a desired emotion for each poem by measuring which emotions were most prevalent in the gathered data, and then normalizing by the rate of emotion words in non-topical data (groups of tweets harvested from Tweet Archivist based on a neutral keyword such as "and").

For Topicality, in addition to choosing tweets based on time range and keyword as we did in the first generation, TwitSonnet creates a trigram frequency measure for each tweet corpus (trigram frequency being a common way of identifying the topic of a document) and gives higher scores to tweets consisting of trigrams with high frequency scores.

For Imagery, we give a higher score to tweets containing stronger primary process imagery, as measured using the Regressive Imagery Dictionary (Provalis, 1990). The Regressive Imagery Dictionary gives higher scores to "primary process" words relating to physical senses, experiences, drives, and the body, and lower scores to more abstract, "secondary process" words. The Regressive Imagery Dictionary is used directly by Simonton in his study of Shakespearean sonnets (1990); Kao and Jurafky (2012) use a related measure.

For the purposes of this study, we did not find a satisfactory method of measuring novelty. Some obvious attempts, such as selecting for unusual trigrams, seemed to only increase the number of off-topic, "random", and nonsensical tweets. In context of our study of poetry experts, the category of Novelty refers to interesting juxtapositions, new thoughts, and subversions of existing concepts, not to this type of "mere novelty". We did reduce repetitiveness by placing a limit on the number of times TwitSonnet was allowed to repeat a poem's keywords, replacing repetitive tweets with the highest ranked alternatives that did not contain the topic keywords.

5.2.2.3 Using the criteria

After rating tweets according to their Topicality, the chosen Emotion, and Imagery, TwitSonnet then combines the seven highest rated pairs of tweets into a sonnet. If seven valid pairs are not found in the data set, TwitSonnet can instead construct a couplet (one pair), a quatrain (two pairs), or a pair of quatrains (four pairs). A sample sonnet is given in Figure 5.2.

In summary, our system is explicitly built to satisfy the domain-specific product-based criteria derived from our CAT experiment (Chapter 4). However, like any system, its success at satisfying them in practice needs to be tested empirically. We will now describe how we have evaluated TwitSonnet.

5.2.3 Evaluating TwitSonnet

We had two goals in evaluating the current version TwitSonnet. First, we wanted to confirm that the effect of the automated scoring was similar to the effect of the crowdsourced scoring. Second, we wanted to improve on the methodology of the previous study by including expert raters, who are more consistent when rating creative artifacts than non-experts (Kaufman et al., 2008).

Review coming tomorrow/this afternoon.
doctor strange was amazing. cant wait for Thor.
closer look at the evolved hero costume
Visually stunning, left me wanting more
What is a Doctor Strange collector corps box?
Check out the latest new movie details!
So excited to see Marvel in the parks!
what was your first Doctor Strange comic? #StrangeTales
I have 10 more tickets to give away
Doctor Strange 8:45 Ill be there
Doctor Strange is pretty, and pretty OK:
gonna lowkey fall asleep in this chair
It better be worth slacking on my dreams!
Doctor Strange (with Christy at Platinum Screens

Figure 5.2: A sample of TwitSonnet's output, regarding the movie "Doctor Strange". (The keyphrase used was "Doctor Strange", and the time range used was the movie's opening weekend.)

5.2.3.1 Method

Domain experts can be difficult to recruit for studies. We recruited expert participants using snowball sampling on the social networks of this thesis' author, who is a published poet under a pen name, and both supervisors. Each participant was given CAD $10, or the equivalent in their local currency, as remuneration.

Participants were asked demographic questions and classified as experts or non-experts. In keeping with the recommendations of Kaufman et al (2008), we based our definition of expertise not in the study of poetry but in experience actively writing successful poetry. Participants who had published poetry in a magazine or collection, read their own poetry at a reading or slam, and/or published digital poetry were considered poetry experts.

The poetry experts consisted of 13 women, 12 men, and 11 non-binary-gendered poets. (While this is a serious overrepresentation of non-binary poets—likely an artifact of the snowball sampling method—we do not expect it to skew our results, as none of the poems in the study pertain to gender or queer/trans* issues.) The median age was 32, ranging from 17 to 56. All but two of the experts were native speakers of English.

The non-experts consisted of 12 women, 19 men, three non-binary, and one non-expert who did not disclose their gender. The median age was 36, ranging from to 21 to 70. 29 of the 35 non-experts were native speakers of English.

As a result of our snowball sampling, most of our "non-expert" participants could actually be considered quasi-experts: they reported that they were regular readers of poetry, had written unpublished poetry for pleasure, taken classes in poetry, listened to poetry podcasts, attended poetry readings, or taught poetry to K-12 students. (An additional form of experience, being a poetry editor for a magazine or other

publication, did not appear among non-experts. Seven of our 36 expert participants reported having been a poetry editor.) Only three participants had no significant experience with poetry, and one of these was a graduate of a prose creative writing program. Thus, we would expect less difference between the experts and non-experts in this study then we would see if the non-experts were completely inexperienced.

Each participant was shown 8 poems in a random order, from the same selection of 8 current events topics and 8 emotions. The topics included three topics from recent movies and television, two astronomy topics, a ban on the "burkini" in France, and two topics relevant to the recent 2016 Summer Olympics. Each topic was associated with an emotion from the NRC Hashtag Emotion Lexicon: anger, anticipation, disgust, fear, joy, sadness, surprise, or trust.

Each of these 8 poems was in turn drawn at random from one of three groups. In Group A, poems were generated using steps 1 and 2 from the TwitSonnet process, but not the remaining steps. In other words, these were our control poems, in which no filtering or reordering based on our four criteria was performed. Poems in Group B were generated using steps 1 through 4 (so they were generated and filtered using our four criteria, but not reordered), and poems in Group C used all five steps including reordering. For each of the 8 poems, participants were then asked the following questions, each on a 5-point scale:

1. "How much do you like this poem?" *(Reaction)*
2. "How creative is this poem?"
3. "How well does this poem express the emotion of [emotion]?" *(Reaction)*
4. "How meaningfully does this poem summarize its topic?" *(Meaning)*
5. "How new and different is this poem?" *(Novelty)*
6. "How successful is the imagery in this poem?" *(Craft)*
7. "How cohesive is the narrative of this poem?" *(Meaning)*

The answers provided at each point of the scale were

- Not at all
- Not much
- A little
- Somewhat
- Very much

Apart from "How creative is this poem?"—which we felt was an irresistible option in a computational creativity project—each of the questions is designed specifically to assess TwitSonnet's success at one of our four domain-specific categories. Our hypothesis was that the poems from Groups B and C would score higher than Group A on at least some questions, and that Group C would score higher than Group B specifically for narrative cohesion. Participants were also given a freeform text box in which to write any other comments they had about the poems.

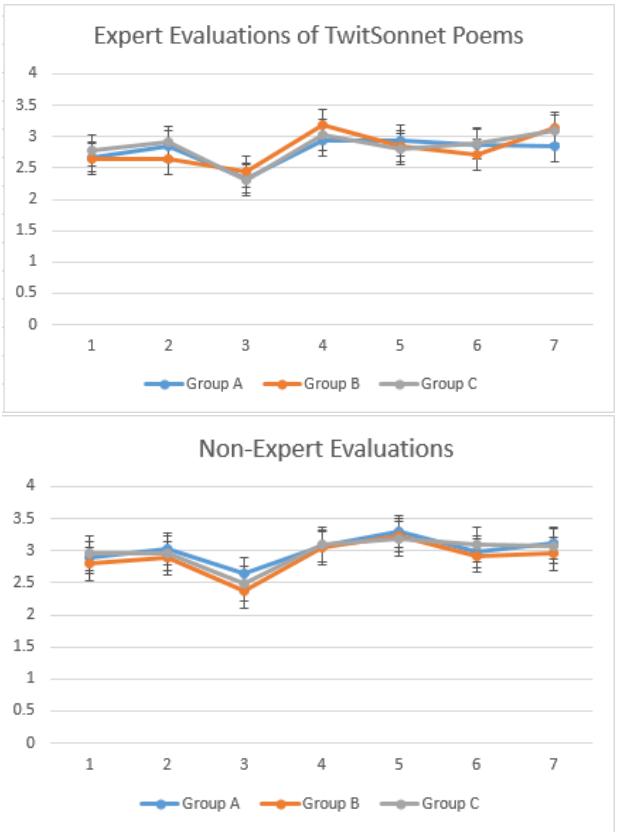

Figure 5.3: Experts' and non-experts' evaluations of TwitSonnet's poems. The X-axis shows the seven evaluation questions in the same order as they are listed in our Method section. The Y-axis shows answers on a 5-point scale, with 5 being the most positive response and 1 the least positive. Error bars show 95% confidence intervals.

5.2.3.2 Results

A one way analysis of variance was performed for each of the seven questions using the survey data from experts, with a Bonferroni correction for seven hypotheses. This analysis showed that the effect of the group from which the poem was taken was not significant for any question, $F(2, 271) = 0.26, 1.54, 0.22, 0.93, 0.36, 0.61$, and 1.41 respectively, $p > 0.5$ in all cases. The results for non-experts were similar.

We compiled the most common freeform comments by experts and nonexperts. Experts stated that the poems seemed random and choppy; often there would be small sections with a satisfying juxtaposition but they would be mixed with other lines that didn't fit. There was too much focus on rhyme and meter at the expense of content, with several experts stating they would have preferred if the poems did not rhyme. There were also too many lines that trailed off in the middle of a sentence or even a word. However, several experts said that they found the idea behind the project very interesting in spite of any criticism they might have of the poems. Nonexperts had fewer comments and responded more to surface features of the poem: for example, several nonexperts said they would have preferred not to see hashtags in the poems, as well as typos, bad punctuation, and other errors. Nonexperts also agreed with experts that the poems lacked coherence.

5.2.4 Discussion

The negative result here is surprising because, in our previous study, the difference between the equivalents of Group A and Group B was statistically significant (Lamb et al., 2015b). There are three possible explanations for this.

First, perhaps the difference is due to a difference in how we performed the evaluation this time (for example, numerical scales vs pairwise preferences). The current study shows a striking lack of difference between groups, which is not attributable merely to the use of a less sensitive statistical method. However, the studies for generations one and three use a pairwise forced choice method and do show statistically significant results, even though freeform comments for generation three (see Section 5.3 below) describe a similar lack of subjective distinguishability between groups. It is possible that, if participants had been forced to express a preference between poems, preferences would have emerged which are too miniscule to appear on a 5-point scale.

A related suggestion is that perhaps the current events topics chosen in this study were not the correct choices. For instance, raters might have had stronger opinions about the emotions expressed in a poem if the poem's topic was more "important" or more polarizing. Such polarizing topics are plentiful in current events, especially as the study was run during the lead-up to the divisive 2016 U.S. presidential election. TwitSonnet's online incarnation did create poems on divisive political topics: an example is shown in Figure 5.4. We chose not to include these poems in the study so as not to conflate a rater's political opinion with their artistic opinion of this poem. This may or may not have been the correct choice.

Second, the main change between the TwitSonnet generation and the first TwitSong generation is that the TwitSonnet automates its line judgments. We have deliberately used computationally simple methods in order to process large numbers of tweets on a large number of topics. It is possible that the key flaw in TwitSonnet's system lies with these line judgment methods and that more sophisticated methods involving machine learning, syntactic parsing, or a knowledge base might get better results. Additionally, although

Final Presidential Debate (10/19)
Donald Trump is master of the head fake
This East Texas pole shows a leftward lean
but goodnight this all debate gave me headache
Much smarter than his brother Crooked John
An interesting debate is taking place
While America tuned in to watch Don
Trump doing the deniro mobster face
started by her very sleazy campaign
Debate Watch Party SAC 305
Donald Trump is LITERALLY insane
watching guy fieris diners drive-ins and dives
That was the sound of women everywhere
Its a humanitarian nightmare

Figure 5.4: A TwitSonnet poem posted online, using the keyword "debate", immediately after the 2016 U.S. presidential election debates.

we test based on the four criteria of Reaction, Meaning, Novelty, and Craft, our actual line ratings are based only on a subset of these criteria. There is more to Craft than rhyme, meter, and imagery, for instance; and just because a line expresses an emotion does not necessarily imply that it will induce that emotion in a reader.

Third, while the focus on this study was on the ranking and ordering steps, the initial filtering step has also improved since the previous study. Humans are unlikely to judge nonsensical tweets as being very topical or as having a clear emotion. Automated judgment is less sensitive to nonsense, and in addition, our filtering step has improved at automatically removing nonsense from both ranked and unranked poems. Thus, it is possible that some of the effect in the previous study was due the ranking step reducing nonsensical tweets, and that this reduction is no longer noticeable in the current study, because the tweets are pre-chosen to avoid this.

Filtering for rhyme and meter (craft) and the use of keywords in data selection (topicality) was already in place in very close to their current form in the first generation, so these steps alone cannot account for our current results, but it should be noted that due to these techniques, even the poems in Group A are not "raw" control poems in the sense of having no attention paid to the four criteria. Neither would, for example, Pentametron's poetry, since it too is selected for rhyme and meter (Bhatnagar, 2012). The use of a *pure* control group - for example, a completely random selection of English-language tweets - would likely produce something closer to a significant result. However, it would not tell us if our filtering techniques, specifically, were working as intended.

5.2.4.1 TwitSonnet as a negative result

As discussed in Chapters 2 and 3.2, falsifiable hypotheses are desirable in evaluating computational creativity, but are inconsistently used. Unfortunately, the testing of falsifiable hypotheses will sometimes

produce a negative result! A negative result does not necessarily invalidate the worth of the project, but it is a sign that the creative system in its current form is not performing as intended.

There are several possible responses to this specific negative result. First, we could try performing a different evaluation. Second, we could modify our line selection techniques and engage in further analysis of existing poems to see which techniques might be most promising.

Third, we could step back and ask ourselves what goals we are working towards with TwitSonnet. A different methodology might serve those goals better. For example, if our goal is to teach a computer to identify poetic lines, we might consider using source text richer in poetic style and technique than Twitter. Even with human raters, generation one of TwitSong had difficulty generating poems with consistent imagery, due to a lack of concrete imagery in the source text. If our goal is to entertain with amusing poetic summaries of news events, we might ask if the present project is the best way to do that. In particular, it is notable that in both this and the previous study, Twitter's informality and conventions such as hashtags were offputting to many participants. These may be aspects of Twitter which make it inherently more difficult as a repository for poetic speech. To verify this interpretation, one option would be to "clean" gathered tweets of hashtags, typos, and other traits that bothered the non-expert raters, before running the study again.

In all cases, a negative result like this one points to a need to reassess and change some aspects of our project, to a greater or lesser degree, so that it fits more precisely with our actual research goals. A negative result in computational creativity also underscores the need for all researchers in this field to use falsifiable hypotheses. If we had not tried to disprove our hypothesis, we might not have discovered that the second generation of TwitSonnet was failing to meet its design goals. We might have had a subjective sense that the poems were not yet as good as they could be—as, indeed, we did before running the study—but we would not have known if this subjective sense was correct. In particular, we would not have known the difference between a situation in which the line selection mechanisms were working as intended but were not conceptually sufficient to constructing a "good" poem, and a situation in which the line selection mechanisms were not improving the poems significantly at all.

5.3 Generation three: The editorial algorithm

5.3.1 Introduction

After the negative result of the second generation of TwitSonnet, we were left pondering what to do for the third generation. As discussed above, it seemed that sonnets made from Twitter might be inherently unsatisfying to readers, and our current work had not produced the hoped-for effect.

Furthermore, another doubt had arisen because by this point in our research we were finished our survey of evaluation techniques and our taxonomy of poetry generation specifically. The first two generations of TwitSong are not as sophisticated as many existing poetry systems, and there is a strong argument to be made that their Process is not as creative as other systems. Instead of creating its own lines, TwitSonnet selects lines according to metrics it is given. In order to make the third generation of TwitSong more creative, we were strongly interested in teaching it to edit its own work. Specifically, we wanted to design a system of targeted edits akin to that of the COLIBRI (2002) and WASP (Gervás, 2013a,b, 2016) systems. We admired WASP's ability to reason about the edits it was making in a more humanlike way, but we

noted that WASP cannot yet specifically edit its poetry to be more topical. Since TwitSong already incorporates optimization for emotion and topicality, a TwitSong which made targeted edits to it work for these properties would be an undeniable advancement in the state of the art.

5.3.2 The mechanisms of generation three

Generation three of TwitSong is built atop the same code for line representation, RhymeSet construction, line judging, and poem construction as the first two generations. Some tweaks were made to each of these, but there is largely continuity. The main change between TwitSonnet and generation three is that generation three uses a mechanism called the Editorial Algorithm, described below, to iteratively change and improve its top-rated lines. Generation three also draws from different source texts, largely news articles, to construct its lines.

We changed the poetic form with which the third generation of TwitSong works, out of a feeling that perhaps sonnets—thought of by the public as a serious, refined form of poetry—were not the best match for poetry that was meant to be topical and light-hearted. Instead, TwitSong 3.0 produces quatrains in Common Meter—an ABAB rhyme scheme with four iambs (eight syllables) in the A lines and three iambs (six syllables) in the B lines. This is the form of many hymns, including "Amazing Grace," as well as other popular poems and songs.

We made two changes to the line rating algorithms in TwitSong 3.0. First, we were frustrated by the muddled meter of the sonnets, but reinstating a strict requirement for iambic pentameter meant discarding too many lines. So we added a fourth poetic criterion, the criterion of Meter. A line is given a score between 0 and 1 based on how closely each of its syllables adheres to the rules of iambic pentameter. Because the stresses of single-syllable words can be difficult to discern, and because the CMU pronouncing dictionary also includes secondary stresses, our Meter scores are not exact. Since human metrical poetry also contains occasional modifications to the meter, we choose to err on the side of permissiveness. However, in practice, selecting lines based on our measure of Meter tends to produce well-formed lines with a strongly iambic feel. This Meter score is then combined with the other normalized scores to produce a line's total score.

Second, we found that the trigram measure of Topicality tended to select for bland turns of phrase which were common in the source text, but did not make it clear what the topic actually was. We modified the trigram measure so that, in addition to rewarding more common trigrams, it also rewards lines that contain the most topical words. These topical words are selected by dividing the frequency of each word in the source text by its frequency in a non-topical comparison text (in this case, the comparison text is a compilation of poems from Poetry Magazine; a factor is added to the frequency to prevent division by zero). The thirty most topical words according to this measure are then chosen and a trigram containing one or more of these topical words receives a bonus to its topicality score, which is larger for the most topical words (i.e. the top ten or twenty) and for trigrams containing multiple topical words.

5.3.2.1 The Editorial Algorithm

Our Editorial Algorithm which we use for TwitSong is a form of genetic algorithm. However, instead of randomly recombining the most successful candidates in each generation—a technique which bears little resemblance to how human poets revise their work—we use a targeted edit at each step, replacing the words in each line that contribute most to the line's worst-performing metric, out of the four metrics of

Topicality, Emotion, Imagery, and Meter. So, rather than adhering strictly to the biological metaphor of a genetic algorithm, we instead use the idea of a genetic algorithm as a base for a more artistically specific form of generation-evaluation loop. We detail this algorithm below.

1. Initialization. As with prior versions of TwitSong, the source text is read and the dictionaries used for each criterion are initialized, including the trigram frequency dictionary and identification of most topical words which are used for the Topicality score. We also initialize an interpolated Markov model (Salzberg et al., 1999) which can be used to generate arbitrary amounts of additional text in the style of the source text. The Markov model can be up to order 3, but can flexibly reduce its order if necessary. If the model generates no results, or only one result, for a 3-gram, then it reduces the 3-gram to a 2-gram or 1-gram. The Markov model is trained to recognize punctuation that could indicate the end of a sentence or line, and runs until it generates an "end of line" marker; in an earlier version that did not use these markers, it was too common for a line to end on a preposition or other unsuitable word. Because the "end of line" marker is not guaranteed to occur after exactly 6 or 8 syllables, the Markov model in practice is run repeatedly until it generates a line that happens to be of the right length. If 100 iterations fail to generate a line of the right length, no line is generated.

2. Line initialization. The source text is read again and this time divided into lines of 6 or 8 syllables, based on appropriate punctuation such as periods, question marks, colons, and commas. Each of these lines is sorted into an appropriate RhymeSet. A group of special lines is also generated purely based on the Markov model's output when given the 30 most topical words and told to iterate on them until it reaches 6 or 8 syllables. Each of these lines is also sorted into a RhymeSet and is then subject to the same processes as all other lines.

3. Scoring. The lines in each RhymeSet are scored based on the four combined metrics, and each RhymeSet is scored based on its best two lines. RhymeSets with a single line are scored, but penalized.

4. Trimming. Each RhymeSet is trimmed in order to increase efficiency and reduce repetition of identical or nearly-identical lines. Any line that is identical to the RhymeSet's top scoring line, or that begins or ends with an identical word, is removed. Optionally, the programmer can also specify removal words that can only appear once in each RhymeSet; if the top scoring line contains one of these words, then any other line containing that word is removed. This is useful for preventing repetitive language. If more than 15 lines remain in the RhymeSet, it is then trimmed down to only its 15 highest scoring lines. The RhymeSets are then re-scored and the 50 highest scoring RhymeSets are kept for the next generation, with RhymeSets of only a single line being removed first.

5. Edit planning. This is the stage at which the Editorial Algorithm identifies which words have the greatest need to be replaced. Each line in the current group of RhymeSets is analyzed based on the four criteria. The criterion with the lowest normalized score, as well as any other criterion which is under a certain threshold, is selected for analysis. Each word in the line is then inspected for its contribution to this criterion, and the lowest performing word is selected for replacement. ("Stop words," such as "the" and "of," are not excluded from this process; the thinking is that, if a stop word is present, there is no *a priori* reason why an alternate version of the line might not use a different sentence structure and have a higher-scoring word there instead.) For example, if Imagery is selected, then words that are very abstract (or "secondary", according to the Regressive Imagery Dictionary) are selected.

6. Word replacement. The selected lines are sent to the Markov model which generates candidate replacement lines, starting with the selected underperforming word and replacing it and all subsequent words. (An earlier prototype of TwitSong replaced only the underperforming word, but this led to choppy and repetitive lines which destroyed the surface continuity that would otherwise be provided by a Markov model; an example is given in Table 5.3.2.1.) Because there is no guarantee that the replacement words will actually be better, the Markov chain generates many candidate replacement lines—20 for each selected starting word. These are then assigned to appropriate RhymeSets.

7. Successive generations. TwitSong repeats steps 3 through 6 to a maximum of 100 generations, or until the average score of the best ten lines stops increasing, whichever comes first. In practice, the program very rarely runs for more than 15-20 generations, and sometimes as few as 3.

8. Poem construction. After there are no more word replacements to be made, the top two lines each from the two highest scoring RhymeSets are selected. These are arranged into a quatrain in Common Meter.

9. Title generation. TwitSong generates its own title for each of its poems, but the title generation mechanism is separate from the rest of the Editorial Algorithm. During the Line Initialization step, in addition to creating the initial RhymeSets out of lines 6 or 8 syllables long, TwitSong also gathers a set of lines from the source text of 3 to 5 syllables without sorting these lines by length or grouping them into RhymeSets. These potential title lines are then scored based on the four combined metrics and checked against the list of most topical words. Ideally, lines containing the first most topical word are selected and the highest scoring such line becomes the title. If there are no such lines, TwitSong will iterate down the list of most topical words. If no potential title line contains any of the 30 most topical words, TwitSong will choose the highest scoring potential title line to be the title.

let wall street start off wall detroit's
and wall street start wall voiced
let wall street start wall street wall point
wall street out loud wall point

Figure 5.5: An early example of a poem from a prototype Editorial Algorithm, using Bernie Sanders' lines from presidential debate transcripts as a source text. In this prototype, pairs of words were replaced during each edit. (An even earlier version, replacing single words, resulted in lines like "let wall wall wall wall wall street".) This problem was avoided by a later protocol in which the target word and everything after it in the line is re-generated at once.

5.3.2.2 Source Texts

TwitSong 3.0's architecture allows it to generate poems more quickly and based on shorter source texts than any previous incarnation. In particular, the use of a Markov chain to generate both special seed lines in the first generation and replacement lines thereafter, means that a relatively small source text can be

used to generate poems. TwitSong 3.0's lower limit seems to be around 20 kilobytes of text, or sometimes a bit less, depending on the properties of the text. Therefore, it can be initialized with only a handful of news articles on a breaking news topic, instead of needing to wait days or weeks for tens of thousands of tweets containing a keyword to appear.

We generated a great number of poems using TwitSong 3.0, mostly based on news articles from the BBC [3], CBC [4], Maclean's [5], and The Guardian [6]. We chose these sources as our mainstays because they are mainstream, professional English language news sources which operate without a paywall. Occasionally we veered into other sources. For instance, when blockbuster movies were released, we collected fan responses to the movies from Tor.com [7] and The Mary Sue [8]. For major holidays, the Wikipedia article describing the holiday was used. We also used alternative, non-news sources for some poems, such as classic novels available on Project Gutenberg[9].

5.3.2.3 The Evolutionary Algorithm in action

As an illustration, we show how the Evolutionary Algorithm uses its word replacement techniques on lines for a poem about the film *Avengers: Infinity War*.

One of the starting lines for this poem is:

> thanos to grow the universe

This line receives high scores for prosody and imagery, but a low score for topicality and a moderately low score for the chosen emotion, surprise. As both topicality and emotion are below their minimum thresholds, the Editorial Algorithm focuses in on both of these for word replacement.

Since topicality is calculated based on trigrams, TwitSong splits this line into its component trigrams:

> thanos to grow / to grow the / grow the universe

The first and last trigrams are selected because they are not found in the trigram dictionary (that is, in the source text; "to grow the" appears in the source text twice). Thus, TwitSong generates a set of candidate replacement lines starting at the beginning of the line, and a set of candidate replacement lines modifying only the last three words.

For emotion, TwitSong splits the line into its component words:

> thanos / to / grow / the / universe

[3] http://bbc.com/news
[4] http://www.cbc.ca/news
[5] http://www.macleans.ca
[6] https://www.theguardian.com/international
[7] https://www.tor.com/
[8] https://www.themarysue.com/
[9] http://www.gutenberg.org/

None of these individual words are very associated with the emotion of surprise in the NRC Hashtag Emotion Lexicon, and some do not appear in the lexicon. Therefore, TwitSong flags all of them, and generates a maximal set of candidate replacement lines (a different set beginning the word replacement at each word).

The completed poem from this run of TwitSong reads:

marvel had the fall of your mouth
luke of this journey through
infinity stone to point out
gags to where thor is too

5.3.3 Evaluation

As in the previous two generations, we ran a study to evaluate TwitSong. Specifically, our goal was to falsifiably test whether or not the Editorial Algorithm and its associated line rating techniques improved TwitSong's poetry.

Emotion	**Frequency**	**Topics**
Disgust	7	Mueller investigation; the Parkland school shooting; Rex Tillerson; Doug Ford's election campaign in Ontario; March For Our Lives; the Stormy Daniels scandal; Viktor Orban's election in Hungary
Fear	6	Winter Olympics (2); Uber self-driving car crash; the Russian election; Austin bombing; NAFTA negotiations; Syrian chemical attack
Anticipation	5	Kim Jong Un's visit to China; Russian spy poisoning; US trade war; North Korea; Michael Cohen warrant
Anger	4	The Cambridge Analytica scandal; Facebook; Tim Hortons; Mark Zuckerberg
Joy	3	Winter Olympics (1); A Wrinkle in Time; Easter on April 1
Sadness	3	Stephen Hawking's death; Good Friday; Humboldt Broncos bus crash
Surprise	1	The Oscars
Trust	1	Black Panther

Table 5.3: Frequency of emotions from the NRC Hashtag Emotion Lexicon assigned to poems on different topics, from the group of 30 topics that were selected for the study. The topics in this table are sorted by associated emotion for ease of reading, and their order does not correspond to the ordering of topics in the study.

5.3.3.1 Method

We assembled three experimental groups of poems: Group A, Group B, and Group C.

Poems from Group A were generated according to the Editorial Algorithm described above. Lines were taken from a source text and generated based on a Markov chain trained on the source text, and the best

Group A
FOR CANADA *(Olympics, joy)* hamelin pointing at the world team made it would be fair swiss stones for pavel is absurd swiss stones for him and there
Group B
WHY IS TRUMP SILENT *(Mueller investigation, disgust)* republican claims he will do flynn pleaded not care less committee has to look into pleaded not to the press
Group C
WAKANDA *(Black Panther, trust)* blackness as we love to her aid killmonger's plan to come conflict the atlantic slave trade sword and it was awesome

Table 5.4: Example poems from the three experimental groups.

lines of each generation were edited with the goal of increasing their score on our four criteria of Topicality, Emotion, Imagery, and Meter.

Poems from Group B were generated with a minimal version of the Editorial Algorithm. Lines were taken from a source text and generated based on a Markov chain trained on the source text. If this resulted in enough RhymeSets to produce a quatrain in Common Meter, the program was stopped there. Otherwise, it was allowed to iterate and perform the Editorial Algorithm for *only* enough generations to produce a valid quatrain. In either case, every line was then assigned a score of zero, and the lines for the quatrain were chosen arbitrarily, as in the control group for the first study (Section 5.1.2.3). Group B was meant as a control group in which the Editorial Algorithm did as little to improve the poems as possible, yet in which the poems were as similar as possible to the poems of Group A in every other respect.

Poems from Group C were generated with a *reversed* Editorial Algorithm. That is to say, the algorithm was run as in Group A, but both the line rating and the edit planning steps were programmed to minimize instead of maximizing the poem's scores. So poems from Group C were the Editorial Algorithm's attempt to make poems that were off-topic, unrelated to the selected emotion, abstract / devoid of imagery, and that failed to conform to an iambic stress pattern.

We chose a set of 30 news topics that were current at the time of the study and generated a Group A, Group B, and Group C poem for each. We then constructed a test set for our study in which, for each of the 30 topics, two of the groups were selected. The order of the news topics was not randomized, but the order of pairings (A vs B, A vs C, B vs A, B vs C, C vs A, or C vs B) was randomized across the set of news topics. All eight of the emotions from the NRC Hashtag Emotion Lexicon were present in our set of poems, but we made no attempt to balance or equalize the appearance of different emotions, instead generally picking the emotion that was most prevalent in articles describing each topic according to the NRC Hashtag Emotion Lexicon, with some normalization and some exceptions (see Table 5.3 for a full list of topics/emotions)

We recruited experimental subjects, as in the previous study, by snowball sampling in order to include a reasonable number of poetry experts in our analysis. Each subject was directed to an online survey in which they were asked about their poetry expertise, presented with each of the 30 pairs of poems, and asked their opinions on the poems. Participants were also asked a few demographic questions and given a freeform text box at the end in which to write other comments about the poetry in the study. The full study, if completed, took about 40 minutes and participants were given 10 Canadian dollars, or the equivalent in their local currency, as remuneration.

Mindful of the methodological issues of the study in Section 5.2, we returned to a pairwise forced choice paradigm for generation three, as we had previously done for the first generation. For each pair of poems in the study, participants were asked the following questions:

- Which poem do you prefer? *(General/Reaction)*
- Which poem is more creative? *(General)*
- Which poem does a better job expressing the emotion of [emotion]? *(Reaction)*
- Which poem does a better job describing the topic of [topic]? *(Meaning)*
- Which poem is more new and different? *(Novelty)*
- Which poem has better imagery? *(Craft)*

5.3.3.2 Results

5.3.3.2.1 Demographics We divided our survey participants into experts and non-experts based on their self-reported experience with poetry. For ease of processing, we used a slightly more stringent (and simpler) definition of expertise: experts were defined as participants whose poetry had been published in a magazine, anthology, collection, etc.

32 poetry experts participated in our study. This included 11 men, 10 women, 9 non-binary poets, and two experts who did not disclose their gender. (As in the previous study, this is a serious overrepresentation of non-binary poets, but we do not expect it to affect our study results as none of the poems in the sample discuss queer/trans* issues.) Their ages ranged from 22 to 57, averaging 38. 28 of the 32 experts were native English speakers.

49 non-experts participated in our study, including 17 men, 27 women, and 5 non-binary participants. Their ages ranged from 17 to 64, averaging 32. 37 of the 49 non-experts were native English speakers.

As in the previous study, many of our non-experts could be considered quasi-experts, as they had some experience with poetry despite not meeting the criteria for being considered an expert. Nearly all (45) of the non-experts had read poetry for pleasure. Many had written poetry for pleasure, read digital poetry, or studied poetry at university. A few non-experts had other significant forms of poetry expertise including writing digital poetry, performing at a poetry slam, teaching poetry, being a poetry editor, publishing criticism/reviews of poetry, or creating an art installation involving poetry. Aside from having had their poetry published, experts reported higher experience than non-experts in nearly every category tested, but no other form of experience with poetry was exclusive to experts.

We also observed high attrition with our survey participants as the survey was rather long. Only 18 experts and 28 non-experts managed to complete every question and make it to the end of the survey. However, since the order of appearance of poems from different groups was randomized, this still left us with a good number of pairwise comparisons between each possible pair of groups for each participant, and did not present a major statistical problem.

5.3.3.2.2 Group comparison We evaluated pairwise preferences between poems by treating them as a binomial distribution; statistical significance is calculated using the binomial theorem for cumulative probability. The null hypothesis is that the probability of choosing a poem from one group over a poem from another, on any question, is 50%. As there are six questions, we applied a Bonferroni correction for multiple hypotheses, resulting in an alpha level of .0083 per test.

Our results are shown in Figures 5.6 and 5.7. As we had hoped, experts significantly preferred poems from Group A to poems from Group B on all six questions,$p < 0.0083$ for all. The differences between Groups B and C were not significant; also surprisingly, neither were any differences between groups A and C.

Non-experts, like experts, significantly preferred poems from Group A to poems from Group B, $p < 0.0083$ for all questions. Unlike experts, non-experts significantly preferred Group C to Group B on all questions, $p < 0.0083$. The differences between Groups A and C were not significant for non-experts.

Rather than the hierarchy with $A > B > C$ that we had expected, it seems that there is little difference between A and C. For experts there is some evidence of a possible hierarchy with $A > C > B$, but with the differences other than $A > B$ too slight to be significant. For non-experts, groups A and C seem to be genuinely statistically the same. This is illustrated in Figures 5.6 and 5.7.

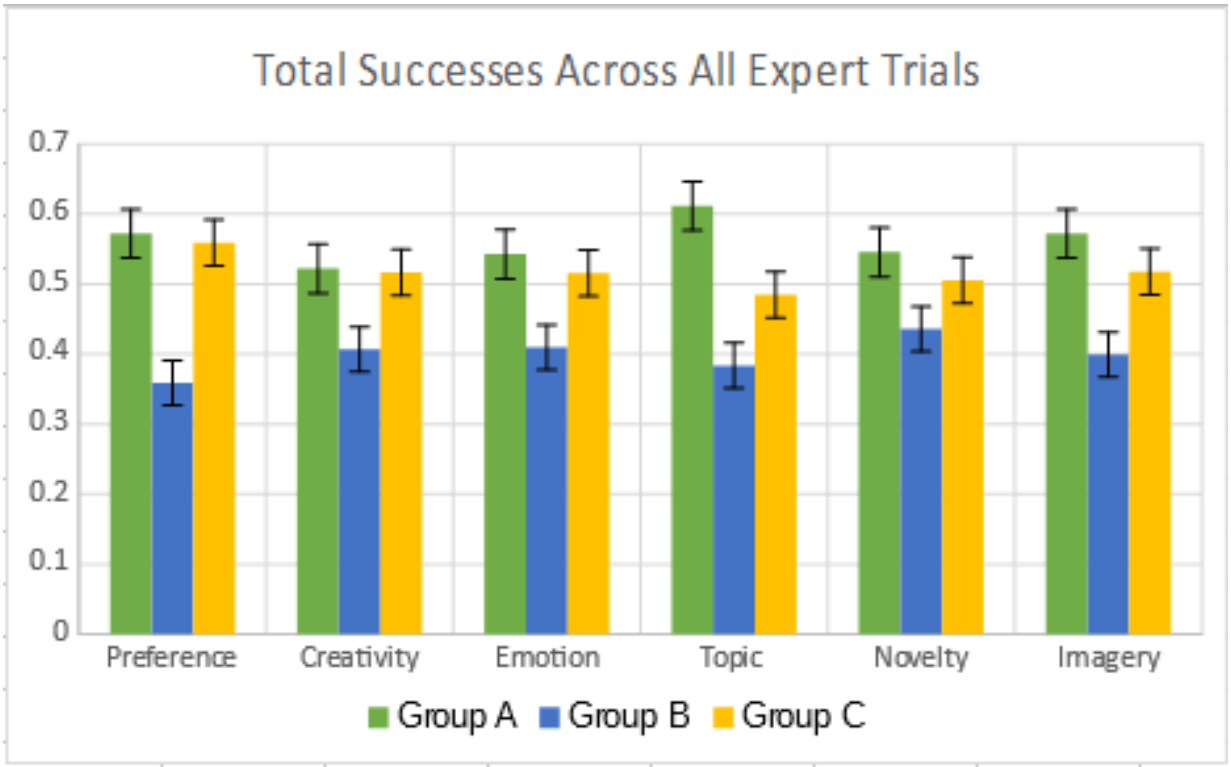

Figure 5.6: Success rates for types of computationally generated poems in pairwise comparisons with other poems, judged by experts. The height of a given bar represents the number of times a poem from that category was selected in preference to any other poem. Error bars represent 95% confidence intervals, prior to Bonferroni correction.

Preference, Creativity, & Novelty	**Emotion**
PLAYERS WOULD NAP *(group C, Humboldt Broncos bus crash, sadness)* players have donated the ones injured but you know how games to come playoff traditions games and panic and now	TRUMP LEAVES HEADS SPINNING *(group C, the Parkland school shooting, disgust)* shooting following his thursday weapons sales to the core weapons like his nra say gun laws have one used for
Topic	**Imagery**
EASTERTIDE *(Easter on April 1, group A, joy)* orthodox easter on sunday liturgy of some sort liturgy of april fools' day easter and are ignored	STEPHEN HAWKING *(Stephen Hawking, Group C, sadness)* stephen was a fun loving guy research to discuss cash if we find evidence of ai rees said to have shed mass

Table 5.5: The poems most strongly preferred by experts on each of our six questions.

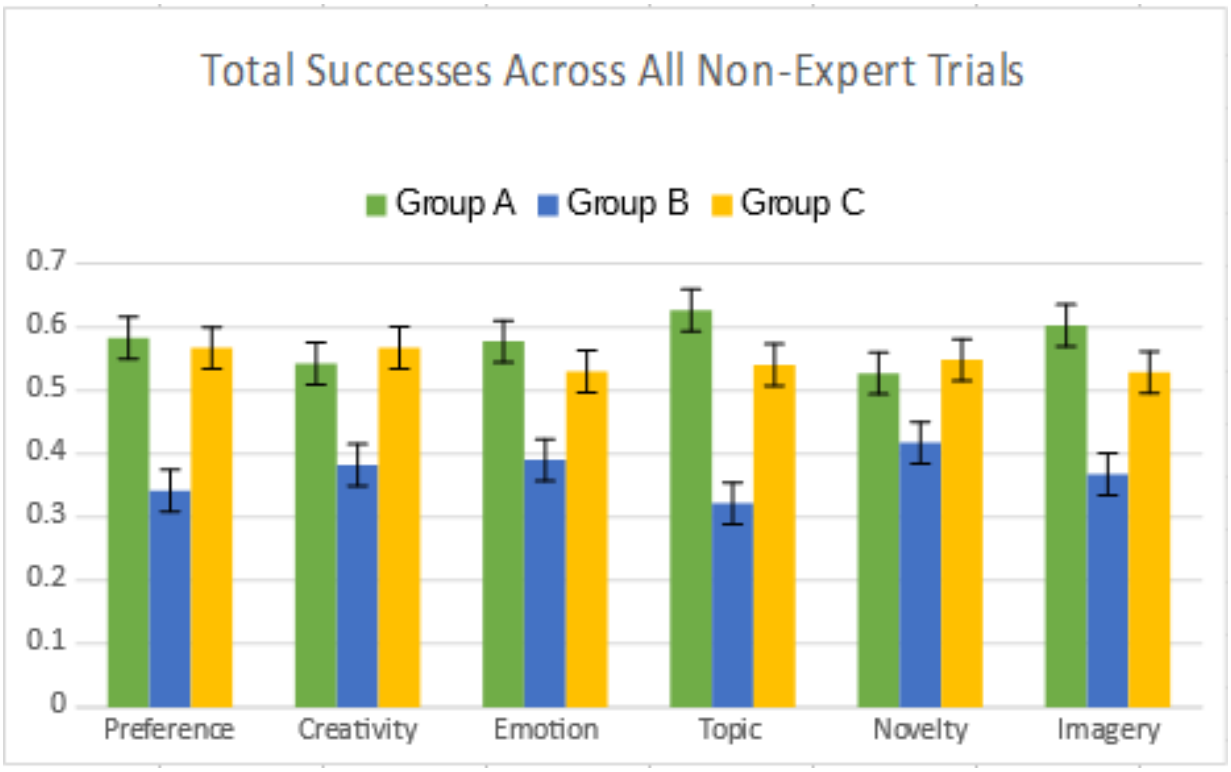

Figure 5.7: Success rates for types of computationally generated poems in pairwise comparisons with other poems, judged by non-experts. The height of a given bar represents the number of times a poem from that category was selected in preference to any other poem. Error bars represent 95% confidence intervals, prior to Bonferroni correction.

5.3.3.2.3 Favorite poems Out of curiosity, we calculated the experts' most highly rated poems in the data set on each of the six questions. Since our study uses pairwise comparisons rather than "objective" quality measure, these are not necessarily participants' favorite poems, but rather, are the poems that most markedly outperformed the poems they were paired with. These are listed in Table 5.5. Interestingly, all but one of these outstanding poems was from Group C. Two of them are Group C poems which were paired with, and markedly outperformed, a poem from Group A.

	Preference	Creativity	Emotion	Topic	Novelty	Imagery
Preference	1					
Creativity	0.855	1				
Emotion	0.563	0.360	1			
Topic	0.525	0.320	0.344	1		
Novelty	0.839	0.861	0.351	0.227	1	
Imagery	0.904	0.837	0.421	0.423	0.892	1

Table 5.6: Correlations (Pearson's R) between answers to each of the six questions, as judged by experts.

5.3.3.2.4 Correlation between questions We also looked at the correlations between the answers to our different questions, to see if our questions were truly capturing different dimensions underlying Product creativity, or if participants were simply choosing the poem they preferred. The results, for experts, are in Table 5.6. All the correlations between questions are above 0, which is not worrisome in and of itself, since it is expected that a preference for a poem in some questions would have a priming effect on preferences in the other questions. However, some correlations are weak to moderate, while others are strong. There is a notable gap between the strongest moderate correlation (preference and emotion, Pearson's r=0.56) and the weakest strong correlation (creativity and imagery, r=0.84).

It appears from these results that the measures of preference, creativity, novelty, and imagery are all strongly intercorrelated with each other, while emotion and topic are more independent. This implies that experts evaluate the poems being studied on three basic dimensions. One is how well the poem represents the target topic; another is how well the poem expresses the target emotion; a third is a more nebulous measure of how "good" the poem is, which includes novelty, imagery, and overall preference. Non-experts exhibited the same pattern as experts, with three underlying dimensions.

5.3.3.2.5 Freeform comments We also counted and categorized the freeform comments made by experts and non-experts. Experts commented more often than non-experts, but there was more unity in the types of comments made by non-experts.

Several experts and several non-experts stated that the poems usually didn't represent the intended emotions very well. Both experts and non-experts also wished that there was a neutral/none/both option in the survey for times when neither poem met its targets particularly well.

Experts, but not non-experts, were concerned about the poems' coherence. Several stated that the poems were incoherent, or that they cared more about the poems' coherence than the items that the survey asked for. Two experts added that some lines were great, but that they were spoiled by their proximity to incogruous or "word salad" lines.

Non-experts were much more likely to make comments about the quality of the poems in general, although they were divided in their responses. Several said that the overall set of poems, or the idea for the study itself, was interesting or cool. Some said that the poems overall are not very good, while others said that some individual poems within the set were quite good. Several non-experts also indicated that the poems were hard to understand or didn't make sense, which may be the non-expert version of complaints about coherence.

One expert commented, "My god, that was awful. The poems were some of the worst computer-generated texts I've ever seen." In contrast, a non-expert said, "This is a really interesting study–I was trying to guess which poems were computer-generated as I did the survey, and I couldn't tell most of the time!" This latter comment is notable since we had intended for it to be clear that *all* poems in the study were computer-generated.

BORDER POST *(group A, NAFTA, fear)* migrants are expected to hold progress is very strong mexico a limited role they harass them along	AND BLACK FRIDAY. *(group A, Good Friday, sadness)* generally closed for details friday in the uk liturgy of jesus with nails easter falls on thursday
MR PUTIN *(group C, Russian election, fear)* violations are putin's tune russians share of which ones kremlin said there is very soon kremlin said relations	

Table 5.7: The three poems that each received one retweet on Twitter during our Press evaluation. (The poem that received one Like is the Stephen Hawking poem which is reproduced in Table 5.5.)

5.3.3.3 Press evaluation

In addition to our survey, we also attempted a Press evaluation to gauge TwitSong's cultural success (see Section 2.6.7, (Jordanous et al., 2015; Jordanous, 2016b; DiPaola et al., 2013; Sheridan et al., 2005; Tresset and Deussen, 2014)). From the two months of mid-March to mid-May, we posted several of TwitSong's poems per week on the Twitter account @uwtwitsong[10]. 54 topics were used, including all 30 of the topics from the Product survey as well as other news topics and some non-news topics such as novels. For each topic, three poems were posted between 9am and 5am Eastern time on the same calendar day—one from Group A, one from Group B, and one from Group C. The exact time and order was randomized. Our intent was to compare the number of Likes and Retweets gathered by each group, and to use this to gauge the relative Press success of each group.

Unfortunately, although we promoted the @uwtwitsong account and gathered over two dozen followers, the poems received almost no Likes or Retweets. Exactly three poems were Retweeted once each, and one received a single Like. (These poems are reproduced in Table 5.7.) This is not enough data on which to

[10] `http://www.twitter.com/uwtwitsong/`

perform any statistical analysis, although two were from Group A and two from Group C. It is, perhaps, an indication that TwitSong is not especially successful from a Press perspective.

5.3.4 Discussion

We were quite surprised by our results. It seems that although the Evolutionary Algorithm does indeed improve poems, it does so even when told to make the poems worse.

We can think of a few ways to interpret this result. One is to conclude that our line rating metrics are useless and that something else about the Evolutionary Algorithm is what improves the poems. However, we are not sure what this would plausibly be. Although our line rating metrics did not seem to improve the poems in the previous study, that study was done with numeric scales, which are less able to pick up small but consistent differences between groups, and this one was done with the more appropriate technique of pairwise comparisons. Although participants complained in both studies that they did not see much difference between the groups, they nevertheless were able to detect, perhaps without realizing it, the difference that there was. This difference must be due to something related to the line rating metrics and how they were used, as there were no other consistent differences between the poems from the three groups. It is conceivable that replacing words in successive evolutionary generations somehow inherently improves the poem even if the words and reasons for replacing them are arbitrary, but we are not sure why this would be the case.

Another possible explanation is that, while the line rating metrics are useful, their reverse versions are also useful. This is most easily explainable with the Meter metric. A line with a perfect score of 1.0 for meter is (as far as the computer can detect) a perfect iambic line. However, the opposite of an iambic line is not simply a line without any meter. Instead, the opposite of an iambic line—with unstressed syllables where the stressed syllables should be, and vice versa—is a trochaic line. But trochaic poetry is as metrically valid as iambic poetry. It is very likely that, while lines from Group B had random stress patterns and lines from Group A were mostly iambic, lines from Group C were probably mostly trochaic, and there are good reasons why a human would prefer both Groups A and C to Group B on this metric. Indeed, looking at the poems from group C, many do contain trochaic or close to trochaic meter, with lines like *game that finish gave a doping* or *shooting following his thursday*.

This explanation is speculative due to a flaw in our experimental design, namely, that we did not explicitly include a question like "Which poem has better meter and rhythm?" to our survey—in spite of the fact that rhythm and meter, every bit as much as Imagery, is a valid subcategory of Craft.

If both Group A and Group C have good meter and Group B does not, then there are two possible explanations for the rest of the results. One is that the answers to the other questions are illusions—that survey participants prefer the poems with better meter, and that the meter tricks them into thinking that all of the other categories are also better. Another possible explanation is that other line rating metrics, other than meter, also exhibit this reverse effect. A poem with low Topicality might contain more unusual trigrams and, thus, more Novelty. A poem with a sufficiently low rating for one emotion might end up exhibiting another, equally interesting emotion. Lines with lower Imagery might use simpler language (stating a situation outright instead of representing it with a sensory image) and therefore be more coherent. This explanation does not completely explain the data; for instance, it does not explain why Groups A and C are both more topical and more novel than Group B.

We suspect that our results are due to a combination of both explanations. Both Group A and Group C contain improvements on the Group B poems to some extent, especially the very visible and visceral improvement of meter, but Group A is slightly more on target with regards to its other specified goals. Experts are more sensitive to this on-targetness, resulting in a ranking where Group A (slightly, non-significantly, but consistently) outperforms Group C, while non-experts are more fully swayed by meter and base their preferences much more on meter than on any other factor we can control. It should be noted that although the differences between Group A and Group C when judged by experts is not large enough to be significant, there is only a 1/64 chance that Group A would outperform Group C on all six questions if the data was truly random.

This explanation is tentative, however. If it is a true explanation then we would expect several predictions to come true in further experiments. First, we would expect that, if we did include a question about Meter (or measured Meter automatically), then Group A and C would prove to have better meter than Group B, and that quite a few if not all of the other questions would be highly correlated with Meter, especially for non-experts. Second, if we were to somehow come up with a better implementation of our line ratings, then the difference between Group A and Group C, at least for experts, would increase.

We note that some of the results from our human poetry study (Section 4.1) were not replicated here. Specifically, non-experts were similar to experts rather than reversing their pattern, and novelty correlated strongly with preference instead of reversing it. We suspect that the similarity between experts and non-experts is due to two factors. First, although the poems are not coherent, they are also not intentionally allusive or obscure, and thus there is no cause for an effect where experts understand the poems and non-experts do not. Second, as mentioned, our non-experts are really mostly quasi-experts—more similar to the XDM students of Section 4.2 than to the completely naive judges of Sections 4.1 and 5.1—and would thus mimic expert opinion more closely than a sample of random people from a service like Crowdflower. As for novelty, we suspect this may be due to the differences between human and computer-generated poetry. Anecdotally, we have found that the poetry of unskilled humans is easy to understand, but mawkish and stereotyped. This sense of obviousness, in our human poetry, was what led non-experts to prefer unskilled, non-novel poetry. However, unskilled computer poetry is usually not particularly obvious; instead, unskilled computer poetry tends towards incoherence. If all or almost all of the poetry in a data set is incoherent, then any discernible thought emerging from the chaos will be both novel and interesting in comparison, in ways that do not make the poetry less accessible to non-experts.

5.3.5 Conclusion

We have now seen three separate incarnations of TwitSong, each of which works on the same principle—find lines in existing text, rate them according to desired qualities, pick the best ones, and arrange them—but each of which has different capabilities and results. We have progressed from a proof of concept that outsources the line ratings to humans, to a more fully automated system generating Twitter sonnets, to a quatrain system which can edit its own work. It is time to reflect on how far we have come and what we have learned from constructing these three systems.

From a **Product** perspective, rigorous tests of TwitSong have yielded mixed results. It is clear, from generation one, that the line selection algorithm works in principle. However, our automated versions of line selection are not especially good. We will discuss this much more, along with possible lines of research for better forms of line selection, in Chapter 6.

The other major problem with TwitSong from a Product perspective is that its poetry is not very coherent. Coherence means more than simply sticking to a topic; coherence means being able to carry a thought from the beginning to the end of a poem without losing track of it. Coherence has to do with syntax and sentence structure as well as staying semantically consistent and avoiding non-sequiturs. In Chapter 6 we will discuss how coherence is a challenge for many other otherwise successful computational poetry systems, not only TwitSong. Operationally defining and achieving coherence in computational poetry remains a topic for future research.

From a **Process** perspective, TwitSong has not been formally evaluated. As described in Section 2.4.5, a formal Process evaluation would require bringing in an outside expert to thoroughly investigate TwitSong's workings, followed by placing it in a category based on its level of Process creativity, or qualitatively describing its strength and weaknesses based on a model like SPECS, or both. We have not performed this evaluation. As TwitSong was initially conceived as a project closer to the engineering-mathematical side of Peréz y Peréz's (2018) continuum, we have always been more interested in Product and Press evaluation for this project than Process.

However, we can argue informally that in some sense, TwitSong has acquired more Process creativity in each incarnation. Using automated instead of human line selection gives more autonomy to generation two than generation one had, and allowing generation three to revise its work gives it yet more. TwitSong is, of course, quite far from being fully autonomous, as the metrics it uses, their weights, the source text chosen, the emotions to target, and all other aspects of its algorithm are still specified by the programmer. However, by revising its own work to improve semantic properties such as topicality and emotion, TwitSong's third generation surpasses WASP (Gervás, 2016), which was previously the state of the art in Process in terms of optimization.

In our Process-based taxonomy of generative poetry, TwitSong occupies the area of Computer Enhancement thanks to its use of optimization. Its use of knowledge / data mining is restricted to the very superficial counting of trigrams and of topical words in its source text. However, some other systems in the Knowledge Representation category use similarly superficial methods to gather topical words (Toivanen et al., 2012, 2014; Tobing and Manurung, 2015; Rashel and Manurung, 2014b). TwitSong's Process could be improved if it used a more sophisticated form of knowledge representation, using techniques like named entity recognition or a vector space model so as to match the current state of the art for knowledge representation in poetry. We feel that it is more important to improve the results of TwitSong's optimization techniques first, rather than adding knowledge representation for knowledge representation's sake (but it might be that better knowledge representation would enable better optimization for Topicality and Emotion).

Finally, from a **Press** perspective, our one attempt at a Press experiment shows that TwitSong is not especially successful from this perspective. A more coherent Product would likely enable TwitSong to attract more readers and attention, and thus improve its Press creativity.

Chapter 6

Discussion, limitations, and future work

6.1 Unsolved questions in poetry evaluation

Our set of four Product criteria—Reaction, Meaning, Novelty, and Craft—with their sub-criteria, is not an empirically validated poetry rating system. They are domain-specific criteria differentiated by what seem to be reasonable conceptual distinctions to us. Their direct base in the responses of poetry experts to computer-generated poetry renders them more solidly evidence based than other domain-specific poetry criteria currently in use. However, much work remains to be done before we can say that we have a valid evidence-based set of criteria for poems.

In particular, our results from the third generation of TwitSong show that the criteria are not necessarily statistically distinct from each other, even for experts. More work needs to be done evaluating experts' ratings of poems on all of the various sub-criteria and seeing if there is a statistical grouping that fits the data better than our initial four-way grouping. After this is done, a standardized questionnaire should be developed which addresses all sub-criteria and which can be tested for its reliability and validity on diverse types of both human and computer-generated poetry.

The questions used in our existing studies are still essentially *ad hoc*, since they only address certain sub-criteria and only in ways which we happened to feel were relevant to the work we were doing. In particular, in the third study (Section 5.3) we failed to address Coherence, which was a major issue for both expert and non-expert raters. However, Coherence is a sub-criterion belonging to our four-criteria model, so this is not a deficiency of our model, merely a deficiency of how we chose to ask questions based on the model. A standardized and validated questionnaire would avoid this issue.

Many computer-generated poetry projects are not yet at a stage where the researchers intend to make finished poetry successfully satisfying all of the criteria. Rather, they are intended as proofs of concept or as explorations of a single criterion. This is not necessarily an argument against using a model such as our four criteria. One could imagine that, after a standardized questionnaire is developed, either the whole questionnaire or a part of the questionnaire should be used. For example, if a researcher was solely

attempting to devise a way of making computer-generated poetry more meaningful, and did not care about the other criteria, then they could still use the standardized questionnaire, but would simply only evaluate the questions assigned to the category of Meaning. Thus, a fully validated method of evaluating the Product creativity of poems would benefit everyone, even if their research goals were narrower than ours.

6.2 Statistical considerations

Although it is customary to describe the work in a thesis as a research program which was planned in advance and then executed, the nature of our research over the years of our thesis work was not quite like this. Instead, work in different areas of the thesis was undergone concurrently, modified, influenced, and revisited in a nonlinear manner, and discoveries made partway through one area of the work shed light on how we should have, or could have done other parts better. There are several purely mathematical and methodological aspects of our research that we would do differently if we were doing it over again.

The statistical methods used changed between different generations of TwitSong, which raises methodological questions. It is difficult to compare one generation of TwitSong to another when each generation is evaluated in a different way: sometimes numeric scales and sometimes pairwise choice, different numbers of options on the numeric scales, and different phrasings, groupings, or selections of questions for the judges surveyed. Also, each generation of TwitSong changes multiple aspects of the poems compared to the previous generation, not only the line generation methods but also the form of the poems, the source material used, and small ad hoc aspects of the program such as the methods used to identify unpronounceable lines, lines that stop in the middle of a word, etc.

This does not invalidate our entire research plan, since our aim was not, precisely speaking, to compare one generation of TwitSong to another. Instead, we viewed each generation as its own self-contained experiment. The goal of Generation Two was not to produce poems that could be directly compared with Generation One, but to demonstrate that the line rating methods used in Generation Two were effective compared to a version of Generation Two without them. Similarly, the goal of Generation Three was not to produce "better" poems than Generation Two, but to demonstrate that the Editorial Algorithm produced better poems than a version of Generation Three that did not edit its lines. However, aside from the fact that neither of these things were demonstrated in the way that we hoped, this lack of direct comparability between generations does limit the scope of the findings from each successive generation.

As for the different uses of different evaluation methods, this is a limitation in research which is in large part meant to demonstrate the proper use of the principles of evaluation that we detail in Chapter 2. However, until a standardized method for computational poetry evaluation is fully developed and validated, it will continue to be the case that all evaluations of the type we have been doing will be in some sense ad hoc. Parameters such as phrasing of the questions and numbers available on a numeric rating scale cannot be derived a priori from theory, and there will be no justification for standardizing them until validity and reliability testing has been performed. However, it is true that using different evaluations for different generations of TwitSong increases the problem of not being able to compare different generations to each other, and simply picking a detailed evaluation methodology and sticking to it would have prevented this. In particular, the choice of specific sub-sets of our four desiderata (imagery as a stand-in for all of Craft, for instance) is arbitrary and could have been improved. In many cases, we cannot justify the use of

particular parameters for our evaluation, such as number of options on a numeric rating scale, because we do not remember why we chose the values for those parameters that we did.

6.3 Developing better line evaluation metrics

Our experiment with the first generation of TwitSong shows clearly that humans choosing lines based on topic and sentiment—and, to a lesser extent, imagery—does measurably improve the resulting poems. Choosing lines based on meter, a more simple numerical calculation, also seems to improve them. (It should be noted that meter is not necessarily a mechanistic calculation; for instance, there are single syllable words which can be stressed or unstressed depending on their context. However, meter can be at least estimated more easily than variables such as meaning.) But the second and third generation experiments show that our automated measures of topic and sentiment are not adequate to the task. An urgent task for improving a system like TwitSong would therefore be to improve the automated measures. Detecting topic and sentiment in text is an open research problem, but more sophisticated efforts than ours certainly already exist; we deliberately chose simple methods in a "quick and dirty" frame of mind, hoping to generate many attempts at poems very rapidly, and it appears that we chose poorly. Among others, Ghosh *et al.*'s neural language model (Ghosh et al., 2017) is a more sophisticated example of generating text based on a chosen emotion.

Furthermore, our line evaluation metrics cover only a fraction of the possible metrics covered under our four criteria. Many other metrics for improving poetry are possible. For instance, Kao and Jurafsky (Kao and Jurafsky, 2015) perform a computational analysis of stylistic differences on 16 features between professional contemporary poetry, amateur contemporary poetry, early 20th century Imagist poetry, and professional 19th century poetry. Some of their results were as follows:

- **Concreteness.** Compared to 19th century poetry, contemporary poetry has more object words, less abstractness, and more concreteness. Amateur poetry is less concrete and less imageable than the 19th century poetry. ("Abstractness", "concreteness", and "imageability" are measures based on various special purpose dictionaries.)
- **Emotion.** Compared to 19th century poetry, contemporary poetry has fewer emotion words and lower emotional arousal. Amateur poetry has similar arousal to the 19th century poetry, but also with even more emotion words and higher valence.
- **Sound features.** Compared to 19th century poetry, contemporary poetry has fewer perfect end rhymes and more assonance (repetition of vowel sounds). Amateur poetry has fewer perfect end rhymes as well, but also more identical end rhymes and less consonance.
- **Vocabulary.** Amateur poets use shorter words and a smaller type-token ratio than 19th century poets. There were no differences between 19th century and contemporary professional poets on these axes.

Kao and Jurafsky attribute the differences between 19th century and contemporary poetry to the influence of Imagism, a movement emerging in the early twentieth century from which many of the axioms taught to beginning poets today have emerged, including the emphasis on concrete imagery (Kao and Jurafsky, 2015). From their results we see that stylistic differences between genres of poetry can be nuanced and involve

more than one axis. We also see that amateur poetry tends to be in some ways "outdated", resembling the linguistic features of older professional poetry more than it resembles the current century's professional poetry, but amateur poetry is also different from professional poetry in other, non-time-dependent ways.

Many other possible textual markers of "good" poetry are possible other than Kao and Jurafsky's. Hirjee and Brown successfully use rhyme patterns detected by RhymeAnalyzer (Hirjee, 2010) to distinguish lyrics written by different rappers, which suggest that internal rhymes and other features detectable with this tool could also be used in other forms of poetry. Kaplan and Blei (Kaplan and Blei, 2007), who are cited by Kao and Jurafsky, use a greater variety of textual measures including orthographic measures (words per poem, words per line, lines per poem, stanzas per poem, word length, lines per stanza); POS frequency and other measures based on POS frequency styles; and internal rhymes. (They use these measures to distinguish the styles of individual poets, rather than to comment on which measures contribute to "good" or professional poetry.)

For more semantically oriented measures, the Linguistic Inquiry and Word Count method (Pennebaker et al., 2001) can be used to measure word frequency in a variety of categories, including words relating to various emotions and to various topics such as the body, social relationships, and cognitive processes such as uncertainty, as well as detecting subtle patterns in function words which can be indicative of personality and emotional state. Dalvean (2015) uses LIWC, along with another set of psycholinguistic word norms, to analyze the difference between amateur and professional contemporary poets. Important characteristics of professional poetry in Dalvean's analysis, when compared to amateur poetry, include more concrete, demonstrative language; more unusual words; fewer emotion words; more number words; and fewer references to time. LIWC is also used to analyze the personality of writers (Tausczik and Pennebaker, 2010) which could lead to the creation of personality variables such as the ones used by InkWell (Gabriel, 2016).

It is worth asking, as many of our fellow researchers did upon seeing our experiment with human poetry (Section 4.1), whether or not the distinction between amateur and professional contemporary poetry is the distinction on which to base computationally creative optimization techniques. After all, an unskilled computer-generated poem may or may not resemble an unskilled human poem in any of these particular ways. Indeed, some researchers have said that they would be delighted for their computer to generate poetry resembling that of a human amateur. Such attitudes are understandable, because a computer that could emulate a human amateur would likely have solved several problems that are endemic to computer poets, especially the problem of coherence (see Section 6.5, below). However, it may not be the case that an unskilled computer is "more amateur" than an unskilled human. Instead, it may be that both the computer and the unskilled human have their own separate ways of differing from a skilled human. Additionally, different unskilled computers may diverge from the skills of professional humans in different ways. For instance, the characteristic patterns of text generated by a Markov model may differ from those of a template or other method.

With all of this in mind, we suggest the following procedure for determining line rating metrics to use in a large project like TwitSong:

1. Make an exhaustive list of all the potential metrics that might theoretically be of interest in your project. These metrics must include an implementation (e.g. "Imagery as measured by the Regressive Imagery Dictionary" as opposed to just "Imagery"). They might reasonably include all of the metrics discussed in the above studies, unless there are theoretical or practical reasons why some of them are not of interest.

2. Identify a reasonably large "base corpus" from which your poetic text will be materially derived and a reasonably large "target corpus" which exemplifies the poetic traits that you wish for your text to acquire. For instance, in the first and second generations of TwitSong, the "base corpus" would be an uncurated set of lines from Twitter; in the third generation, it would be a set of news articles. A target corpus for TwitSong is not expressly defined, but a corpus of contemporary English poetry would be a reasonable choice. One might alternately choose a corpus of, for example, humorous poetry in English, or of serious contemporary poetry specifically based on the news, such as the poetry published in Rattle's Poets Respond[1].

3. Run a statistical regression similar to Dalvean's (2015) to identify which of these metrics are actually useful for distinguishing one corpus from another. (Care should be taken when selecting semantic metrics if the topics of the base and target corpus differ. For instance, when distinguishing news articles from contemporary English poetry, it might be expected that news articles would contain more words related to events, more present tense and third person, etc—not because these words inherently make them less "poetic," but because these words are characteristic of how news is written and the target corpus is not news. Metrics that could provide confounding results in this way should be removed before performing the regression.)

4. Choose the metrics indicated by the regression model for purposes of line selection and editing in your system.

This method would become cumbersome in some circumstances, such as when there were many different unrelated types of base or target corpus, but it would lead to better results. It would also not solve all line metric implementation problems with TwitSong. For example, it would be difficult to use this method to implement topicality, since the target corpus is likely chosen for stylistic reasons rather than for being more on-topic than the base corpus. Still, we frequently thought of doing an analysis like this for TwitSong but chose to work on other parts of the system instead. We regret this choice.

As a side note, Kao and Jurafsky's (2015) and Dalvean's (2015) analyses both point to less emotion words as a hallmark of better contemporary poetry, while TwitSong, as well as many other computational poetry systems, specifically seeks words related to a specific emotion. (Indeed, Kao, Jurafsky, and Dalvean would seem to contradict the non-expert raters who strongly preferred positive emotion in the first generation of Twitsong—though, of course, those were non-experts.) It may be that these goals can be reconciled somehow. As Kao and Jurafsky (2015) point out, the goal of Imagist poetry is not to be devoid of emotion, but rather, to express emotions implicitly by describing a concrete image which evokes them. A dictionary such as the NRC Hashtag Emotion Lexicon includes many such concrete words as well as directly emotional words. It may be that by maximizing an emotional measure that includes concrete words, but minimizing one one more like the LIWC measure of Emotion Words—or by aiming for words within a certain, non-maximal emotional range—a computer could achieve this balance. It may also be that the two sides of such a paired implementation would simply cancel each other out. An analysis like the one described above would be necessary in order to discern whether or not a particular manipulation of dictionaries is a good practice.

[1] `https://www.rattle.com/respond/`

6.4 Developing intelligent editing strategies

As might be obvious from reading Section 5.3.2.3, our line editing strategies, independently of the metrics used to judge the lines, could use some work. Our current editing strategy is fairly liberal in what it selects to replace, and each candidate replacement line replaces both the selected word(s) and everything after them with arbitrarily generated Markovian text. The hope is that, with enough of these line regenerations, some of them will happen to have better scores on our line metrics. This strategy does work adequately for line metrics like meter which are implemented adequately, but it is not especially sophisticated. A better strategy would be more selective—which would be easier to achieve with better line rating metrics. More importantly, an ideal strategy would have some way of ensuring that its replacements are actually an improvement, at least in some narrow way. For example, with imagery, one could imagine an alternate Markov chain which is constrained to only generate concrete words, or which is heavily weighted in favor of such words.

6.5 Coherence

The biggest problem with TwitSong for many readers is not the individual line metrics, but the poems' overall coherence. While coherence was not directly addressed in the third generation's evaluation, it is a part of our set of four poetry criteria (a sub-criterion of Meaning). Coherence is a complex trait of a text with both syntactic features (e.g. grammatical correctness) and semantic ones ("making sense", or having the meanings of words relate to each other in a way that feels natural and easy to follow, as well as following the development of a single idea throughout the text).

TwitSong is not alone among poetry systems for struggling with coherence. While it is hard to assess the coherence of poetry in a language we do not speak, readers will note that nearly all of the English poetry excerpts in Section 3.2 are incoherent. Early versions of McGonagall (Manurung, 1999) are syntactically coherent, but are semantically coherent only when the semantics are specifically handcrafted. Stereotrope (Veale, 2013a) is coherent, but at the cost of feeling repetitive and trite, which caused poetry experts in Section 4.2 to dislike it more than most of the incoherent poems. InkWell (Gabriel, 2016) is coherent enough to pass for human with a group of poetry experts, but it is easier to write passably coherent haiku than a coherent sonnet or quatrain, as human readers expect a haiku to be a short sequence of loosely connected, potentially fragmentary images. Goodwin's (2016) neural network poetry can be quite coherent, but Goodwin states that the results are uneven. Overall, coherence is still an important unsolved problem in computer-generated poetry.

TwitSong's first and second generations produce work that is reasonably grammatically correct, since its phrasings within a line are taken from tweets written by humans, though this does not produce overall coherence across the poem's different lines. TwitSong's third generation uses a Markov model, and produces text that feels notably Markovian and rambling. It could be made more grammatical by choosing a different mere generation method. For example, it could use a context-free grammar. Or, like ePoGeeS (Roque, 2011), it could use a class-based n-gram model to ensure that appropriate parts of speech follow each other. We would not expect this to solve the overall coherence problem, but there is such a thing as one version of a system having more or less coherence than another.

A knowledge base is another way to try to ensure coherence, though, as we have seen in Section 3.2.2, the use of knowledge representation does not in and of itself guarantee this. Using knowledge representation

would also represent a significant change to the architecture of TwitSong, as the knowledge would have to be factored in to the line generation and editing techniques somehow. A minimally invasive modification might be to use knowledge representation, either through a vector space model or through parsing and named entity recognition on individual base texts, as the measure of Topicality, but while this might easily improve Topicality line judgments, it would likely not make poems coherent on its own.

More interesting, and less well-explored, are methods for ensuring coherence across the different lines of a poem by investigating poetic structure. For example, one could augment Kao and Jurafsky's (2015) and Dalvean's (2015) methods of choosing desired line properties by also studying how these properties change across the course of a poem, particularly a longer poem like a sonnet. Does a human poem display about the same level of topicality throughout, or does its level of topicality tend to increase (or decrease, or move in some more complex but predictable pattern) as the poem goes on? Simonton (1990) applies this type of analysis to Shakespearean sonnets, finding that sonnets tend to increase in their number of unique words and number of syllables per word, until the final couplet, where these measures (as well as primary process imagery) drop again. But in the most successful sonnets, these effects are reduced. We have not encountered other work in the digital humanities which measures the progression of other metrics across the course of the poem, or which studies other (non-sonnet) forms of poetry in this way.

An even more promising way to create coherence across the lines of a poem might emerge from argumentation mining (Palau and Moens, 2009), which explores how different statements are used to support the main ideas of a text. Argumentation mining was developed for persuasive speech, but has also been applied to narrative (Bex and Bench-Capon, 2014). A poetic application of argumentation mining, whether for analysis or generation, would be extremely interesting. So would an application of computational rhetoric, which analyzes text for classic rhetorical structures such as anaphora and chiasmus (Harris and DiMarco, 2009).

6.6 Black boxes vs white boxes

Given the good results of Goodwin (2016) and the Chinese community, one might reasonably ask why we do not switch entirely to neural networks for poetry generation, and classify efforts like TwitSong as obsolete.

We feel that, with so many unsolved questions remaining in computational creativity, both neural network-based and rule-based approaches are useful in different ways for solving these problems. The advantage of rule-based systems is that they are not "black boxes". A system like TwitSong can (potentially) justify why it made the creative decisions that it did. This is not to say that TwitSong passes Guckelsberger's (2017) recursive "why?" test; merely that it is possible, as we did partially in Section 5.3.2.3, to trace the operations that are made on a particular line and to see precisely where a word was changed because it was considered not topical enough, not emotional enough, etc. A neural network cannot do this, and can largely only be judged on its output. Again referencing Pérez y Pérez (2018), this "black box" nature does not make work from the engineering-mathematical perspective impossible, but it certainly makes cognitive-social insights difficult.

Neural networks may at present give better results, but the gap is not yet large enough to justify abandoning rule-based systems altogether. Rule-based systems like InkWell (Gabriel, 2016), which make traceable decisions about lines in a manner analogous to TwitSong's, have also on occasion produced good

results. And even some very recent neural network-based poetry systems, such as Loller-Anderson and Gambäck's (2018), still struggle with the same basic problems of coherence as most rule-based systems. Both approaches have untapped promise, and both require further development.

6.7 Interactivity

A final pipe dream that we have had for this system is to make TwitSong or something like it into a fully interactive system with a graphical user interface. A user could choose a source text, choose between any other parameters that have multiple options (such as poetic form or emotion expressed), and also choose *weights* for the different line evaluation metrics. For example, a user could decide that they want to emphasize meter but do not care about imagery, or that emotion is more important than topic, or vice versa. This would likely be a thing to work on after improving and diversifying the line evaluation metrics as described above, but it would lead to a usable, useful co-creative system in which the human user has fine-grained control over what sort of thing is generated, but TwitSong uses its information processing capabilities to provide novel and valuable output that the human could not have composed on their own. Our results with TwitSong were not advanced enough to allow us to usefully implement this dream. But we predict that, in the next 10 years, this type of system, with humans and computers working to augment each other's abilities, will be one of the most valuable products to come out of the computational creativity community.

Chapter 7

Conclusion

We have now surveyed the theory and practice, from an interdisciplinary perspective, of how to evaluate computational creativity; surveyed current trends in computer-generated poetry and fit them into a larger taxonomy of poetry techniques; performed two important experiments into the role of expertise in poetry evaluation; and developed and tested three separate versions of a found poetry generation system.

Our evaluation survey is the most comprehensive survey of its type to date and provides ample material for future researchers to use, both in following current evaluation best practices and in using insights from psychology and philosophy to guide the future development of new evaluations. Our poetry taxonomy is broader in scope than other similar taxonomies and focuses on the artistic purposes of any computational techniques used, rather than on algorithmic minutiae. Our survey of the state of the art in poetry shows that many interesting Process techniques are being used, but only a few produce coherent and polished Products. There is still a great deal of room for the practice of computational poetry to evolve.

In our evaluation experiments, we have underscored the importance of expert judges in poetry evaluation. Experts are consistently better judges than non-experts, but in contemporary poetry, the problem is particularly acute, with non-experts' preferences being the reverse of experts'. Most "good" contemporary poetry preferred by experts is hard for non-experts to understand, and non-experts therefore prefer more understandable "bad" work. This preference for understandable work and for less novelty overwhelms any other distinction between proposed poetry evaluation criteria, at least for non-experts.

Testing the preferences of expert judges on computer-generated poetry, we found that some agreement existed but that it was limited in scope. We used the written comments of the experts to identify a new set of criteria for computer-generated poetry: Reaction, Meaning, Novelty, and Craft, along with several sub-criteria for each of these. While this set of criteria has yet to be developed into a reliable and validity-tested evaluation tool, it provides a starting point which is more evidence-based than other proposed sets of evaluation criteria for poetry.

Finally, taking all this work into account, we have developed and tested three versions of our own poetry system, TwitSong. The first generation of TwitSong is a human-in-the-loop system and a proof of concept which shows that, by taking a source text and selecting lines that are better on certain metrics, a better poem can be constructed. The second and third generations build on this proof of concept by automating line judgments based on the same set of metrics. The second generation builds found poetry sonnets out

of Twitter, while the third generation builds quatrains out of news articles. The major innovation of the third generation is that it edits its own work, through an Editorial Algorithm which is inspired by WASP (Gervás, 2016) but which goes further than WASP does, as the third generation of TwitSong is able to make edits based on semantic traits such as emotion and topicality. The performance of both generations is mixed, both due to a lack of coherence and due to less sophisticated ways of operationalizing the line evaluation metrics. However, experiments with the third generation show that its Editorial Algorithm does improve the resulting poetry in at least some ways.

In a research domain as young as computer-generated poetry, there is always far more left undone than done, as even a successful line of research will yield more questions for further research than answers. We hope that our work in this thesis is a bountiful and productive source of such questions and of inspiration for future researchers in this field.

We end by asking TwitSong itself to provide a closing statement, in the form of a joyful quatrain using the draft version of this thesis as a source text.

CREATIVE HUMANS
and the lines in the vector space
and palmer russian spy
based on the four combined in place
we will look very high

References

M.H. Abrams. Poetry, theories of. In *Princeton Encyclopedia of Poetry and Poetics, Enlarged Edition*, pages 639–649. Princeton University Press, 1974.

Selcuk Acar, Cyndi Burnett, and John F. Cabra. Ingredients of creativity: Originality and more. *Creativity Research Journal*, 29(2):133–144, 2017.

Manex Agirrezabal, Bertol Arrieta, Aitzol Astigarraga, and Mans Hulden. POS-tag based poetry generation with WordNet. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 162–166. Association for Computational Linguistics, 2013.

Wendy Aguilar and R Pérez y Pérez. Criteria for evaluating early creative behavior in computational agents. In *Proceedings of the Fifth International Conference on Computational Creativity*, pages 284–287, Ljubljana, Slovenia, 2014. Association for Computational Creativity.

Teresa Amabile. *Componential theory of creativity*. Harvard Business School, Boston, MA, 2012.

Teresa M Amabile. A consensual technique for creativity assessment. In *The Social Psychology of Creativity*, pages 37–63. Springer, 1983a.

Teresa M Amabile. The social psychology of creativity: a componential conceptualization. *Journal of Personality and Social Psychology*, 45(2):357, 1983b.

Aitzol Astigarraga, José María Martínez-Otzeta, Igor Rodriguez Rodriguez, Basilio Sierra, and Elena Lazkano. Poet's little helper: A methodology for computer-based poetry generation. a case study for the Basque language. In *Proceedings of the Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2017)*, pages 2–10, 2017.

M Dorothee Augustin, Claus-Christian Carbon, and Johan Wagemans. Artful terms: a study on aesthetic word usage for visual art versus film and music. *i-Perception*, 3(5):319, 2012.

b arco cultural centre. Poetweet, 2013. `http://poetweet.com.br/`, accessed May 2018.

John Baer. The case for domain specificity of creativity. *Creativity Research Journal*, 11(2):173–177, 1998.

John Baer. How divergent thinking tests mislead us: are the Torrance Tests still relevant in the 21st century? the Division 10 debate. *Psychology of Aesthetics, Creativity, and the Arts*, 5(4):309, 2011.

John Baer. Domain specificity and the limits of creativity theory. *The Journal of Creative Behavior*, 46 (1):16–29, 2012.

John Baer and Sharon S McKool. Assessing creativity using the consensual assessment technique. *Handbook of Assessment Technologies, Methods and Applications in Higher Education*, pages 65–77, 2009.

Anna E. Balakian. Dadaism. In *Princeton Encyclopedia of Poetry and Poetics, Enlarged Edition*, page 180. Princeton University Press, 1974.

Gabriele Barbieri, François Pachet, Pierre Roy, and Mirko Degli Esposti. Markov constraints for generating lyrics with style. In *Proceedings of the 20th European Conference on Artificial Intelligence*, pages 115–120. IOS Press, 2012.

Christopher Bartel. Originality and value. *British Journal of Aesthetics*, 25:169–184, 1985.

Monroe C Beardsley. On the creation of art. *Journal of Aesthetics and Art Criticism*, 23(3):291–304, 1965.

Tarek R Besold. The unnoticed creativity revolutions: bringing problem-solving back into computational creativity. In *Proceedings of the AISB 3rd International Symposium on Computational Creativity*, pages 1–8, Sheffield, UK, 2016. CRC Press, Taylor & Francis Group.

Floris Bex and Trevor JM Bench-Capon. Understanding narratives with argumentation. In *COMMA*, pages 11–18, 2014.

Ranjit Bhatnagar. Pentametron, 2012. `http://pentametron.com`, accessed December 9, 2015.

Debarun Bhattacharjya. Preference models for creative artifacts and systems. In *Proceedings of the Seventh International Conference on Computational Creativity*, pages 52–59, Paris, France, 2016. Association for Computational Creativity.

Jannece Blijlevens, Clementine Thurgood, Paul Hekkert, Lin-Lin Chen, Helmut Leder, and T.W. Allan Whitfield. The aesthetic pleasure in design scale: The development of a scale to measure aesthetic pleasure for designed artifacts. *Psychology of Aesthetics, Creativity, and the Arts*, 11(1):86–98, 2017.

Margaret A Boden. *The Creative Mind: Myths and Mechanisms*. Psychology Press, Hove, UK, 1990.

Alessandro Bollo and Luca Dal Pozzolo. Analysis of visitor behaviour inside the museum: an empirical study. In *Proceedings of the Eighth International Conference on Arts and Cultural Management*, volume 2, Montreal, Canada, 2005. International Association of Arts and Cultural Management.

Samira Bourgeois-Bougrine, Vlad Glaveanu, Marion Botella, Katell Guillou, Pierre Marc De Biasi, and Todd Lubart. The creativity maze: exploring creativity in screenplay writing. *Psychology of Aesthetics, Creativity, and the Arts*, 8(4):384, 2014.

Oliver Bown. Generative and adaptive creativity: a unified approach to creativity in nature, humans and machines. In *Computers and Creativity*, pages 361–381. Springer, Berlin, 2012.

Oliver Bown. Empirically grounding the evaluation of creative systems: incorporating interaction design. In *Proceedings of the Fifth International Conference on Computational Creativity*, pages 112–119, Ljubljana, Slovenia, 2014. Association for Computational Creativity.

Oliver Bown. Player responses to a live algorithm: conceptualising computational creativity without recourse to human comparisons? In *Proceedings of the Sixth International Conference on Computational Creativity*, pages 126–133. Association for Computational Creativity, 2015.

Kevin Burns. Computing the creativeness of amusing advertisements: a Bayesian model of Burma-Shave's muse. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 29(01):109–128, 2015.

Donald T Campbell. Blind variation and selective retentions in creative thought as in other knowledge processes. *Psychological Review*, 67(6):380, 1960.

Linda Candy and Zafer Bilda. Understanding and evaluating creativity. In *Proceedings of the Seventh ACM Conference on Creativity and Cognition*, pages 497–498, Georgia Tech University, USA, 2009. ACM.

Jim Carpenter. Public override void, 2004. `https://slought.org/resources/public_override_void`, accessed December 9, 2015.

Jim Carpenter. etc4, 2007. Blog post. `http://theprostheticimagination.blogspot.ca/2007/07/etc4.html`.

Erin A Carroll and Celine Latulipe. The creativity support index. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*, pages 4009–4014, Boston, US, 2009. ACM.

Shelly H. Carson, Jordan B. Peterson, and Daniel M. Higgins. Reliability, validity, and factor structure of the creative achievement questionnaire. *Creativity Research Journal*, 17(1):37–50, 2005.

William Chamberlain. *The Policeman's Beard is Half Constructed*. Warner Books, 1984.

Heather Chan and Dan A Ventura. Automatic composition of themed mood pieces. In *Proceedings of the Fifth International Joint Workshop on Computational Creativity*, pages 19–115,28, Universidad Complutense de Madrid, Spain, 2008. Association for Computational Creativity.

John Charnley, Alison Pease, and Simon Colton. On the notion of framing in computational creativity. In *Proceedings of the Third International Conference on Computational Creativity*, pages 77–82, Dublin, Ireland, 2012. Association for Computational Creativity.

John William Charnley, Simon Colton, and Maria Teresa Llano. The FloWr Framework: Automated flowchart construction, optimisation and alteration for creative systems. In *Proceedings of the Fifth International Conference on Computational Creativity*. Association for Computational Creativity, 2014.

Wayne Clements. Poetry beyond the Turing test. In *Proceedings of Electronic Visualization and the Arts (EVA 2016)*, 2016.

Tom Collins and Robin Laney. Computer-generated stylistic compositions with long-term repetitive and phrasal structure. *Journal of Creative Music Systems*, 1, 2017.

Simon Colton. Creativity versus the perception of creativity in computational systems. In *AAAI Spring Symposium: Creative Intelligent Systems*, pages 14–20, Palo Alto, US, 2008. Association for the Advancement of Artificial Intelligence.

Simon Colton. The painting fool: stories from building an automated painter. In *Computers and Creativity*, pages 3–38. Springer, Berlin, 2012.

Simon Colton and Geraint A Wiggins. Computational creativity: the final frontier? In *Proceedings of the 20th European Conference on Artificial Intelligence*, pages 21–26, Montpellier, France, 2012. IOS Press.

Simon Colton, A Pease, and J Charnley. Computational creativity theory: The FACE and IDEA descriptive models. In *Proceedings of the Second International Conference on Computational Creativity*, pages 90–95, Mexico City, Mexico, 2011. Association for Computational Creativity.

Simon Colton, Jacob Goodwin, and Tony Veale. Full face poetry generation. In *Proceedings of the Third International Conference on Computational Creativity*, pages 95–102. Association for Computational Creativity, 2012.

Simon Colton, Alison Pease, Joseph Corneli, Michael Cook, and Teresa Llano. Assessing progress in building autonomously creative systems. In *Proceedings of the Fifth International Conference on Computational Creativity*, pages 137–145, Ljubljana, Slovenia, 2014. Association for Computational creativity.

Simon Colton, Teresa Llano, Rose Hepworth, John Charnley, Cat Gale, Archie Baron, François Pachet, Pierre Roy, Pablo Gervás, Nick Collins, et al. The beyond the fence musical and computer says show documentary. 2016.

Simon Colton, Alison Pease, and Rob Saunders. Issues of authenticity in autonomously creative systems. In *Proceedings of the Ninth International Conference on Computational Creativity*. Association for Computational Creativity, 2018.

Michael Cook and Simon Colton. Neighbouring communities: Interaction, lessons and opportunities. 2018a.

Michael Cook and Simon Colton. Redesigning computationally creative systems for continuous creation. In *Proceedings of the Ninth International Conference on Computational Creativity*. Association for Computational Creativity, 2018b.

Joseph Corneli, Anna Jordanous, Rosie Shepperd, Maria Teresa Llano, Joanna Misztal, Simon Colton, and Christian Guckelsberger. Computational poetry workshop: making sense of work in progress. In *Proceedings of the Sixth International Conference on Computational Creativity*, pages 268–275, Park City, US, 2015. Association for Computational Creativity.

David Cropley and Arthur Cropley. Elements of a universal aesthetic of creativity. *Psychology of Aesthetics, Creativity, and the Arts*, 2(3):155–161, 2008.

David H Cropley, James C Kaufman, and Arthur J Cropley. Malevolent creativity: A functional model of creativity in terrorism and crime. *Creativity Research Journal*, 20(2):105–115, 2008.

DH Cropley and AJ Cropley. Engineering creativity: a systems concept of functional creativity. In *Creativity Across Domains: Faces of the Muse*, pages 169–185. Lawrence Erlbaum Associates, Inc, Denmark, 2005.

Mihaly Csikszentmihalyi. *Creativity: flow and the psychology of discovery and invention*. Harper Collins, New York, 1996.

Mihaly Csikszentmihalyi. Implications of a systems perspective for the study of creativity. In *Handbook of Creativity*, pages 313–338. Cambridge University Press, Cambridge, UK, 1999.

Joao M Cunha, Joao Gonçalves, Pedro Martins, Penousal Machado, and Amílcar Cardoso. A pig, an angel and a cactus walk into a blender: a descriptive approach to visual blending. In *Proceedings of the Eighth International Conference on Computational Creativity*, pages 80–87, Atlanta, US, 2017. Association for Computational Creativity.

Palle Dahlstedt. Between material and ideas: a process-based spatial model of artistic creativity. In *Computers and Creativity*, pages 205–233. Springer, Berlin, 2012.

Michael Dalvean. Ranking contemporary American poems. *Literary and Linguistic Computing*, 30(1): 6–19, 2015.

Amitava Das and Björn Gambäck. Poetic machine: computational creativity for automatic poetry generation in Bengali. In *Proceedings of the Fifth International Conference on Computational Creativity, ICCC*, pages 230–238, Ljubljana, Slovenia, 2014. Association for Computational Creativity.

Subrata Dasgupta. Contesting (Simonton's) blind variation, selective retention theory of creativity. *Creativity Research Journal*, 23(2):166–182, 2011.

Belén Díaz-Agudo, Pablo Gervás, and Pedro A González-Calero. Poetry generation in COLIBRI. In *Advances in Case-Based Reasoning*, pages 73–87. Springer, 2002.

Jennifer Diedrich, Mathias Benedek, Emanuel Jauk, and Aljoscha C Neubauer. Are creative ideas novel and useful? *Psychology of Aesthetics, Creativity, and the Arts*, 9(1):35, 2015.

Steve DiPaola, Graeme McCaig, Kristin Carlson, Sara Salevati, and Nathan Sorenson. Adaptation of an autonomous creative evolutionary system for real-world design application based on creative cognition. In *Proceedings of the Fourth International Conference on Computational Creativity*, pages 40–47, Sydney, Australia, 2013. Association for Computational Creativity.

Alan Dorin and Kevin B Korb. Creativity refined: bypassing the gatekeepers of appropriateness and value. In *Computers and Creativity*, pages 339–360. Springer, Berlin, 2012.

Jeremy Douglass. Numeracy and electronic poetry. *Journal of Mathematics and the Arts*, 8(1-2):13–23, 2014.

Shlomo Dubnov, Kevin Burns, and Yasushi Kiyoki. Cross-cultural aesthetics: analyses and experiments in verbal and visual arts. In *Cross-Cultural Multimedia Computing*, pages 21–41. Springer, Switzerland, 2016.

edde addad. comments on "A taxonomy of generative poetry techniques", 2018. `https://gnoetrydaily.wordpress.com/2018/03/25/comments-on-a-taxonomy-of-generative-poetry-techniques/`, accessed April 2018.

Ernest Edmonds, Lizzie Muller, and Matthew Connell. On creative engagement. *Visual Communication*, 5(3):307–322, 2006.

Arne Eigenfeldt. Generative music for live musicians: an unnatural selection. In *Proceedings of the Sixth International Conference on Computational Creativity*, pages 142–149. Brigham Young University, 2015.

Arne Eigenfeldt, Oliver Bown, Andrew R Brown, and Toby Gifford. Distributed musical decision-making in an ensemble of musebots: dramatic changes and endings. In *Proceedings of the Eighth International Conference on Computational Creativity*, pages 88–95, Atlanta, US, 2017. Association for Computational Creativity.

Ahmed Elgammal and Babak Saleh. Quantifying creativity in art networks. In *Proceedings of the Sixth International Conference on Computational Creativity*, pages 39–46, Park City, US, 2015. Association for Computational Creativity.

Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. CAN: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms. In *Proceedings of the Eighth International Conference on Computational Creativity*, pages 96–103, Atlanta, US, 2017. Association for Computational Creativity.

eRoGK7 et al. Gnoetry daily, 2011. `https://gnoetrydaily.wordpress.com`, accessed December 9, 2015.

Frieda Fayena-Tawil, Aaron Kozbelt, and Lemonia Sitaras. Think global, act local: a protocol analysis comparison of artists' and nonartists' cognitions, metacognitions, and evaluations while drawing. *Psychology of Aesthetics, Creativity, and the Arts*, 5(2):135, 2011.

Celso França, Luıs Fabrıcio W Góes, Alvaro Amorim, Rodrigo Rocha, and Alysson Ribeiro da Silva. Regent-dependent creativity: a domain independent metric for the assessment of creative artifacts. In *Proceedings of the Seventh International Conference on Computational Creativity*, pages 68–76, Paris, France, 2016. Association for Computational Creativity.

Ronald S Friedman and Christa L Taylor. Exploring emotional responses to computationally-created music. *Psychology of Aesthetics, Creativity, and the Arts*, 8(1):87, 2014.

Chris T Funkhouser. *New directions in digital poetry.* A&C Black, 2012.

Christopher Thompson Funkhouser. *Prehistoric digital poetry: an archaeology of forms, 1959-1995.* University Alabama Press, 2007.

Richard P Gabriel. in the control room of the banquet. In *Proceedings of the 2016 ACM International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*, pages 250–268. ACM, 2016.

Prasad Gade, Mary Galvin, James O'Sullivan, Paul Walsh, and Órla Murphy. Reactions to imagery generated using computational aesthetic measures. *Leonardo*, 50(5), 2017.

Philip Galanter. Computational aesthetic evaluation: past and future. In *Computers and Creativity*, pages 255–293. Springer, Berlin, 2012.

Rodrigo García, Pablo Gervás, Raquel Hervás, Rafael Pérez, and Fernando ArÃmbula. A framework for the er computational creativity model. In *MICAI 2006: Advances in Artificial Intelligence*, pages 70–80, Apizaco, Mexico, 2006. Springer.

Berys Gaut. The philosophy of creativity. *Philosophy Compass*, 5(12):1034–1046, 2010.

Pablo Gervás. An expert system for the composition of formal Spanish poetry. *Knowledge-Based Systems*, 14(3):181–188, 2001.

Pablo Gervás. Exploring quantitative evaluations of the creativity of automatic poets. In *Workshop on Creative Systems, Approaches to Creativity in Artificial Intelligence and Cognitive Science, 15th European Conference on Artificial Intelligence*, Lyon, France, 2002. European Association for Artificial Intelligence.

Pablo Gervás. Engineering linguistic creativity: Bird flight and jet planes. In *Proceedings of the NAACL HLT 2010 Second Workshop on Computational Approaches to Linguistic Creativity*, pages 23–30, Los Angeles, US, 2010. Association for Computational Linguistics.

Pablo Gervás. Computational modelling of poetry generation. In *Artificial Intelligence and Poetry Symposium, AISB Convention*, Exeter University, UK, 2013a. Society for the Study of Artificial Intelligence and the Simulation of Behaviour.

Pablo Gervás. Evolutionary elaboration of daily news as a poetic stanza. In *Proceedings of the IX Congreso Español de Metaheurísticas, Algoritmos Evolutivos y Bioinspirados-MAEB*, pages 229–238, 2013b.

Pablo Gervás. Constrained creation of poetic forms during theme-driven exploration of a domain defined by an n-gram model. *Connection Science*, 28(2):111–130, 2016.

Pablo Gervás. Comparative evaluation of elementary plot generation procedures. In *Proceedings of the 6th International Workshop on Computational Creativity, Concept Invention, and General Intelligence*, 2017.

Pablo Gervás. Template-free construction of rhyming poems with thematic cohesion. In *Proceedings of the Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2017)*, pages 21–28, 2017.

Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. Generating topical poetry. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1191. Association for Computational Linguistics, 2016.

Andrei Gheorghe. The longest poem in the world, 2013. `http://www.longestpoemintheworld.com/`, accessed May 2018.

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. Affect-LM: A neural language model for customizable affective text generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 634–642, 2017.

Vlad Petre Glăveanu. Creativity as a sociocultural act. *The Journal of Creative Behavior*, 49(3):165–180, 2015.

Loss Pequeo Glazier. E-poetry: An international digital poetry festival, 2016. `http://epc.buffalo.edu/e-poetry/archive/`, accessed February 25, 2016.

Fernand Gobet and Herbert A Simon. Expert chess memory: revisiting the chunking hypothesis. *Memory*, 6(3):225–255, 1998.

Kenneth Goldsmith. Flarf is Dionysus. Conceptual Writing is Apollo. *Poetry*, July/August 2009.

Hugo Gonçalo Oliveira. Tra-la-lyrics 2.0: Automatic generation of song lyrics on a semantic domain. *Journal of Artificial General Intelligence*, 6(1):87–110, 2015.

João Gonçalves, Pedro Martins, and Amílcar Cardoso. Blend City, BlendVille. In *Proceedings of the Eighth International Conference on Computational Creativity*, pages 112–119, Atlanta, US, 2017. Association for Computational Creativity.

Ross Goodwin. Adventures in narrated reality. In *Medium*, 2016. Retrieved November 29, 2016.

Nada Gordon. Unicorn believers don't declare fatwas. *Poetry*, July/August 2009.

Kazjon Grace, Mary Lou Maher, Maryam Mohseni, and Rafael Pérez y Pérez. Encouraging p-creative behaviour with computational curiosity. In *Proceedings of the Eighth International Conference on Computational Creativity*, pages 120–127, Atlanta, US, 2017. Association for Computational Creativity.

Oskar Gross, Hannu Toivonen, Jukka M Toivanen, and Alessandro Valitutti. Lexical creativity from word associations. In *Proceedings of the Seventh International Conference on Knowledge, Information and Creativity Support Systems*, pages 35–42, Melbourne, Australia, 2012. IEEE.

Christian Guckelsberger, Christophe Salge, and Simon Colton. Addressing the "why?" in computational creativity: a non-anthropocentric, minimal model of intentional creative agency. In *Proceedings of the Eighth International Conference on Computational Creativity*, pages 128–135, Atlanta, US, 2017. Association for Computational Creativity.

Brion Gysin et al. Cut-ups: A project for disastrous success. In *The third mind*, pages 42–51. 1978.

Mika Hämäläinen. Harnessing NLG to create Finnish poetry automatically. In *Proceedings of the Ninth International Conference on Computational Creativity*. Association for Computational Creativity, 2018.

Otto Hantula and Simo Linkola. Towards goal-aware collaboration in artistic agent societies. In *Proceedings of the Ninth International Conference on Computational Creativity*. Association for Computational Creativity, 2018.

Sarah Harmon. Figure8: a novel system for generating and evaluating figurative language. In *Proceedings of the Sixth International Conference on Computational Creativity*, pages 71–77, Park City, US, 2015. Association for Computational Creativity.

David M Harrington. On the usefulness of "value" in the definition of creativity: A commentary. *Creativity Research Journal*, 30(1):118–121, 2018.

Jacob Harris. Times haiku: serendipitous poetry from the New York Times, 2013. `http://haiku.nytimes.com/`, accessed December 9, 2015.

Randy Harris and Chrysanne DiMarco. Constructing a rhetorical figuration ontology. In *Persuasive Technology and Digital Behaviour Intervention Symposium*, pages 47–52, 2009.

Eliška Hartlová and F Nack. Mobile social poetry with tweets. *Bachelor thesis, University of Amsterdam*, 2013.

Maximilian Droog Hayes and Geraint A Wiggins. Adding semantics to statistical generation for poetic creativity. In *Sixth International Conference on Computational Creativity*. Association for Computational Creativity, 2015. Late-breaking abstract.

Jing He, Ming Zhou, and Long Jiang. Generating Chinese classical poems with statistical machine translation models. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. AAAI Publications, 2012.

Eric Heep and Ajay Kapur. Extracting visual information to generate sonic art installation and performance. In *Proceedings of the 21st International Symposium on Electronic Art*, 2015.

Paul Hekkert, Dirk Snelders, and Piet CW Wieringen. "Most advanced, yet acceptable": typicality and novelty as joint predictors of aesthetic preference in industrial design. *British Journal of Psychology*, 94 (1):111–124, 2003.

Jiun-Jhy Her. An analytical framework for facilitating interactivity between participants and interactive artwork: case studies in MRT stations. *Digital Creativity*, 25(2):113–125, 2014.

Hussein Hirjee. Rhyme, rhythm, and rhubarb: Using probabilistic methods to analyze hip hop, poetry, and misheard lyrics. 2010.

Hussein Hirjee and Daniel G Brown. Using automated rhyme detection to characterize rhyming style in rap music. *Empirical Musicology Review*, 2010.

I Infantino, A Augello, A Manfré, G Pilato, and F Vella. Robodanza: live performances of a creative dancing humanoid. In *Proceedings of the Seventh International Conference on Computational Creativity*, pages 388–395, Paris, France, 2016. Association for Computational Creativity.

Arthur M Jacobs. Towards a neurocognitive poetics model of literary reading. In *Cognitive Neuroscience of Natural Language Use*, pages 135–159. Cambridge University Press, Cambridge, UK, 2015.

DA Johner, D Bedwell, C Graham, W Lemmon, O Martinez, and AK Goel. Using human computation to acquire novel methods for addressing visual analogy problems on intelligence tests. In *Proceedings of the Sixth International Conference on Computational Creativity*, pages 23–30, Park City, US, 2015. Association for Computational Creativity.

Anna Jordanous. A standardised procedure for evaluating creative systems: computational creativity evaluation based on what it is to be creative. *Cognitive Computation*, 4(3):246–279, 2012a.

Anna Jordanous. Stepping back to progress forwards: setting standards for meta-evaluation of computational creativity. In *Proceedings of the Fifth International Conference on Computational Creativity*, pages 129–136, Ljubljana, Slovenia, 2014. Association for Computational creativity.

Anna Jordanous. Four pppperspectives on computational creativity in theory and in practice. *Connection Science*, 28(2):194–216, 2016a.

Anna Jordanous. The longer term value of creativity judgements in computational creativity. In *Proceedings of the AISB Symposium on Computational Creativity*, pages 16–23, Sheffield, UK, 2016b. AISB.

Anna Jordanous. Personal communication on Twitter, 2016c. `http://twitter.com/annajordanous/status/747766290648080384`.

Anna Jordanous. Creativity vs quality: why the distinction matters when evaluating computational creativity systems. In *The 5th Computational Creativity Symposium at the AISB Convention*, 2018.

Anna Jordanous, Daniel Allington, and Byron Dueck. Measuring cultural value using social network analysis: a case study on valuing electronic musicians. In *Proceedings of the Sixth International Conference on Computational Creativity*, pages 110–117, Park City, US, 2015. Association for Computational Creativity.

Anna Katerina Jordanous. *Evaluating computational creativity: a standardised procedure for evaluating creative systems and its application*. PhD thesis, University of Sussex, 2012b.

Patrik N Juslin, Laura S Sakka, Gonçalo T Barradas, and Simon Liljeström. No accounting for taste? idiographic models of aesthetic judgment in music. *Psychology of Aesthetics, Creativity, and the Arts*, 10(2):157, 2016.

Stefano Kalonaris. Satisficing goals and methods in human-machine music improvisations: Experiments with Dory. *Journal of Creative Music Systems*, 2(1):21, 2018.

Anna Kantosalo and Hannu Toivonen. Modes for creative human-computer collaboration: alternating and task-divided co-creativity. In *Proceedings of the Seventh International Conference on Computational Creativity*, pages 77–84, Paris, France, 2016. Association for Computational Creativity.

Anna Kantosalo, Jukka M Toivanen, Ping Xiao, and Hannu Toivonen. From isolation to involvement: Adapting machine creativity software to support human-computer co-creation. In *Proceedings of the Fifth International Conference on Computational Creativity*, pages 1–7. Association for Computational Creativity, 2014.

Anna Aurora Kantosalo, Jukka Mikael Toivanen, Hannu Tauno Tapani Toivonen, et al. Interaction evaluation for human-computer co-creativity: a case study. In *Proceedings of the Sixth International Conference on Computational Creativity*, pages 276–283, Park City, US, 2015. Association for Computational Creativity.

Justine Kao and Dan Jurafsky. A computational analysis of style, affect, and imagery in contemporary poetry. In *NAACL Workshop on Computational Linguistics for Literature*, pages 8–17, 2012.

Justine T Kao and Dan Jurafsky. A computational analysis of poetic style. *LiLT (Linguistic Issues in Language Technology)*, 12, 2015.

David M Kaplan and David M Blei. A computational approach to style in American poetry. In *Seventh IEEE International Conference on Data Mining*, pages 553–558, 2007.

Pythagoras Karampiperis, Antonis Koukourikos, and Evangelia Koliopoulou. Towards machines for measuring creativity: the use of computational tools in storytelling activities. In *Proceedings of the 14th International Conference on Advanced Learning Technologies*, pages 508–512, Athens, Greece, 2014. IEEE.

Andrej Karpathy. The unreasonable effectiveness of recurrent neural networks, 2015. qtd. in (Goodwin, 2016).

James C. Kaufman. Counting the muses: Development of the Kaufman Domains of Creativity Scale (K-DOCS). *Psychology of Aesthetics, Creativity, and the Arts*, 6(4):298–308, 2012.

James C Kaufman and John Baer. Beyond new and appropriate: who decides what is creative? *Creativity Research Journal*, 24(1):83–91, 2012.

James C Kaufman and Ronald A Beghetto. Beyond big and little: the four c model of creativity. *Review of General Psychology*, 13(1):1, 2009.

James C Kaufman, John Baer, Jason C Cole, and Janel D Sexton. A comparison of expert and nonexpert raters using the consensual assessment technique. *Creativity Research Journal*, 20(2):171–178, 2008.

James C Kaufman, John Baer, and Jason C Cole. Expertise, domains, and the consensual assessment technique. *The Journal of Creative Behavior*, 43(4):223–233, 2009.

Kyungil Kim, Jinhee Bae, Myung-Woo Nho, and Chang Hwan Lee. How do experts and novices differ? relation versus attribute and thinking versus feeling in language use. *Psychology of Aesthetics, Creativity, and the Arts*, 5(4):379, 2011.

Alexis Kirke and Eduardo Miranda. Emotional and multi-agent systems in computer-aided writing and poetry. In *Proceedings of the Artificial Intelligence and Poetry Symposium*, Exeter, UK, 2013. AISB.

Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. The future of crowd work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 1301–1318, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1331-5. doi: 10.1145/2441776.2441923. URL `http://doi.acm.org/10.1145/2441776.2441923`.

Alison Knowles and James Tenney. A sheet from 'The House', a computer poem, 1968. qtd. in (Funkhouser, 2007).

Maria E Kronfeldner. Darwinian "blind" hypothesis formation revisited. *Synthese*, 175(2):193–218, 2010.

Ray Kurzweil. Ray Kurzweil's cybernetic poet, 2001. `http://www.kurzweilcyberart.com/poetry`.

Carolyn Lamb, Daniel G Brown, and Charles LA Clarke. Human competence in creativity evaluation. In *Proceedings of the Sixth International Conference on Computational Creativity*, pages 102–109, Park City, US, 2015a. Association for Computational Creativity.

Carolyn Lamb, Daniel G Brown, and Charles LA Clarke. Evaluating digital poetry: Insights from the CAT. In *Proceedings of the Seventh International Conference on Computational Creativity*. Association for Computational Creativity, 2016a.

Carolyn Lamb, Daniel G Brown, and Charles LA Clarke. A taxonomy of generative poetry techniques. In *Proceedings of Bridges 2016: Mathematics, Music, Art, Architecture, Culture*, 2016b.

Carolyn Lamb, Daniel Brown, and Charles Clarke. Incorporating novelty, meaning, reaction and craft into computational poetry: a negative experimental result. In *Proceedings of the Eighth International Conference on Computational Creativity*. Association for Computational Creativity, 2017a.

Carolyn Lamb, Daniel G Brown, and Charles LA Clarke. A taxonomy of generative poetry techniques. *Journal of Mathematics and the Arts*, 11(3):159–179, 2017b.

Carolyn Lamb, Daniel G Brown, and Charles LA Clarke. Evaluating computational creativity: An interdisciplinary tutorial. *ACM Computing Surveys (CSUR)*, 51(2):28, 2018.

Carolyn E Lamb, Daniel G Brown, and Charles LA Clarke. Can human assistance improve a computational poet? *Proceedings of Bridges 2015: Mathematics, Music, Art, Architecture, Culture*, pages 37–44, 2015b.

Stefan Lattner, Maarten Grachten, and Gerhard Widmer. Imposing higher-level structure in polyphonic music generation using convolutional restricted Boltzmann machines and constraints. *Journal of Creative Music Systems*, 2:31, 2018.

Jon O Lauring, Matthew Pelowski, Michael Forster, Matthias Gondan, Maurice Ptito, and Ron Kupers. Well, if they like it... effects of social groups' ratings and price information on the appreciation of art. *Psychology of Aesthetics, Creativity, and the Arts*, 10(3):344, 2016.

Helmut Leder and Marcos Nadal. Ten years of a model of aesthetic appreciation and aesthetic judgments: the aesthetic episode–developments and challenges in empirical aesthetics. *British Journal of Psychology*, 105(4):443–464, 2014.

Helmut Leder, Benno Belke, Andries Oeberst, and Dorothee Augustin. A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology*, 95(4):489–508, 2004.

Helmut Leder, Gernot Gerger, David Brieber, and Norbert Schwarz. What makes an art expert? Emotion and evaluation in art appreciation. *Cognition and Emotion*, 28(6):1137–1147, 2014.

John Lee, Ying Cheuk Hui, and Yin Hei Kong. Knowledge-rich, computer-assisted composition of Chinese couplets. *Digital Scholarship in the Humanities*, 31(1):152–163, 2016.

Joel Lehman and Kenneth O Stanley. Beyond open-endedness: quantifying impressiveness. In *Artificial Life*, volume 13, pages 75–82, Cambridge, US, 2012. MIT Press.

Andrea Lingentfelter. Personal communication, June 2017.

Simo Linkola, Tapio Takala, and Hannu Toivonen. Novelty-seeking multi-agent systems. In *Proceedings of the Seventh International Conference on Computational Creativity*, pages 1–8, Paris, France, 2016. Association for Computational Creativity.

Simo Linkola, Anna Kantosalo, Tomi Männistö, and Hannu Toivonen. Aspects of self-awareness: an anatomy of metacreative systems. In *Proceedings of the Eighth International conference on Computational Creativity*, pages 189–196, Atlanta, US, 2017. Association for Computational Creativity.

Maria Teresa Llano, Rose Hepworth, Simon Colton, Jeremy Gow, John Charnley, N Lavrac, M Znidaršic, Matic Perovšek, Mark Granroth-Wilding, and Stephen Clark. Baseline methods for automated fictional ideation. In *Proceedings of the Fifth International Conference on Computational Creativity*, pages 211–219, Ljubljana, Slovenia, 2014. Association for Computational Creativity.

Malte Loller-Andersen and Björn Gambäck. Deep learning-based poetry generation given visual input. In *Proceedings of the Ninth International Conference on Computational Creativity*. Association for Computational Creativity, 2018.

Róisın Loughran and Michael ONeill. Generative music evaluation: why do we limit to human? In *Proceedings of the First Conference on Computer Simulation of Musical Creativity*, Huddersfield, UK, 2016. University of Huddersfield.

Róisın Loughran and Michael ONeill. Application domains considered in computational creativity. In *Proceedings of the Eighth International Conference on Computational Creativity*, Atlanta, US, 2017. Association for Computational Creativity.

Todd I Lubart. Creativity across cultures. In *Handbook of Creativity*, pages 339–350. Cambridge University Press, Cambridge, UK, 1999.

Theo Lutz. Stochastiche texte. *Augenblick*, 4(1):3–9, 1959. qtd. in (Roque, 2011); translated to English by Helen MacCormack, 2005.

Mary-Anne Mace and Tony Ward. Modeling the creative process: A grounded theory analysis of creativity in the domain of art making. *Creativity Research Journal*, 14(2):179–192, 2002.

Eric Malmi, Pyry Takala, Hannu Toivonen, Tapani Raiko, and Aristides Gionis. DopeLearning: a computational approach to rap lyrics generation. *arXiv preprint arXiv:1505.04771*, 2015.

Hisar Manurung, Graeme Ritchie, and Henry Thompson. Towards a computational model of poetry generation. Technical report, The University of Edinburgh, 2000.

Hisar Maruli Manurung. Chart generation of rhythm-patterned text. In *Proceedings of the First International Workshop on Literature in Cognition and Computers*, pages 15–19, 1999.

Ruli Manurung, Graeme Ritchie, and Henry Thompson. Using genetic algorithms to create meaningful poetic text. *Journal of Experimental and Theoretical Artificial Intelligence*, 24(1):43–64, 2012.

Jon McCormack and Mark dInverno. On the future of computers and creativity. In *AISB Symposium on Computational Creativity*, pages 1–4, London, UK, 2014. Society for the Study of Artificial Intelligence and the Simulation of Behaviour.

Kyle McDonald. Neural nets for generating music. In *Artists and Machine Intelligence*, 2017. `https://medium.com/artists-and-machine-intelligence/neural-nets-for-generating-music-f46dffac21c0` Accessed April 2018.

Kenric McDowell. Ai poetry hits the road. In *Artists and Machine Intelligence*, 2017. `https://medium.com/artists-and-machine-intelligence/ai-poetry-hits-the-road-eb685dfc1544` Accessed April 2018.

Kenrick McDowell. Music, art & machine intelligence 2016 conference proceedings. In *Medium*, 2016. Retrieved December 2, 2016.

Jeffrey D McGovern and Gavin Scott. EloquentRobot: A tool for automatic poetry generation. In *Proceedings of the Seventh ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2016.

Stephen McGregor, Matthew Purver, and Geraint Wiggins. Process based evaluation of computer generated poetry. In *The INLG 2016 Workshop on Computational Creativity in Natural Language Generation*, page 51, Edinburgh, UK, 2016. ACM SIGGEN.

Alexander S. McKay, Maciej Karwowski, and James C. Kaufman. Measuring the muses: Validating the Kaufman Domains of Creativity Scale (K-DOCS). *Psychology of Aesthetics, Creativity, and the Arts*, 11(2):216–230, 2017.

Lewis Mckeown and Anna Jordanous. An evaluation of the impact of constraints on the perceived creativity of narrative generating software. In *Proceedings of the Ninth International Conference on Computational Creativity*. Association for Computational Creativity, 2018.

Marvin L Minsky. Why people think computers can't. *AI Magazine*, 3(4):3, 1982.

Joanna Misztal and Bipin Indurkhya. Poetry generation system with an emotional personality. In *Proceedings of the Fifth International Conference on Computational Creativity*, pages 72–81, 2014.

D. Moffat and M. Kelly. An investigation into people's bias against computational creativity in music composition. In *The Third Joint Workshop on Computational Creativity*, ECAI 2006, Trento, Italy, aug 2006.

René Mogensen. Evaluating an improvising computer-implementation as a partial creativity in a music performance system. *Journal of Creative Music Systems*, 2, 2017.

Saif M Mohammad and Svetlana Kiritchenko. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326, 2015. Full lexicon available at `http://saifmohammad.com/WebPages/lexicons.html`.

Kristine Monteith, Bruce Brown, Dan Ventura, and Tony Martinez. Automatic generation of music for inducing emotive response. In *Proceedings of the First International Conference on Computational Creativity*, pages 140–149, Lisbon, Portugal, 2010. Association for Computational Creativity.

Nick Montfort and Stephanie Strickland. Sea and spar between. *Dear Navigator*, 2, 2010.

Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: going deeper into neural networks. *Google Research Blog*, 20, 2015.

John Morris. Haiku—at random, 1973. qtd. in (Funkhouser, 2007).

Martin Mumford and Dan Ventura. The man behind the curtain: overcoming skepticism about creative computing. In *Proceedings of the Sixth International Conference on Computational Creativity*, pages 1–7, Park City, US, 2015. Association for Computational Creativity.

Frieder Nake. Construction and intuition: creativity in early computer art. In *Computers and Creativity*, pages 61–94. Springer, Berlin, Germany, 2012.

Santiago Negrete-Yankelevich and Nora Morales-Zaragoza. The apprentice framework: planning, assessing creativity. In *Proceedings of the Fifth International Conference on Computational Creativity*, pages 280–283, Ljubljana, Slovenia, 2014. Association for Computational creativity.

Yael Netzer, David Gabay, Yoav Goldberg, and Michael Elhadad. Gaiku: generating haiku with word associations norms. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, pages 32–39. Association for Computational Linguistics, 2009.

David Norton, Derrall Heath, and Dan Ventura. Establishing appreciation in a creative system. In *Proceedings of the First International Conference on Computational Creativity*, pages 26–35, Lisbon, Portugal, 2010. Association for Computational Creativity.

David Norton, Derrall Heath, and Dan Ventura. An artistic dialogue with the artificial. In *Proceedings of the 8th ACM Conference on Creativity and Cognition*, pages 31–40, Atlanta, US, 2011. ACM.

David Norton, Derrall Heath, and Dan Ventura. Finding creativity in an artificial artist. *The Journal of Creative Behavior*, 47(2):106–124, 2013.

David Norton, Derrall Heath, and Dan Ventura. Accounting for bias in the evaluation of creative computational systems: an assessment of DARCI. In *Proceedings of the Sixth International Conference on Computational Creativity*, pages 31–38, Park City, US, June–July 2015. Association for Computational Creativity.

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

Hugo Gonçalo Oliveira. O Poeta Artificial 2.0: Increasing meaningfulness in a poetry generation twitter bot. In *Proceedings of the Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2017)*, pages 11–20, 2017a.

Hugo Gonçalo Oliveira. A survey on intelligent poetry generation: Languages, features, techniques, reutilisation and evaluation. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 11–20, 2017b.

Hugo Gonçalo Oliveira and Ana Oliveira Alves. Poetry from concept maps–yet another adaptation of PoeTryMe's flexible architecture. In *Proceedings of 7th International Conference on Computational Creativity, ICCC*, 2016.

Ana-Maria Olteţeanu and Zoe Falomir. comRAT-C: A computational compound remote associates test solver based on language data and its comparison to human performance. *Pattern Recognition Letters*, 67:81–90, 2015.

Ana-Maria Olteţeanu and Zoe Falomir. Object replacement and object composition in a creative cognitive system: towards a computational solver of the alternative uses test. *Cognitive Systems Research*, 39: 15–32, 2016.

Sarah Opolka, Philipp Obermeier, and Torsten Schaub. Automatic genre-dependent composition using answer set programming. In *Proceedings of the 21st International Symposium on Electronic Art*, 2015.

François Pachet and Pierre Roy. Markov constraints: steerable generation of Markov sequences. *Constraints*, 16(2):148–172, 2011.

François Pachet, Pierre Roy, Julian Moreira, and Mark d'Inverno. Reflexive loopers for solo musical improvisation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2205–2208. ACM, 2013.

Raquel Mochales Palau and Marie-Francine Moens. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 98–107. ACM, 2009.

Alexandre Papadopoulos, Pierre Roy, and François Pachet. Assisted lead sheet composition using flowcomposer. In *International Conference on Principles and Practice of Constraint Programming*, pages 769–785. Springer, 2016.

Philippe Pasquier, Adam Burnett, and James Maxwell. Investigating listener bias against musical metacreativity. In *Proceedings of the Seventh International Conference on Computational Creativity*, pages 42–51, Paris, France, 2016. Association for Computational Creativity.

Marcus Pearce and Geraint Wiggins. Towards a framework for the evaluation of machine compositions. In *Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in the Arts and Sciences*, pages 22–32, York, UK, 2001. Society for the Study of Artificial Intelligence and the Simulation of Behaviour.

Marcus Pearce, David Meredith, and Geraint Wiggins. Motivations and methodologies for automation of the compositional process. *Musicae Scientiae*, 6(2):119–147, 2002.

Marcus T Pearce and Geraint A Wiggins. Evaluating cognitive models of musical composition. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, pages 73–80, 2007.

Alison Pease and Simon Colton. On impact and evaluation in computational creativity: A discussion of the Turing test and an alternative proposal. In *Proceedings of the AISB symposium on AI and Philosophy*, pages 15–22, York, UK, 2011. Society for the Study of Artificial Intelligence and the Simulation of Behaviour.

Alison Pease, Daniel Winterstein, and Simon Colton. Evaluating machine creativity. In *Workshop on Creative Systems, 4th International Conference on Case Based Reasoning*, pages 129–137, Vancouver, Canada, 2001. ACM.

James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

Graham Percival, Satoru Fukayama, and Masataka Goto. Song2Quartet: a system for generating string quartet cover songs from polyphonic audio of popular music. In *Proceedings of the 16th ISMIR conference*, pages 114–120. The International Society of Music Information Retrieval, 2015.

Francisco C Pereira, Mateus Mendes, P Gervás, and Amılcar Cardoso. Experiments with assessment of creative systems: an application of Ritchie's criteria. In *Proceedings of the workshop on computational creativity, 19th international joint conference on artificial intelligence*, page 05, Edinburgh, UK, 2005. ACM.

Rafael Pérez y Pérez. The computational creativity continuum. In *Proceedings of the Ninth International Conference on Computational Creativity*. Association for Computational Creativity, 2018.

Jonathan A Plucker and Joseph S Renzulli. Psychometric approaches to the study of human creativity. In *Handbook of Creativity*, pages 35–61. Cambridge University Press, Cambridge, UK, 1999.

Poetry Foundation. Poetry magazine discussion guide. `http://www.poetryfoundation.org/poetrymagazine/guide/89`, February 2014. Accessed: 2015-02-03.

Emma Policastro and Howard Gardner. From case studies to robust generalizations: an approach to the study of creativity. In *Handbook of Creativity*, pages 213–225. Cambridge University Press, Cambridge, UK, 1999.

Peter Potash, Alexey Romanov, and Anna Rumshisky. Evaluating creative language generation: The case of rap lyric ghostwriting. *arXiv preprint arXiv:1612.03205*, 2016.

Provalis. Regressive imagery dictionary. https://provalisresearch.com/products/content-analysis-software/wordstat-dictionary/regressive-imagery-dictionary/, 1990.

Oscar Puerto and David Thue. A model of inter-musician communication for artificial musical intelligence. In *Proceedings of the Eighth International Conference on Computational Creativity*, pages 221–228, Atlanta, US, 2017. Association for Computational Creativity.

Fahrurrozi Rahman and Ruli Manurung. Multiobjective optimization for meaningful metrical poetry. In *Proceedings of the Second International Conference on Computational Creativity*, pages 4–9, 2011.

Ananth Ramakrishnan A and Sobha Lalitha Devi. An alternate approach towards meaningful lyric generation in Tamil. In *Proceedings of the NAACL HLT 2010 Second Workshop on Computational Approaches to Linguistic Creativity*, pages 31–39. Association for Computational Linguistics, 2010.

Fam Rashel and Ruli Manurung. Pemuisi: a constraint satisfaction-based generator of topical Indonesian poetry. In *Proceedings of the Fifth International Conference on Computational Creativity*, pages 82–90, Ljubljana, Slovenia, 2014a. Association for Computational Creativity.

Fam Rashel and Ruli Manurung. Pemuisi: a constraint satisfaction-based generator of topical Indonesian poetry. In *Proceedings of the Fifth International Conference on Computational Creativity*, pages 82–90, 2014b.

Mel Rhodes. An analysis of creativity. *Phi Delta Kappan*, 42(7):305–310, 1961.

Mark O Riedl and R Michael Young. Story planning as exploratory creativity: techniques for expanding the narrative search space. *New Generation Computing*, 24(3):303–323, 2006.

Graeme Ritchie. Assessing creativity. In *Proc. of AISB01 Symposium*, 2001.

Graeme Ritchie. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines*, 17(1):67–99, 2007.

Iván Guerrero Román and Rafael Pérez y Pérez. Social Mexica: a computer model for social norms in narratives. In *Proceedings of the Fifth International Conference on Computational Creativity*, pages 192–200, Ljubljana, Slovenia, 2014. Association for Computational Creativity.

Antonio Roque. Language technology enables a poetics of interactive generation. *Journal of Electronic Publishing*, 14(2), 2011.

Kimiko Ryokai, Noriko Misra, and Yoshinori Hara. Artistic distance: body movements as launching points for art inquiry. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 679–686, Seoul, Korea, 2015. ACM.

Eugene Sadler-Smith. Wallas' four-stage model of the creative process: more than meets the eye? *Creativity Research Journal*, 27(4):342–352, 2015.

Steven L Salzberg, Mihaela Pertea, Arthur L Delcher, Malcolm J Gardner, and Hervé Tettelin. Interpolated markov models for eukaryotic gene finding. *Genomics*, 59(1):24–31, 1999.

Jonathan Sammartino and Stephen E Palmer. Aesthetic issues in spatial composition: representational fit and the role of semantic context. *Perception*, 41(12):1434, 2012.

Rob Saunders, Petra Gemeinboeck, Adrian Lombard, Dan Bourke, and A Baki Kocaballi. Curious whispers: an embodied artificial creative system. In *Proceedings of the First International Conference on Computational Creativity*, pages 100–109, Lisbon, Portugal, 2010. Association for Computational Creativity.

Udo Schlegel, Eren Cakmak, Juri Buchmller, and Daniel A. Keim. G-rap: interactive text synthesis using recurrent neural network suggestions. In *ESANN 2018 proceedings*, 2018.

Oscar Schwartz and Benjamin Laird. bot or not. `http://botpoet.com/`, 2015. Accessed: 2015-09-08.

Marco Scirea, Peter Eklund, and Julian Togelius. Toward a context sensitive music generator for affective state expression. In *Proceedings of the Sixth International Conference on Computational Creativity*. Brigham Young University, 2015. Late-breaking abstract.

John R Searle. Minds, brains, and programs. *Behavioral and brain sciences*, 3(03):417–424, 1980.

Jennifer G Sheridan, Alan Dix, Simon Lock, and Alice Bayliss. Understanding interaction in ubiquitous guerrilla performances in playful arenas. In *People and Computers XVIII—Design for Life: Proceedings of HCI 2004*, pages 3–17. Springer, Bornemouth, UK, 2005.

Hirohito Shibata and Koichi Hori. A system to support long-term creative thinking in daily life and its evaluation. In *Proceedings of the Fourth Conference on Creativity & Cognition*, pages 142–149, Loughborough, UK, 2002. ACM.

Disha Shrivastava, Anirban Laha, and Karthik Sankaranarayanan. A machine learning approach for evaluating creative artifacts. *arXiv preprint arXiv:1707.05499*, 2017.

Paul J Silvia. Emotional responses to art: From collation and arousal to cognition and emotion. *Review of general psychology*, 9(4):342, 2005.

Dean Keith Simonton. Lexical choices and aesthetic success: A computer content analysis of 154 Shakespeare sonnets. *Computers and the Humanities*, 24(4):251–264, 1990.

Dean Keith Simonton. Creativity and discovery as blind variation: Campbell's (1960) BVSR model after the half-century mark. *Review of General Psychology*, 15(2):158, 2011.

Divya Singh, Margareta Ackerman, and Rafael Pérez y Pérez. A ballad of the Mexicas: Automated lyrical narrative writing. In *Eighth International Conference on Computational Creativity*, 2017.

Jeffrey K. Smith and Lisa F. Smith. The 1.5 criterion model of creativity: Where less is more, more or less. *Journal of Creative Behavior*, 51, 2017.

Michael R Smith, Ryan S Hintze, and Dan Ventura. Nehovah: a neologism creator nomen ipsum. In *Proceedings of the Fifth International Conference on Computational Creativity*, pages 173–181, Ljubljana, Slovenia, 2014. Association for Computational Creativity.

Tim Smithers. Autonomy in robots and other agents. *Brain and Cognition*, 34(1):88–106, 1997.

Von-Wun Soo, Tung-Yi Lai, Kai-Ju Wu, and Yu-Po Hsu. Generate modern style Chinese poems based on common sense and evolutionary computation. In *2015 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pages 315–322. IEEE, 2015.

Robert J Sternberg. Whence creativity? *The Journal of Creative Behavior*, 51(4):289–292, 2017.

Robert J. Sternberg. What's wrong with creativity testing? *Journal of Creative Behavior*, 0, 2018.

Robert J. Sternberg, James C. Kaufman, and Jean E. Pretz. The propulsion model of creative contributions applied to the arts and letters. *Journal of Creative Behavior*, 35(2):75–101, 2001.

Martin Storme, Todd Lubart, Nils Myszkowski, Ping Chung Cheung, Toby Tong, , and Sing Lau. A cross-cultural study of task specificity in creativity. *Journal of Creative Behavior*, 51(3), 2015.

Charlie Stross. Lovecraft.pl, December 2013. Blog post. `http://www.antipope.org/charlie/blog-static/2013/12/lovebiblepl.html`, accessed December 21, 2016.

Bob L. Sturm and Oded Ben-Tal. Taking the models back to music practice: Evaluating generative transcription models built using deep learning. *Journal of Creative Music Systems*, 2(1), 2017.

Luchen Tan, Haotian Zhang, Charles Clarke, and Mark Smucker. Lexical comparison between Wikipedia and Twitter corpora by using word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 657–661, 2015.

Luchen Tan, Adam Roegiest, Charles LA Clarke, and Jimmy Lin. Simple dynamic emission strategies for microblog filtering. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1009–1012. ACM, 2016.

A. Tapscott, J. Gómez, C. León, J. Smailović, M. Žnidaršič, and P. Gervás. Empirical evidence of the limits of automatic assessment of fictional ideation. In *Proceedings of the Fifth International Workshop on Computational Creativity, Concept Invention, and General Intelligence at ESSLLI*, pages 58–71, Bozen-Bolzano, Italy, 2016. Association for Logic, Language and Information.

Yla R Tausczik and James W Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

Brandon Tearse, Peter Mawhorter, Michael Mateas, and Noah Wardrip-Fruin. Experimental results from a rational reconstruction of minstrel. In *In Proceedings of the 2nd International Conference on Computational Creativity*, pages 54–59, Mexico City, Mexico, 2011. Association for Computational Creativity.

Berty C Tobing and Ruli Manurung. A chart generation system for topical metrical poetry. In *Proceedings of the Sixth International Conference on Computational Creativity June*, pages 308–314, 2015.

Jukka Toivanen, Hannu Toivonen, Alessandro Valitutti, Oskar Gross, et al. Corpus-based generation of content and form in poetry. In *Proceedings of the Third International Conference on Computational Creativity*, pages 211–215, Dublin, Ireland, 2012. Association for Computational Creativity.

Jukka M Toivanen, Matti Järvisalo, and Hannu Toivonen. Harnessing constraint programming for poetry composition. In *Proceedings of the Fourth International Conference on Computational Creativity*, pages 160–167. computationalcreativity.net, 2013.

Jukka M Toivanen, Oskar Gross, and Hannu Toivonen. The officer is taller than you, who race yourself! Using document specific word associations in poetry generation. In *Proceedings of the Fifth International Conference on Computational Creativity*, pages 355–359. Association for Computational Creativity, 2014.

Ellis Paul Torrance. *Torrance tests of creative thinking.* Personnel Press, Incorporated, Princeton, US, 1968.

Patrick Tresset and Oliver Deussen. Artistically skilled embodied agents. In *Proceedings of the AISB Symposium on Computational Creativity*, London, UK, 2014. Society for the Study of Artificial Intelligence and the Simulation of Behaviour.

University of Helsinki Computer Science Department. Works of artistic nature, 2016. `https://www.cs.helsinki.fi/discovery/works-artistic-nature`, accessed December 6, 2016.

Frank van der Velde, Roger A Wolf, Martin Schmettow, and Deniece S Nazareth. A semantic map for evaluating creativity. In *Proceedings of the Sixth International Conference on Computational Creativity June*, page 94, 2015.

Oshin Vartanian, Alenoush Vartanian, Roger E. Beaty, Emily C. Nusbaum, and Kristen Blackler. Revered today, loved tomorrow: Expert creativity ratings predict popularity of architects' works 50 years later. *Psychology of Aesthetics, Creativity, and the Arts*, 11:386–391, 2017.

Tony Veale. Less rhyme, more reason: Knowledge-based poetry generation with feeling, insight and wit. In *Proceedings of the Fourth International Conference on Computational Creativity*, pages 152–159, 2013a.

Tony Veale. Linguistic readymades and creative reuse. *Journal of Integrated Design and Process Science*, 17(4):37–51, 2013b.

Tony Veale. Game of tropes: exploring the placebo effect in computational creativity. In *Proceedings of the Sixth International Conference on Computational Creativity*, pages 78–85, Park City, US, 2015. Association for Computational Creativity.

Tony Veale and Yanfen Hao. Exploiting readymades in linguistic creativity: a system demonstration of the jigsaw bard. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations*, pages 14–19. Association for Computational Linguistics, 2011.

Dan Ventura. Mere generation: essential barometer or dated concept? In *Proceedings of the Seventh International Conference on Computational Creativity*, pages 17–24, Paris, France, 2016. Association for Computational Creativity.

Dan A Ventura. A reductio ad absurdum experiment in sufficiency for evaluating (computational) creative systems. In *Proceedings of the Fifth International Joint Workshop on Computational Creativity*, pages 11–19, Madrid, Spain, 2008. Association for Computational Creativity.

Tiago Gonzaga Videira, Bruce Pennycook, and Jorge Martins Rosa. Formalizing fado: A contribution to automatic song-making. *Journal of Creative Music Systems*, 1(2), 2017.

Graham Wallas. *The art of thought.* London: Jonathan Cape, 1926.

Annalu Waller, Rolf Black, David A OMara, Helen Pain, Graeme Ritchie, and Ruli Manurung. Evaluating the STANDUP pun generating software with children with cerebral palsy. *ACM Transactions on Accessible Computing (TACCESS)*, 1(3):16, 2009.

Qixin Wang, Tianyi Luo, Dong Wang, and Chao Xing. Chinese Song iambics generation with neural attention-based model. *arXiv preprint arXiv:1604.06274*, 2016a.

Zhe Wang, Wei He, Hua Wu, Haiyang Wu, Wei Li, Haifeng Wang, and Enhong Chen. Chinese poetry generation with planning based neural network. *arXiv preprint arXiv:1610.09889*, 2016b.

Thomas B Ward, Steven M Smith, and Ronald A Finke. Creative cognition. In *Handbook of creativity*, pages 189–212. Cambridge University Press, Cambridge, UK, 1999.

Robert W Weisberg. On the usefulness of value in the definition of creativity. *Creativity Research Journal*, 27(2):111–124, 2015.

Geraint A Wiggins. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems*, 19(7):449–458, 2006.

M Tsan Wong, A Hon Wai Chun, Qing Li, SY Chen, and Anping Xu. Automatic haiku generation using VSM. In *WSEAS International Conference. Proceedings. Mathematics and Computers in Science and Engineering*. World Scientific and Engineering Academy and Society, 2008.

Brandon Wood. Tweet haikus, 2013. `https://twitter.com/tweethaikuscom`, accessed May 2018.

Linli Xu, Liang Jiang, Chuan Qin, Zhe Wang, and Dongfang Du. How images inspire poems: Generating classical Chinese poetry from images with memory networks. *CoRR*, abs/1803.02994, 2018. URL `http://arxiv.org/abs/1803.02994`.

Rui Yan. i, poet: Automatic poetry composition through recurrent neural networks with iterative polishing schema. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI Press, 2016.

Rui Yan, Han Jiang, Mirella Lapata, Shou-De Lin, Xueqiang Lv, and Xiaoming Li. i, poet: Automatic chinese poetry composition through a generative summarization framework under constrained optimization. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*. AAAI Press, 2013.

Hui Yang, Ian Soboroff, Li Xiong, Charles LA Clarke, and Simson L Garfinkel. Privacy-preserving IR 2016: Differential privacy, search, and social media. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1247–1248. ACM, 2016.

Xiaopeng Yang, Xiaowen Lin, Shunda Suo, and Ming Li. Generating thematic Chinese poetry using conditional variational autoencoders with hybrid decoders. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI-ECAI*, 2018.

Xiaoyuan Yi, Ruoyu Li, and Maosong Sun. Generating Chinese classical poems with RNN encoder-decoder. *arXiv preprint arXiv:1604.01537*, 2016.

Victor Zappi and Andrew McPherson. The D-Box: how to rethink a digital musical instrument. In *Proceedings of the 21st International Symposium on Electronic Art*, 2015.

Mo H Zareei, Dale A Carnegie, and Ajay Kapur. Noise square: physical sonification of cellular automata through mechatronic sound-sculpture. In *Proceedings of the 21st International Symposium on Electronic Art*, 2015.

Xingxing Zhang and Mirella Lapata. Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 670–680. Association for Computational Linguistics, 2014.