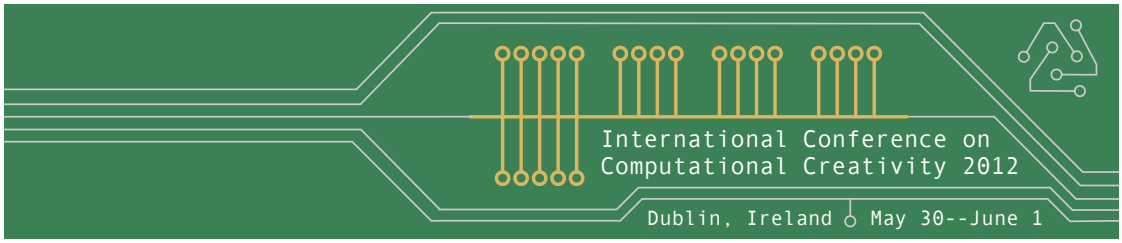


Proceedings of the Third International Conference on

Computational Creativity

edited by
Mary Lou Maher, Kristian Hammond, Alison Pease, Rafael Pérez y Pérez, Dan Ventura and Geraint Wiggins

Dublin, Ireland

May 2012

University College Dublin
Dublin, Ireland

http://computationalcreativity.net/iccc2012

About the logo: this year's conference logo references three aspects of Irish history and culture: The golden traces on the circuit board spell "ICCC" in Ogham, an Old Irish alphabet. The upper-right corner includes a Celtic knot, the three points of which emphasize the third meeting of the ICCC. The color palette of green, white, and gold echoes the Irish tricolor.

First published 2012

TITLE: PROCEEDINGS OF THE THIRD INTERNATIONAL CONFERENCE ON COMPUTATIONAL CREATIVITY

EDITORS: MARY LOU MAHER, KRISTIAN HAMMOND, ALISON PEASE, RAFAEL PÉREZ Y PÉREZ, DAN VENTURA, GERAINT WIGGINS

ISBN: 978-1-905254668

Preface

The Third International Conference on Computational Creativity 2012 represents a growth and maturity of a conference series that builds on a series of workshops held over ten years and the first two international conferences: the first held in Portugal in 2010 and the second held in Mexico in 2011. The purpose of this conference series is to make a scientific contribution to the field of computational creativity through discussion and publication on progress in fully autonomous creative systems, modeling human and computational creativity, computational support for human creativity, simulating creativity, and human/machine interaction in creative endeavors. Contributions come from many relevant disciplines, including computer science, artificial intelligence, engineering design, cognitive science, psychology, and art.

This year the conference received 59 paper submissions and 11 demonstration submissions. The peer review process for paper submissions has two steps: All paper submissions were reviewed by three members of the Program Committee and these reviews were further reviewed and compared by the Senior Program Committee. All demonstration submissions were reviewed by the Senior Program Committee. The committees accepted 34 papers from authors representing 18 countries: Australia, Austria, Canada, Finland, Germany, India, Indonesia, Ireland, Israel, Italy, Japan, Mexico, Poland, Singapore, Slovenia, Spain, UK, and USA.

In order to provide a snapshot of current progress in computational creativity and a glimpse of next steps, the conference invites and encourages two kinds of paper submissions: regular papers addressing foundational issues, describing original research on creative systems development and modeling, and position papers describing work-in-progress or research directions for computational creativity. The conference includes a balance of the two: 18 regular papers and 16 position papers. As in previous years, the conference also includes demonstrations in which conference attendees can play with specific implementations of computational creativity. The conference is organized into sessions that reflect the topics of interest this year: analogy and conceptual blending, creativity and search, generative systems, evaluating computational creativity, cognition and creativity, and language and creativity.

The collection of papers in this conference proceedings shows a maturity in the field through new examples of computational creativity and theoretical advances in understanding generative systems and evaluation of computational creativity. The conference series demonstrates success as we see publications that build on the advances of previous years through references to papers published in this conference series. We look forward to this publication providing the foundation for future developments in computational creativity.

Mary Lou, Kris, Alison, Rafael, Dan and Geraint

May 2012

Conference Chairs
General Chair: Pablo Gervás, Universidad Complutense de Madrid
Local Chair: Tony Veale, University College Dublin
Program Chair: Mary Lou Maher, University of Maryland
Publicity Chair: Kyle Jennings, University of California, Berkeley

Local Organizing Committee
Tony Veale, University College Dublin
Yanfen Hao, University College Dublin
Alejandra Lopez Fernandez, University College Dublin

Senior Program Committee
Kristian Hammond, Northwestern University
Alison Pease, University of Edinburgh
Rafael Pérez y Pérez, Autonomous Metropolitan University, México
Dan Ventura, Brigham Young University
Geraint Wiggins, Goldsmiths, University of London

Steering Committee
Amílcar Cardoso, University of Coimbra, Portugal
Simon Colton, Imperial College London, UK
Pablo Gervás, Universidad Complutense de Madrid, Spain
Nick Montfort, Massachusetts Institute of Technology
Alison Pease, University of Edinburgh, UK
Rafael Pérez y Pérez, Autonomous Metropolitan University, México
Graeme Ritchie, University of Aberdeen, UK
Rob Saunders, University of Sydney, Australia
Dan Ventura, Brigham Young University, USA
Tony Veale, University College, Dublin, Eire
Geraint A. Wiggins, Goldsmiths, University of London, UK

Program Committee
John Barnden, University of Birmingham
Oliver Bown, Monash University
David Brown, Worcester Polytechnic Institute
Nick Bryan-Kinns, Queen Mary, University of London
Win Burleson, Arizona State University
F. Amílcar Cardoso, Universidade de Coimbra
Kenny Chow, Hong Kong Polytechnic University
Simon Colton, Imperial College London
Roger Dannenberg, Carnegie Mellon University
Douglas Fisher, Vanderbilt University
John Gero, George Mason University
Ashok Goel, Georgia Institute of Technology
Paulo Gomes, Universidade de Coimbra
Andres Gomez de Silva, Instituto Tecnológico Autónomo de México
Kaz Grace, University of Sydney
Robert Keller, Harvey Mudd College
Henry Lieberman, Massachusetts Institute of Technology
Birte Loenneker-Rodman, Across Systems GmbH
Ramon López de Mántaras, IIIA-CSIC
Brian Magerko, Georgia Institute of Technology
Ruli Manurung, University of Indonesia
David C. Moffat, Glasgow Caledonian University

Contents

---

* RP = Regular Papers
* PP = Position Papers

From Conceptual "Mash-ups" to "Bad-ass" Blends:

A Robust Computational Model of Conceptual Blending

Tony Veale
School of Computer Science and Informatics
University College Dublin, Belfield D4, Ireland.
Tony.Veale@UCD.ie

Abstract

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

The Plumbing of Creative Thought

We can think of comparisons as pipes that carry salient information from a source to a target concept. Some pipes are fatter than others, and thus convey more information: think of resonant metaphors or rich analogies that yield deeper meaning the more you look at them. By convention, pipes carry information in one direction only, from source to target. But creativity is no respecter of convention, and creative comparisons are sometimes a two-way affair.

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

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

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

This *Pitt 2.0* doesn't actually exist, of course. Hawke's description is a *conceptual blend* that constructs a whole new source concept in its own counterfactual space. Blending is pervasive in modern culture, and can be seen in everything from cartoons to movies to popular fiction, while the elements of a blend can come from any domain of experience, from classic novels to 140-character tweets to individual words. As defined by the cognitive linguists Gilles Fauconnier and Mark Turner (1998, 2002), conceptual blending combines the smoothness of metaphor with the structural complexity and organizing power of analogy. We can think of blending as a cognitive operation in which conceptual ingredients do not flow in a single direction, but are thoroughly stirred together, to create a new structure with its own emergent meanings.

The *Kristofferson-as-Pitt* blend shows just how complex a conceptual blend can be, while nonetheless remaining intelligible to a reader: when we interpret these constructs,

we are not aware of any special challenge being posed, or of any special machinery being engaged. Nonetheless, this kind of blend poses significant problems for our computers and their current linguistic/cognitive-modelling abilities. In this paper we propose a computional middle-ground, called a *conceptual mash-up*, that captures some of the power and utility of a conceptual blend, but in a form that is practical and robust to implement on a computer. From this starting point we can begin to make progress toward the larger goal of creative computational systems that – to use Hawke's word – can formulate truly *badass* blends of their own.

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

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

Related Work and Ideas

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

Veale and O'Donoghue (2000) proposed an SMT-based model of conceptual blending that was perhaps the first computational model of the phenomenon. The model, called *Sapper*, addresses many of the problems faced by SME – such as deciding for itself which knowledge is relevant to a blend – but succumbs to others, such as the need for a hand-crafted knowledge base. Pereira (2007) presents an alternative computational model that combines SMT with other computational techniques, such as using genetic algorithms to search the space of possible blends. Pereira's model was applied both to linguistic problems (such as the interpretation of novel noun-noun compounds) and to visual problems, such as the generation of novel monsters/creatures for video games. Nonetheless, Pereira's approach was just as reliant on hand-crafted knowledge. To explore the computational uses of blending without such a reliance on specially-crafted knowledge, Veale (2006) showed how blending theory can be used to understand novel portmanteau words – or "formal" blends – such as "Feminazi" (Feminist + Nazi). This approach, called *Zeitgeist*, automatically harvested and interpreted portmanteau blends from Wikipedia, using only Wikipedia itself and Wordnet (Fellbaum, 1998) as resources.

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

generation and comprehension.

Shutova (2010) combines elements of several of these approaches. She annotates verbal metaphors in corpora (such as "to *stir* excitement", where the verb "stir" is used metaphorically) with the corresponding conceptual metaphors identified in Lakoff and Johnson (1980). Statistical clustering techniques are then used to generalize from the annotated exemplars, allowing the system to recognize other metaphors in the same vein (e.g. "he *swallowed* his anger"). These clusters can also be analyzed to identify literal paraphrases for a given metaphor (such as "to *provoke* excitement" or "*suppress* anger"). Shutova's approach is noteworthy for the way it operates with Lakoff and Johnson's inventory of conceptual metaphors without actually using an explicit knowledge representation.

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

"Milking" Knowledge from the Web

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

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

For instance, the topic *pirate* yields the query "*why do pirates **", to retrieve snippets that include these questions:

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

Parsing the 2nd question above, we obtain its logical form:

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

A computational system needs a critical mass of such commonsense knowledge before it can be usefully applied to problems such as conceptual blending. Ideally, we could extract a large body of everyday musings from the query logs of a search engine like Google, since many users persist in using full NL questions as Web queries. Yet such logs are jealously guarded, not least on concerns about privacy. Nonetheless, engines like Google do expose the most common queries in the form of text completions: as one types a query into the search box, Google anticipates the user's query by matching it against past queries, and offers a variety of popular completions.

In an approach we call *Google milking*, we coax completions from the Google search box for a long list of strings with the prefix "why do", such as "why do a" (which prompts "*why do a*nimals hibernate?"), and "why do aa" (which prompts "*why do aa* batteries leak?"). We use a manual trie-driven approach, using the input "why do X" to determine if any completions are available for a topic prefixed with X, before then drilling deeper with "*why do X*a" … "*why do X*z". Though laborious, this process taps into a veritable mother lode of nuggets of conventional wisdom. Two weeks of milking yields approx. 25,000 of the most common questions on the Web, for over 2,000 topics, providing critical mass for the processes to come.

Conceptual "Mash-ups"

Google milking yields these frequent questions about *poets*

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

Querying the web directly, the system finds other common presuppositions about poets, such as "*why do poets die poor?*" and "*why do poets die young?*", precisely the kind

of knowledge that shapes our stereotypical view of poets yet which one is unlikely to find in a dictionary. Now suppose a user asks the system to explore the ramifications of the blend *Philosophers are Poets*: this prompts the system to introspectively ask "*how are philosophers like poets?*". This question spawns others, which are produced by replacing the subject of the *poet*-specific questions above, yielding new introspective questions such as "*do philosophers write poetry?*", "*do philosophers use metaphors?*", and "*do philosophers write about love?*".

Each repurposed question can be answered by again appealing to the web: the system simply looks for evidence that the hypothesis in question (such as "*philosophers use metaphors*") is used in one or more web texts. In this case, the Google API finds supporting documents for the following hypotheses: "*philosophers die poor*" (3 results), "*philosophers die young*" (6 results), "*philosophers use metaphors*" (156 results), and "*philosophers write about love*" (just 2 results). The goal is not to show that these behaviors are as salient for philosophers as they are for poets, rather that they can be meaningful for philosophers.

We refer to the construct *Philosophers are Poets* as a *conceptual mash-up*, since knowledge about a source, *poet*, has been mashed-up with a given target, *philosopher*, to yield a new knowledge network for the latter. Conceptual mash-ups are a specific kind of conceptual blend, one that is easily constructed via simple computational processes.

To generate a mash-up, the system starts from a given target T and searches for the source concepts $S_1 \ldots S_n$ that might plausibly yield a meaningful blend. A locality assumption limits the scale of the search space for sources, by assuming that T must exhibit a pragmatic similarity to any vehicle S_i. Budanitsky and Hirst (2006) describe a raft of term-similarity measures based on WordNet (Fellbaum, 1998), but what is needed for blending is a generative measure: one that can quantify the similarity of T to S as well as suggest a range of likely S's for any given topic T.

We construct such a measure via corpus analysis, since a measure trained on corpora can easily be made corpus-specific and thus domain- or context-specific. The Google ngrams (Brants and Franz, 2006) provide a large collection of word sequences from Web texts. Looking to the 3-grams, we extract coordinations of generic nouns of the form "Xs and Ys". For each coordination, such as "*tables and chairs*" or "*artists and scientists*", X is considered a pragmatic (rather than semantic) neighbor of Y, and vice versa. When identifying blend sources for a topic T, we consider the neighbors of T as candidate sources for a blend. Furthermore, if we consider the neighbors of T to be features of T, then a vector space representation for topics can be constructed, such that the vector for a topic T contains all of the neighbors of T that are identified in the Google 3-grams. In turn, this vector representation allows us to calculate the similarity of a topic T to a source S, and rank the neighbors $S_1 \ldots S_n$ of T by their similarity to T.

Intuitively, writers use the pattern "Xs and Ys" to denote an ad-hoc category, so topics linked by this pattern are not just similar but truly comparable, or even interchangeable. Potential sources for T are ranked by their perceived similarity to T, as described above. Thus, when generating mash-ups for *philosopher*, the most highly ranked sources suggested via the Google 3-grams are: *scholar, epistemologist, ethicist, moralist, naturalist, scientist, doctor, pundit, savant, explorer, intellectual* and *lover*.

Multi-Source Mash-Ups

The problem of finding good sources for a topic T is highly under-constrained, and depends on the contextual goals of the speaker. However, when blending is used for knowledge acquisition, multi-source mash-ups allow us to blend a range of sources into a rich, context-free structure. If $S_1 \ldots S_n$ are the n closest neighbors of T as ranked by similarity to T, then a mash-up can be constructed to describe the semantic potential of T by collating all of the questions from which the system derives its knowledge of $S_1 \ldots S_n$, and by repurposing each for T. A complete mashup collates questions from all the neighbors of a topic, while a 10-neighbor mashup for *philosopher*, say, would collate all the questions possessed for *scholar ... explorer* and then insert *philosopher* as the subject of each. In this way a conceptual picture of *philosopher* could be created, by drawing on beliefs such as *naturalists tend to be pessimistic* and *humanists care about morality*.

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

The generality of a question is related to the number of times it appears in our knowledge-base with different subjects. Thus, "*why do ___ wear black*" appears 21 times, while "*why do ___ wear black hats*" and "*why do ___ wear white coats*" each just appear twice. When a mash-up for a topic T is presented to the user, each imported question Q is ranked according to two criteria: Q_{count}, the number of neighbors of T that suggest Q; and Q_{sim}, the similarity of T to its most similar neighbor that suggests Q (as calculated

using a WordNet-based metric; see Seco et al., 2006). Both combine to give a single salience measure $Q_{salience}$ in (1):

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

Note that Q_{count} is always greater than 0, since each question Q must be suggested by at least one neighbor of T. Note also that salience is not a measure of surprise, but of aptness, so the larger Q_{count}, the larger $Q_{salience}$. It is time-consuming to test every question in a mash-up against web content, as a mash-up of *m* questions requires *m* web queries. It is more practical to choose a cut-off *w* and simply test the top *w* questions, as ranked by salience in (1). In the next section we evaluate the ranking of questions in a mash-up, and estimate the likelihood of successful knowledge transfer from one topic to another.

Empirical Evaluation

Our corpus-attested, neighborhood-based approach to similarity does not use WordNet, but is capable of replicating the same semantic divisions made by WordNet. In earlier work, Almuhareb and Poesio (2004) extracted features for concepts from text-patterns found on the web. These authors tested the efficacy of the extracted features by using them to cluster 214 words taken from 13 semantic categories in WordNet (henceforth, we denote this experimental setup as AP214), and report a cluster purity of **0.85** in replicating the category structures of WordNet. But if the neighbors of a term are instead used as features for that term, and if a term is also considered to be its own neighbor, then an even higher purity/accuracy of **0.934** is achieved on AP214. Using neighbors as features in this way requires a vector space of just 8,300 features for AP214, whereas Almuhareb and Poesio's original approach to AP214 used approx. 60,000 features.

The locality assumption underlying this notion of a pragmatic neighborhood constrains the number of sources that can contribute to a multi-source mash-up. Knowledge of a source S can be transferred to topic T only if S and T are neighbors, as identified via corpus analysis. Yet, the Google 3-grams suggest a wealth of neighboring terms, so locality does not unduly hinder the transfer of knowledge. Consider a test-set of 10 common terms, *artist, scientist, terrorist, computer, gene, virus, spider, vampire, athlete* and *camera*, where knowledge harvested for each of these terms is transferred via mash-ups to all of their neighbors. For instance, "*why do artists use Macs?*" suggests "*musicians use Macs*" as a hypothesis because *artists* and musicians are close neighbors, semantically (in WordNet) and pragmatically (in the Google n-grams); this hypothesis is in turn validated by 5,700 web hits. In total, 410,000 hypotheses are generated from these 10 test terms, and when posed as web queries to validate their content, approx. 90,000 (21%) are validated by usage in web texts.

Just as knowledge tends to cluster into pragmatic neighborhoods, hypotheses likewise tend to be validated in clusters. As shown in Figure 1, the probability that a hypothesis is valid for a topic T grows with the number of neighbors of T for which it is known to be valid (Q_{count}).

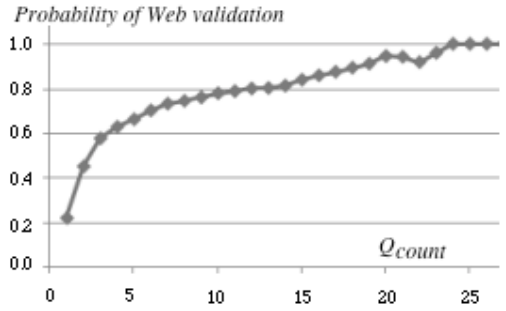


Figure 1. *Likelihood of a hypothesis in a mash-up being validated via web search (y-axis) for hypotheses that are suggested by Q_{count} neighbors (x-axis).*

Unsurprisingly, close neighbors with a high similarity to the topic exert a greater influence than more remote neighbors. Figure 2 shows that the probability of a hypothesis for a topic being validated by web usage grows with the number of the topic's neighbors that suggest it *and* its similarity to the closest of these neighbors ($Q_{salience}$).

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

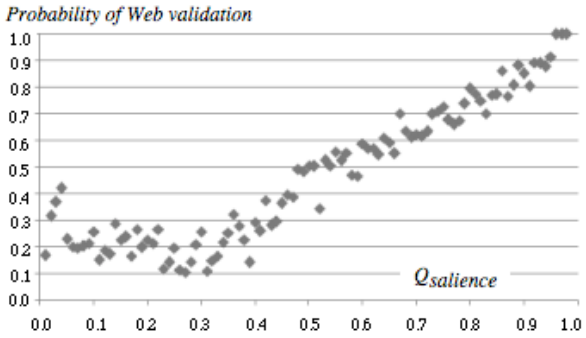


Figure 2. *Likelihood of a hypothesis in a mash-up being validated via web search (y-axis) for hypotheses with a particular $Q_{salience}$ measure (x-axis).*

Figure 3 shows the average rate of web validation for the top-ranked hypotheses (ranked by salience) of complete mash-ups generated for each of our 10 test terms from *all* of their neighbors. Since these are common terms, they have many neighbors that suggest many hypotheses. On average, 85% of the top 20 hypotheses in each mash-up are

validated on by web search as plausible, while just 1 in 4 of the top 60 hypotheses in a mashup is not web-validated.

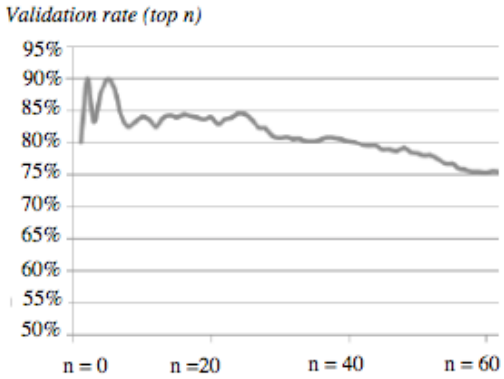


Figure 3. *Average % of top-n hypotheses in a mash-up (as ranked by $Q_{salience}$) that are validated by Web search.*

Figures 1 – 3 show that the system is capable of extracting knowledge from the web which can be successfully transferred to neighboring terms via metaphors and mash-ups, and then meaningfully ranked by salience. But just how useful is this knowledge? To determine if it is the kind of knowledge that is useful for categorization – and thus the kind that captures the perceived essence of a concept – we use it to replicate the AP214 categorization test of Poesio and Almuhareb (2004). Recall that AP214 tests the ability of a feature–set / representation to support the category distinctions imposed by WordNet, so that 214 words can be clustered back into the 13 WordNet categories from which they are taken. Thus, for each of these 214 words, we harvest questions from the Web, and treat each question body as an atomic feature of its subject.

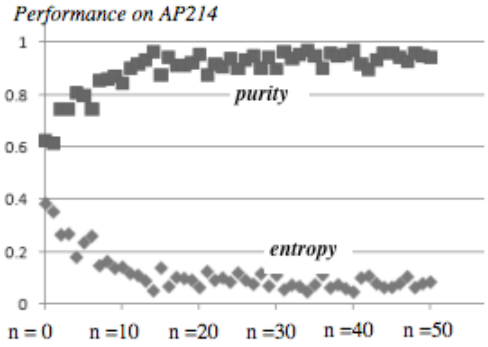


Figure 4. *Performance on AP214 improves as knowledge is transferred from the **n** closest neighbors of a term.*

Clustering over these features alone offers poor accuracy when reconstructing WordNet categories, yielding a cluster purity of just over 0.5. One AP214 category in particular, for time units like *week* and *year*, offers no traction to the question-based approach, and accuracy / purity increases to 0.6 when this category is excluded. People, it seems, rarely question the conceptual status of an abstract temporal unit.

But as knowledge is gradually transferred to the terms in AP214 from their corpus-attested neighbors, so that each term is represented as a conceptual mash-up of its *n* nearest neighbors, categorization markedly improves. Figure 4 shows the increasing accuracy of the system on AP214 (excluding the vexing *time* category) when using mashups of increasing numbers of neighbors. Blends really do bolster our knowledge of a topic with insights that are relevant to categorization.

Conclusions: A Metaphor-Eye to the Future

We have shown here how common questions on the web can provide the world knowledge needed to drive a robust, if limited, form of blending called *conceptual mash-ups*. The ensuing powers of introspection, though basic, can be used to speculate upon the conceptual make-up of a given topic, not only in individual metaphors but in rich, informative mash-ups of multiple concepts.

The web is central to this approach: not only are questions harvested from the web (e.g., via Google "milking"), but newly-formed hypotheses are validated by means of simple web queries. The approach is practical, robust and quantifiable, and uses an explicit knowledge representation that can be acquired on demand for a given topic. Most importantly, the approach makes a virtue of blending, and argues that we should view blending not as a problem of language but as a *tool* of creative thinking.

The ideas described here have been computationally realized in a web application called *Metaphor-Eyes*. Figure 5 overleaf provides a snapshot of the system in action. The user enters a query – in this case the provocative assertion "*Google is a cult*" – and the system provides an interpretation based on a mash-up of its knowledge of the source (cults) and of the target (Google). Two kinds of knowledge are used to provide the interpretation of Figure 5. The first is common-sense knowledge of cults, of the kind that we expect most adults to possess. This knowledge includes widely-held stereotypical beliefs such as that cults are lead by gurus, that they worship gods and enforce beliefs, and that they recruit new members, especially celebrities, which often act as apologists for the cult. The system possesses no stereotypical beliefs about Google, but using the Google 2-grams (somewhat ironically, in this case), it can find linguistic evidence for the notions of a *Google guru*, a *Google god* and a *Google apologist*. The corresponding stereotypical beliefs about cults are then projected into the new blend space of *Google-as-a-cult*.

Metaphor-Eyes derives a certain robustness from its somewhat superficial treatment of blends as mash-ups. In essence, the system manipulates conceptual-level objects (ideas, blends) by using language-level objects (strings, phrases, collocations) as proxies: a combination at the concept-level is deemed to make sense if a corresponding combination at the language-level can be found in a corpus (or in the Google n-grams). As such, any creativity

exhibited by the system is often facile or glib. Because the system looks for conceptual novelty in the veneer of surface language, it follows in the path of humour systems that attempt to generate interesting semantic phenomena by operating at the punning level of words and their sounds.

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

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

References

Almuhareb, A. and Poesio, M. (2004) Attribute-Based and Value-Based Clustering: An Evaluation. *In Proceedings Of EMNLP'2004*, pp 158-165.

Barnden, J. A. 2006. Artificial Intelligence, figurative language and cognitive linguistics. *G. Kristiansen, M. Achard, R. Dirven, and F. J. Ruiz de Mendoza Ibanez (Eds.), Cognitive Linguistics: Current Application and Future Perspectives*, 431-459. Berlin: Mouton de Gruyter.

Brants, T. and Franz, A. 2006. Web 1T 5-gram Version 1. *Linguistic Data Consortium*.

Budanitsky, A. and Hirst, G. 2006. Evaluating WordNet-based Measures of Lexical Semantic Related- ness. *Computational Linguistics*, 32(1):13-47.

Falkenhainer, B., Forbus, K. and Gentner, D. 1989. Structure-Mapping Engine: Algorithm and Examples. *Artificial Intelligence*, 41:1-63.

Gilles Fauconnier and Mark Turner. (1998). Conceptual Integration Networks. *Cognitive Science,* 22(2):133–187.

Gilles Fauconnier and Mark Turner. (2002). *The Way We Think. Conceptual Blending and the Mind's Hidden Complexities.* Basic Books.

Fellbaum, C. (ed.) 2008. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge.

Gentner, D. 1983, Structure-mapping: A Theoretical Framework. *Cognitive Science* 7:155–170.

Hearst, M. 1992. Automatic acquisition of hyponyms from large text corpora. In Proc. of the 14th International Conference on Computational Linguistics, pp 539–545.

Lakoff, G. and Johnson, M. 1980. *Metaphors we live by*. University of Chicago Press.

Pasca, M. and Van Durme, B. 2007. What You Seek is What You Get: Extraction of Class Attributes from Query Logs. In *Proceedings of IJCAI-07, the 20th International Joint Conference on Artificial Intelligence*.

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

Seco, N., Veale, T. and Hayes, J. 2004. An Intrinsic Information Content Metric for Semantic Similarity in WordNet. In the proceedings of ECAI 2004, the 16th European Conference on Artificial Intelligence. Valencia, Spain. John Wiley

Shutova, E. 2010. Metaphor Identification Using Verb and Noun Clustering. *In Proceedings of the 23rd International Conference on Computational Linguistics*, 1001-1010.

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

Veale, T. and D. O'Donoghue. (2000). Computation and Blending. *Cognitive Linguistics*, 11(3-4):253-281.

Veale, T. 2006. Tracking the Lexical Zeitgeist with Wikipedia and WordNet. *Proceedings of ECAI-2006, the 17th European Conference on Artificial Intelligence*.

Veale, T. and Hao, Y. 2007a. Making Lexical Ontologies Functional and Context-Sensitive. In *Proceedings of the 46th Ann. Meeting of Assoc. of Computational Linguistics*.

Veale T. and Hao, Y. 2007b. Comprehending and generating apt metaphors: a web-driven, case-based approach to figurative language. In *Proceedings of AAAI'2007, the 22nd national conference on Artificial intelligence*, pp.1471-1476.

Veale, T. and Li, G. 2011. Creative Introspection and Knowledge Acquisition: Learning about the world thru introspective questions and exploratory metaphors. In Proc. of AAAI'2011, the 25th Conference of the Association for the Advancement of Artificial Intelligence.

Veale, T. 2012 *Exploding the Creativity Myth: The Computational Foundations of Linguistic Creativity*. London: Bloomsbury/Continuum.

Figure 5. *A screen-shot from the computational system* Metaphor-Eyes*, which implements the model described in this paper.* Metaphor-Eyes *shows how we can use conceptual mash-ups to explore what-ifs and to stimulate human creativity. (Note: Because the system has no prior ontological knowledge about Google, each entry above shows a default score of 100 and a support/similarity measure of 0). Please visit* http://Afflatus.UCD.ie *to interact with the* Metaphor-Eyes *system for yourself, or to find out more about the system's XML functionality.*

Goal-Driven Conceptual Blending: A Computational Approach for Creativity

Boyang Li, Alexander Zook, Nicholas Davis and Mark O. Riedl
School of Interactive Computing, Georgia Institute of Technology, Atlanta GA, 30308 USA
{boyangli, a.zook, ndavis35, riedl}@gatech.edu

Abstract

Conceptual blending has been proposed as a creative cognitive process, but most theories focus on the analysis of existing blends rather than mechanisms for the efficient construction of novel blends. While conceptual blending is a powerful model for creativity, there are many challenges related to the computational application of blending. Inspired by recent theoretical research, we argue that contexts and context-induced goals provide insights into algorithm design for creative systems using conceptual blending. We present two case studies of creative systems that use goals and contexts to efficiently produce novel, creative artifacts in the domains of story generation and virtual characters engaged in pretend play respectively.

Introduction

Conceptual blending has been proposed as a fundamental cognitive process, responsible for the creation of a broad range of creative artifacts. (Fauconnier and Turner 1998, 2002; Grady 2000; Hutchins 2005). Fauconnier and Turner (1998, 2002) proposed that conceptual blending involves the merger of two or more input spaces into a blended space. A griffin, for example, may be considered as a blend of an eagle and a lion. Each of the input spaces contains a number of concepts and their inter-connections. These concepts are selectively merged and projected into the blend space. After that, additional structures may emerge in the blend by pattern completion and further elaboration.

Among artifacts created by conceptual blending, we distinguish two types of blends, namely *semiotic expressions* and *standalone concepts*. Semiotic expressions are used in communication to highlight certain aspects of or shed light on one of the input spaces. An often discussed semiotic expression is "this surgeon is a butcher" (cf. Grady, Oakley, and Coulson 1999; Brandt and Brandt 2002; Veale and O'Donoghue 2000). The input spaces of the expression are the conceptual spaces of the surgeon and the butcher respectively. The surgeon in the blend possesses the brutal attitude of the butcher, forming a criticism of the surgeon.

This paper focuses on the construction of the second type of blend, which we call standalone concepts. An example is the lightsaber from *Star Wars*: a lightsaber blends together a sword and a laser emitter, but it is an independent concept that does not inform hearers about the properties of swords or laser emitters. During blend creation, contents are still projected from the input spaces into the blend, but the blend is not meant to convey information about the input spaces.

We share the belief that theories of creativity should be computable (Johnson-Laird 2002), yet most accounts of blending focus on analyses of existing blends and have not fully described how novel blends are constructed cognitively or algorithmically. In particular, three key procedures required for blending lack sufficient details necessary for efficient computation: (1) the selection of input spaces, (2) the selective projection of elements of input spaces into the blend, and (3) the stopping criteria for blend elaboration. Inefficiencies in these procedures can lead to significant difficulties in finding appropriate blends and elaborations. For example, a simplistic algorithm may produce all possible combinations of elements from all input spaces, resulting in a combinatorial explosion of possible blends.

We argue that these three main procedures must algorithmically make use of the *context* and *goals* of the blend being constructed. Brandt and Brandt (2002) proposed communication contexts and goals as the driving force behind the three key procedures, but their analysis is limited to semiotic expressions. We extend their theory to the construction of novel standalone concepts and provide computational justifications through two case studies of working computational systems in the domains of story generation and pretend play. Our systems construct blends in a goal-driven and context-driven manner. As integral aspects of the conceptual blending process, contexts and goals provide concrete computational benefits by pruning search spaces and improving average-case performance.

The Theories of Conceptual Blending

This section reviews the theories of conceptual blending described by Fauconnier and Turner (1998, 2002) and Brandt and Brandt (2002). We compare these two accounts side by side and identify some underspecified parts in the theories, which a working system must address.

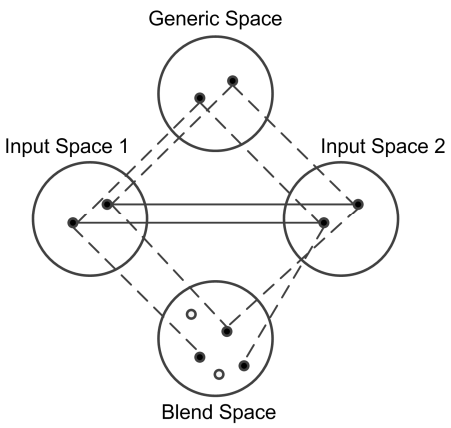


Figure 1. The four-space blending theory, adapted from Fauconnier & Turner (2002).

In the original blending theory (BT) by Fauconnier and Turner (1998, 2002), conceptual blending takes two or more mental spaces as inputs. Mental spaces are dynamically constructed during a discourse (e.g. conversation) to contain relevant concepts. The input space of the surgeon, for example, includes the surgeon and relevant entities, such as his scalpel, patient and so on. Elements in one input space are then mapped to their counterparts in another input space, using mapping rules such as identity or analogy. In the surgeon-as-butcher example, the cleaver of the butcher is mapped to the scalpel of the surgeon; the dead animal is mapped to the patient, and so forth. Elements from input spaces are selectively projected into a *blend space*. The *generic space* captures the structural similarities between input spaces. In addition to elements projected from input spaces, the blend space can also contain emergent structures created by pattern completion or elaboration. This four-space formulation is shown in Figure 1, where big circles denote mental spaces, black dots represent elements in the spaces, solid lines are the mappings between inputs, and dashed lines denote correspondences among the elements in the four spaces. Hollow dots denote emergent structures in the blend.

Fauconnier and Turner, however, did not specify how elements from these input spaces could be chosen during the selective projection. Although eight optimality principles—human scale, topology, pattern completion, integration, vital relations, unpacking, web, and relevance—were proposed as quality measures, they can only evaluate the quality of a complete blend *after* it is constructed. This suggests a computational approach where all possible blends have to be generated and tested individually, called a neo-Darwinian algorithm by Johnson-Laird (2002). A neo-Darwinian algorithm could lead to a combinational explosion of options and is infeasible for large input spaces. In contrast, what Johnson-Laird calls a neo-Lamarckian approach generates only valuable products by applying quality constraints on the search space. To do so in blending requires a mechanism

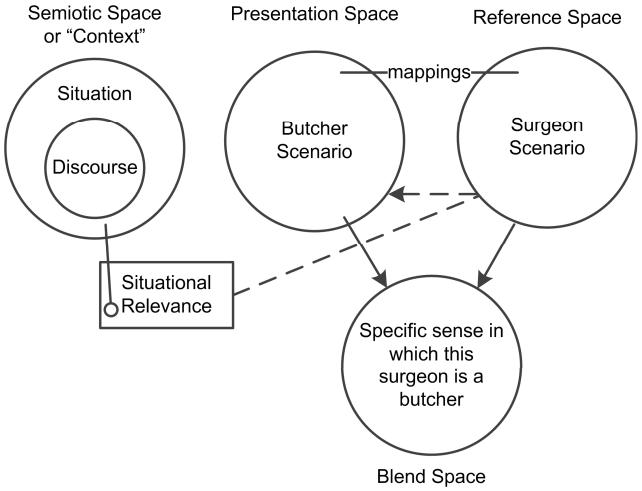


Figure 2. The context-dependent blending theory, adapted from Brandt and Brandt (2002).

that selects elements from input spaces effectively during the generation process. Note that projection occurs after inter-space mappings are built, so the complexity of analogy making should not be conflated with the complexity of projection.

Moreover, BT does not provide detailed procedures for the effective retrieval of input spaces, nor for the elaboration of blends. A full computational implementation of the blending theory should select input spaces by itself, rather than assume them as given. To create a powerful criticism of the surgeon, a system should decide to blend it with a butcher, rather than a driver or a school teacher. A creative system may possess a huge amount of knowledge, so an efficient selection procedure for input spaces is necessary for it to operate within reasonable time limits. The same argument goes for elaboration. Neither a human nor a computational system should elaborate a blend endlessly. We usually do not wish to simulate the entire world's reaction to an irresponsible surgeon, which will require excessive computational power or time.

In summary, the original blending theory left much ambiguity in three key procedures: (1) the selection of input spaces, (2) the selection of elements for projection, and (3) the stopping criterion for blend elaboration. However, a complete working implementation of BT must contain these procedures.

Context-Dependent Blending

Brandt and Brandt (2002) pointed out that blends used in communication do not have fixed meanings. Rather, they are real-world phenomena that can only be analyzed in the context of the discourse during which they were uttered. Under different circumstances, the same utterance "this surgeon is a butcher" can mean different things. For example, if a soldier referred to a battlefield medic as a butcher, he may be highlighting the fact that he has to perform an astonishing number of amputations. This

interpretation is vastly different than the ethical judgment about a surgeon's skills after a failed surgery.

To account for this plurality of meanings, blend construction must not be based solely on the attributes of the input domains, but on the context. Hence, Brandt and Brandt abandoned the context-free generic space and proposed a context-driven blend construction process. A simplified version of Brandt and Brandt's theory is shown in Figure 2, where the dashed arrow indicates that the situational relevance prompts the retrieval of the two input spaces in order and solid arrows represent selective projection. First, the communicative situation (context) retrieves the input spaces in order. The situation implies a first input space, the *reference space* based in the actual real-world situation. For example, if we are talking about a particular surgery, the input space of the surgeon becomes available. Based on the goal of communication (e.g. conveying that the surgeon is irresponsible), a *presentation space* including the figurative entities (e.g. a butcher) is then retrieved and mapped to the reference space. Second, the goal of communication, captured through the situational relevance, determines what elements from the two spaces are projected into the blend. If we want to accuse the surgeon of being irresponsible, we should project the careless attitude of the butcher into the blend. Third, elaboration of blends also depends on the context and the goal, fleshing out a blended space until sufficient detail is achieved for the guiding goal. In summary, the sequential retrieval of input spaces, the selection of projected elements, and the stopping criteria are all driven by contexts and goals.

Blending Novel Concepts

In addition to metaphorical expressions, blending also yields novel concepts independent of the input spaces, such as the lightsaber. We believe the two cases differ in the relationship between the inputs spaces and the blend.

Blends used in communication are usually meant to underscore certain aspects of, or to attach new properties to an input space. In our example, the surgeon who possesses a butcher's attitude in the blend space becomes a criticism of the surgeon in the input. The linkage between the blend and the input spaces is essential for understanding.

In contrast, standalone concepts are blends that are not meant to convey meaning about their input spaces. For example, the birth stork is a blend of the metaphor birth-is-arrival, air travel, and the stork (Fauconnier and Turner 2002, ch. 14), but it does not tell us much about air travel or storks in general. As another example, the lightsaber in *Star Wars* is clearly a blend of a sword and a laser emitter, but it is not meant to tell us anything about swords or laser, even though understanding laser and swords can help us understand lightsabers. There are links going from the input spaces to the blend spaces, but not vice versa.

Standalone concepts created from blending are commonly seen in stories and story-related activities. Dragons, for example, possess features of snakes, large cats and birds of prey, and the instinctive fear of the three is postulated to be its origin (Jones 2002). Many other mythical creatures are combinations of common animals. Novel concepts created from blending also appear in modern science fiction. The popular Japanese sci-fi manga *Doraemon* (Fujio 1974-1996) contains several gadgets made this way, such as a toy telephone that can transmit flu instead of voice.

The idea of an independent concept does not contradict the notion of contextualized meaning. The meaning of a novel concept, when used as a semiotic sign, can still change depending on the context. One may say "my new kitchen knife is a lightsaber" to emphasize its sharpness. In Brandt and Brandt's framework, this meaning is created from the blend of the kitchen knife and the lightsaber, which is a different blend than the lightsaber itself. The concept lightsaber, on the other hand, is as independent as the concept butcher. For another example, the birth stork is now often a humorous symbol for baby births rather than used to explain births. The contexts and meanings vary, but the concepts remain relatively constant.

We extend Brandt and Brandt's contextualized blending theory to the construction of novel concepts. Below we present two computational systems that generate novel concepts in fictions and pretend play. While Brandt and Brandt's (2002) theory initially was not meant to explain such blends, we show that a goal and context driven approach can account for their generation and bring computational benefits.

Previous Implementations

Most previous work on computational algorithms for conceptual blending are based on the theories of Fauconnier and Turner and thus do not fully account for input space selection, selective projection, or blend elaboration. It is common for computational blending algorithms to ignore one or more of these stages. The Alloy blending engine, as part of the Griot poem generator (Goguen and Harrell 2004), is the earliest implementation of BT that we are aware of. Input spaces are manually coded as symbolic expressions. Projection into the blend is based on a structural mapping between input spaces. Without the guidance of goals, any element from the input spaces may be projected. The authors found the number of possible blends is exponential to the number of relations in the input spaces, but did not discuss methods to prune these spaces. Hervás et al. (2006) proposed a process that enriches texts of stories which can be considered as conceptual blends. They proposed that readers' familarity with input spaces, as part of the communication context, is important in selecting input spaces and elements for projection without specifying an algorithm.

Martinez et al. (2011) proposed a blending algorithm where input spaces are sets of axioms. Compatible axioms are selected for projection. The system does not consider goals to directly build blends that meet specific purposes. Rather, it "enumerates alternatives ranked by the complexity of the underlying mappings". Yamada et al. (2011) generate motion of dancing characters by

representing motion as wavelet equations and blending is a weighted summation of wavelet coefficients. The process does not involve any of the three key procedures mentioned earlier. Thagard and Steward (2011) proposed an implementation of blending at the neural level, which also does not consider goals.

The Divago system (Pereira 2007) is noteworthy in that it considers goals, but only uses them indirectly during construction of standalone concepts. Note that the input spaces are given to—rather than selected by—Divago. In Divago, goals do not directly participate in the selection of elements being projected. If an element *a* is mapped to *b*, one of 4 things can be projected into the blend: *a*, *b*, *a*/*b*, or nothing. This leads to a combinatorial explosion of possible blends regardless of the complexity of finding an inter-space analogical mapping. To effectively search an exponentially growing space, Divago utilizes a genetic algorithm (GA) that stochastically samples the space of possible blends. From a population of blends, Divago first selects those with high scores as computed by an evaluation function, consisting of the weighted sum of the eight optimality criteria introduced earlier. One of the criteria, relevance, is interpreted as goal satisfaction. After evaluation, highly ranked blends are randomly modified to create the next population. Imitating biological evolution, this process repeats in the hope of finding a near-optimal blend after sufficient number of iterations. Note however that a high score does not guarantee goal satisfaction because the evaluation function contains several, possibly competing criteria.

To elaborate a blend, Divago fires any production rules whose premises are true to add content into the blend. Divago also supplements details to the blend based on its similarity with other frames. For example, if a blend is similar enough to the bird frame, Divago will grant it the ability to fly. However, given enough rules and frames, rule firing and pattern matching can potentially go on indefinitely, as it does not specify any explicit stopping criteria for elaboration based on the notions of meaningfulness or necessity.

In the light of the above analysis, it is clear that goals and context can guide a computational conceptual blending process for the purpose of creating novel, standalone concepts. However, context and goals must be used directly to ensure successful blends and to focus search efficiently. In the next section, we study two creative systems that utilize contexts and goals to effectively realize the three procedures in conceptual blending.

Two Case Studies

This section presents two systems that implement blending theory in a goal and context driven manner. We describe these systems through the lens of the context-driven blending theory. The first system builds fictional gadgets in computer-generated stories. The second system constructs objects used in pretend play that combine features of a desired fantasy-world object with a real-world object at hand. The two systems address the three aforementioned problems—input space selection, selective projection, and elaboration—in a neo-Lamarckian manner by employing constraints introduced by the domain of application and the specific goal to be achieved. The first case study focuses on selective projection and elaboration. The second case study focuses on input space selection.

Generating Gadgets in Fictions

As a vibrant research field, artificial intelligence (AI) story generation aspires to create intelligent systems that can create and tell novel stories. Most current approaches to story generation are restricted to generating stories for static, hand-authored micro-worlds, manipulating given characters and objects to produce stories (e.g. Cavazza, Charles, and Mead 2002; Gervás et al. 2005; Riedl and Young 2010; Ontañón and Zhu 2010). These systems can be likened to jigsaw puzzle solvers who play only with given pieces and never dream of inventing a new piece. These story generation systems are not able to tell stories with novel objects or gadgets; they cannot tell *Star Wars* if the idea of lightsaber is not supplied ahead of time.

The aim of the gadget generation algorithm (Li and Riedl 2011a, 2011b) is to break out of these static world configurations and create new types of objects previously unknown to the system. Our approach was initially presented as a combination of partial-order planning (Weld 1994) and analogical reasoning. Here we point out its connection to conceptual blending. The algorithm blends existing concepts to generate novel standalone concepts as unforeseen gadgets in support of a goal derived from a story context.

To generate a gadget, the algorithm reasons about how the gadget should be used in the context of a story, including events that happen immediately before and after its usage. These events are captured as the behavior of the gadget, represented as a temporally ordered sequence of actions. Given a goal derived from a story (e.g., a character must become infected by a flu virus), the algorithm iteratively constructs the gadget's behavior by working backward from the goal using actions and entities from various input spaces. Goals are first-order logic predicates such as `infected-by(bob, virus)`. An action is an operator that requires certain predicates as preconditions and asserts some predicates as effects. The gadget generator works with a conventional story generator, which supplies goals considered appropriate for a gadget to achieve. The final behavior of the gadget must achieve these goals.

Goals are first used to identify the input spaces. The reference space includes objects in the goal predicates and relevant concepts. For example, the reference space implied by the goal `infected-by(bob, virus)` includes concepts such as flu viruses, a character named Bob, and actions such as coughing and curing. In fact, the reference space exists only conceptually and is not separated from the rest of the knowledge in the system. It highlights that knowledge structures closely related to the goal play more important roles in projection than the rest. The system

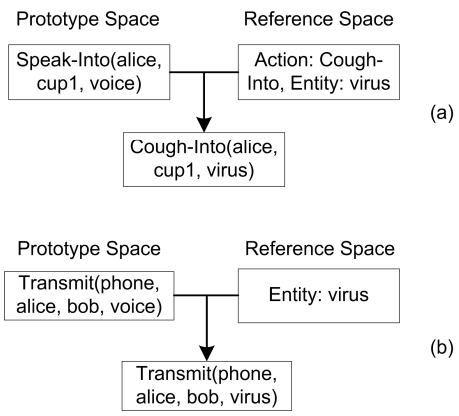


Figure 3. Two projections involved in the generation of the flu-transmitting gadget phone.

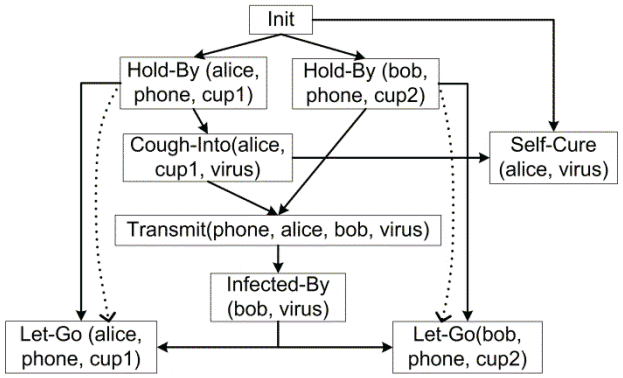


Figure 4. The behavior of the flu-transmitting gadget phone

creates the second input space by retrieving an object from its knowledge base of many known objects that achieve predicates analogous to the specified goals. This object is called the prototype for the gadget. At this time, the list of known potential prototypes is small, and we go down the list from the best guess. In the long term, a robust mechanism to find the best prototype is needed (e.g. Wolverton 1994). The second input space, the prototype space, includes the prototype object, its behavior, as well as actions used and entities referred to in the behavior. For example, the behavior of a toy phone object refers to the entity `voice`, which becomes part of the space.

The behavior of the gadget, or the blend space, is built by projecting actions from the two input spaces selectively. This projection is driven by goals in a backward-chaining, iterative process. Following the partial-order planning representation, actions have causal requirements—predicates that must be true in the micro-world for the action to be performed (also called *preconditions*). When an action is brought in the blend space to solve one goal, its own causal requirements are added to the set of goals. Actions continue to be projected to satisfy goals until all goals are satisfied or determined to be fundamental properties of the gadget itself. Note that when there are multiple possible ways to achieve a goal, the algorithm tries the best first and backtracks when mistakes are made. The selective projection process is thus completely goal-driven. Due to space constraints, we refer interested reader to (Li and Riedl 2011b) for details of the algorithm.

There are several methods to project actions from the input spaces into the blend space. First, the algorithm can project an action from the prototype space without any changes. Second, an action from either space can be projected with arguments from both spaces, constituting a form of blending. Figure 3(a) shows the action `Cough-Into` being projected with arguments from both input spaces. This action is mapped with the action `Speak-Into` in the prototype space because of identified analogical similarities. The analogical reasoning engine Sapper (Veale and O'Donoghue 2000) is used to establish analogical mapping across input spaces. Third, a special case of blending occurs when an action from the prototype space is combined with illegal arguments from the reference space to create a new action previously not allowed. As illustrated in Figure 3(b), the `Transmit` action of a toy phone is now allowed to transmit `virus` instead of `voice`, which otherwise would not have been a legal assignment of parameters. The algorithm ignores the rules of the micro-world to achieve goals of an imaginary gadget phone that can transmit flu virus from one person to another. This allows a generated gadget to achieve the impossible, producing an object not conceived before.

Figure 4 shows the behavior of a gadget phone generated by the algorithm, which one can use to give her flu to someone else and free herself from it. The goal predicates it achieves are `infected-by (bob, virus)` and `not(infected-by (alice, virus))`. Each box represents one action. Solid arrows denote temporal precedence and dotted arrows denote closure actions. This gadget appeared in the Doraemon manga. Li and Riedl (2011a) also describes an example not from the manga.

Elaboration in the blend space is simulated with the use of closure actions. Closure actions are not necessary for the gadget's goal, but restore some goal-irrelevant aspects of the story world to an ordinary state. For example, in Figure 4, the actions where people put down the phone after use (the `Let-Go`s) are closure actions. Closure actions are manually labeled in prototype behaviors and projected into the blend space. Their use is motivated by the desire to restore an ordinary state after using the gadget, which creates a sense of denouement in the story. Although the elaboration is not directly driven by goals, it is motivated by the domain of storytelling.

Finally, the gadget is verified by incorporating the gadget behaviors into the story from which the goals were originally derived. If any necessary conditions of the gadget's behavior cannot be achieved in the story, the gadget does not work and has to be modified further by going through additional rounds of blending. When this happens, a second prototype is retrieved as yet another input space and blended with the current gadget to make it even more powerful. The algorithm cannot create two gadgets simultaneously.

By using means-ends search guided by goals, the search space of possible actions to be added to the behavior of the gadget is pruned, resulting in improved average-case efficiency of the algorithm. Each iteration of the algorithm tries to satisfy a goal in the blend space, and will thus only consider actions that can achieve that goal. While the total number of actions in both input spaces may be large, usually only a small fraction of them can be projected to achieve any given goal. In contrast, a naïve selective projection algorithm will attempt to project all actions in any input space using all possible projection methods. Although the goal-driven best-first search in the worst case has to consider all possibilities in the search space, in an average case it only considers a small portion of the total possibilities (Weld 1994). A goal-driven blending algorithm also has the added benefit of ensuring that any result of the algorithm is guaranteed to meet all of the acceptability requirements.

In summary, the gadget generation system implements a goal-driven model of blending, including relatively efficient selective projection (due to pruning of the search space) and elaboration procedures. These procedures are driven by the gadget's goal, the particular story that serves as the context, and the general domain of storytelling. However, this system has not fully investigated the selection of input spaces, which will be illustrated in the next case study.

A Virtual Character for Pretend Play

Our second case study illustrates the use of goals in blending with a specific focus on the selection of appropriate input spaces. In this case, a real-world object is selected to represent an object from a fantasy world, as required in children's pretend play. A goal specifies means to appropriately prune potential input spaces and select one option based on contextual constraints of similarity. Below we describe how our pretend play system can be viewed through the lens of conceptual blending; full details on the system are presented in (Zook, Riedl, and Magerko 2011).

In *pretend play*, children construct and enact story scripts and roles with real-life objects (Nourot 1998). Examples of pretend play include lightsaber duels with cardboard tubes, holding pretend tea parties with stuffed animal guests and acting as a group of pirates sailing on a couch. When enacting these scripts, pretenders have goals of using particular objects from a fictional world, but are limited to using the real-world objects that are ready at hand. Pretend players imaginatively overlay the fictional object onto the real-world object to create a blend, i.e. a pretend object existing in both the fictional world and the real world. Pretend play research has found children project traits of the fictional object on the real object. As an example, children engaged in a lightsaber duel with cardboard tubes may make buzzing noises when they swing the tube. In general, the pretending process involves identifying real-world objects as stand-ins for the fictional object, and selectively projecting traits of the fictional object onto the real object. The objects are input spaces to the blend. The construction process must account for how input spaces relate to a larger context of a target pretend play activity, pruning the options considered.

Building computational systems that can engage in pretend requires the capacity to construct the objects used in these scripts (Zook, Riedl, and Magerko 2011). To formalize the problem, the play activity (lightsaber duel) provides a structuring situation such that a pretender—human or agent—selects a real world object (cardboard tube) as a presentation space for a given reference space of a fictional object (lightsaber). That is, the goal is to find a presentation input space that most closely matches the reference input space. Once the input spaces are selected, the blending process takes the most relevant aspects of the fictional object for the activity (buzzing), which are imposed onto the real object in the blend space for use in play (swinging a cardboard tube while buzzing). This process starts with a situation in the fictional world and a specific fictional object (the lightsaber I am using in the duel), and seeks a presentation object in the real world to effectively manifest the fictional object.

To reason about numerous objects in the fictional world and the real world, we need a computational representation of objects and their attributes. Lakoff and Johnson (1980) proposed that the salient perceptual, motor-activity, and purposive features of objects affect how humans interact with them. We model objects in both the fictional and real domains using selected attributes in these categories. Following prototype theory (Rosch 1978), these attributes are assigned fuzzy values to represent a real-valued ([0, 1]) range of degree of membership (DOM). As an example, a lightsaber may have a 0.8 DOM value for the perceptual feature of being blue (very blue, except for the handle), 0.9 DOM value for the motor-activity feature of ease of handling (very easy to hold and swing), and 0.1 DOM value for the purpose of supporting weights (unsuitable for propping up heavy objects).

Iconic attributes are salient attributes of an object that distinguish it from similar objects within the same category. These attributes help to resolve the potential ambiguity of which fictional object is being represented by a given real world object. For example, if a pretender grabs an object and begins making buzzing noises, it may be unclear if they are signaling that they are holding a buzzing lightsaber or shooting a laser pistol. An iconic posture of handling, however, makes this difference clear. Iconic attributes help participants in pretend play interpret other players' behaviors and intentions.

The computational play system algorithm has three steps for context and goal driven blending: (1) select a real-world object based on the pretending goal and context; (2) select the set of fictional object attributes to project; and (3) project these attribute values into the blend. The first step is the selection of the input space—the real-world object—to be blended with the fictional object.

Selecting an input space uses the pretending goal to first prune impossible input spaces and then search for the optimal input space among those that remain. Conceptually

this process is similar to the surface-level filtering process used in the MAC/FAC system (Forbus, Gentner and Law 1995) with the modification of using fuzzy attributes rather than predicates. The filtering quickly removes from consideration all real-world objects that differ too significantly from the desired fictional object on a single attribute. The goal specifies attributes that are important to the pretend play object and the extent to which they must be preserved. Thus, when seeking a lightsaber, all real objects that are too difficult to handle would be ignored, as they cannot serve as useful lightsabers during the play. Computationally, all real objects are compared to the desired fictional object on the set of relevant attributes, and those that differ by a specified threshold are pruned. For example, when considering ease of handling, a cardboard tube would be kept as a candidate real object for a lightsaber, while a wooden log would be discarded.

After filtering, the remaining real-world objects are searched for the single real-world object with minimum difference from the desired fictional object. Computationally, the pretend play system exhaustively searches all available real-world objects and calculates the Euclidean distances between the attributes of each candidate real object and the fictional object. The real-world object with the minimal distance from the fictional object is then selected. In this domain, pruning appears to sufficiently constrain the set of potential input spaces to enable subsequent exhaustive search, although alternative search techniques are likely applicable.

Once the input spaces are selected, the second step is to bridge the remaining distance between the real-world object and the fictional object by determining the set of iconic attributes that capture characteristic attributes of the fictional object. These will then be mapped back to the real-world object so that the agent can play with the real-world object as a placeholder for the fictional object. Iconic attributes of an object are those attributes that are most different from other objects under consideration; they capture which features are relevant to the pretend goal. The level of *iconicity* of an attribute for the desired pretend object is calculated as the sum of Euclidean distances between that attribute and the same attribute of all other objects in the fictional world. Iconicity values are normalized within categories of objects and an attribute is considered iconic for an object if it falls in the proximity of the maximum value.

In the third stage, blending occurs by projecting iconic values of the desired fictional object onto the selected real object. By default, the blend space contains all attributes of the real object. All iconic attributes of the pretend object are projected into the real object, replacing the original values. This process captures the notion that most action and reasoning should treat the real world as the base with the pretend domain layered onto this base.

In the pretend play algorithm, context and goals are utilized to filter the set of possible input spaces to only those that are most crucial for the use of a real-world object for pretend play. By pruning the set of possible objects according to their relevance to a goal, the process avoids the naïve consideration of all possible combinations of spaces to use for blending. Selective projection of attributes is achieved by searching for the most iconic attributes of the presentation input space (the fictional object) to be blended with attributes of the reference input space (the real-world object). While this search is performed in a brute-force manner, the number of attributes that cross the acceptability threshold for iconicity are relatively limited.

Discussion and Conclusions

As a powerful mechanism for creativity, conceptual blending is capable of synthesizing known concepts into new concepts. Much existing theoretical work focuses on the blending phenomenon and identifying the input spaces and the blend without a mechanism for blending. This paper presents our first efforts at building a complete computational outline for conceptual blending systems. We identify three major procedures in conceptual blending: (1) the selection of input spaces, (2) the mechanism for selective projection of input space attributes into the blend space, and (3) the sufficiency condition for pattern completion and elaboration. These components play vital roles in conceptual blending and have significant implications for the efficiency of computation. In our analysis of computational implementations of blending theory, we found few systems fully account for all three processes.

Brandt and Brandt (2002) argued that the construction of semiotic expressions as blends are cued by communication contexts and guided by the specific communicative goal. We argue that context and goals can provide the basis for rigorous and efficient computational algorithms for the three main processes described above. We present two computational systems that utilize goals and context to guide generation of standalone conceptual blends. The gadget generation algorithm mainly demonstrates procedures (2) and (3) by utilizing goals to select concepts from the input spaces and elaborating them. The pretend play work likewise uses context to determine which input spaces to select and which concepts from the input spaces to project into the blend, illustrating procedures (1) and (2).

These two case studies suggest the three main procedures can be implemented efficiently by employing constraints introduced by their respective domain of application, the contexts of the solutions, and the specific goals the solutions must achieve. Our analysis shows that goals can be used to prune the search space and improve average-case performance. Although our implementations are deterministic, we believe determinism and goals are not a bundled package. A goal-driven procedure may not be completely deterministic or even optimal. Future work is needed to reduce the effort required to author the knowledge representations used by our systems (e.g. with crowdsourcing (Li et al. 2012)).

Boden (2004) raised the questions of whether computers can appear to be creative, and whether computational

systems can help us understand creativity. By implementing theories of creativity, we are forced to consider procedural details which theories sometimes do not cover. We believe a computational approach can help expose underspecified components or flaws in existing theories, hint at their solution, or even lead to their remedy. A computational approach to creativity will strengthen our confidence in answering yes to both of Boden's questions.

Acknowledgement

This work was supported by the National Science Foundation under Grant No. IIS-1002748. We thank Per Aage Brandt and the anonymous reviewers for valuable inputs.

References

Boden, M. A. 2004. *The Creative Mind: Myths and mechanisms*. 2nd ed: Routledge.

Brandt, L., and Brandt, P. A. 2002. Making Sense of a blend. *Apparatur* 4:62-71.

Cavazza, M., Charles, F., and Mead, S. J. 2002. Planning Characters' Behavior in Interactive Storytelling. *Journal of Visualization and Computer Animation* 13 (2):121 - 131.

Fauconnier, G., and Turner, M. 1998. Conceptual Integration Networks. *Cognitive Science* 22 (2):133-187.

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

Forbus, K.D., Gentner, D., and Law, K. 1995. MAC/FAC: A model of similarity-based retrieval. *Cognitive Science* 19(2).

Fujio, F. F. 1974-1996. *Doraemon*. Vol. 1-45. Tokyo: Shogakukan.

Gervás, P., Díaz-Agudo, B., Peinado, F., Hervás, R. 2005. Story Plot Generation based on CBR. *Knowledge Based Systems*. 18.

Goguen, J., and Harrell, F. 2004. Foundations for Active Multimedia Narrative: Semiotic spaces and structural blending. *Interaction Studies: Social Behaviour and Communication in Biological and Articial Systems*.

Grady, J. 2000. Cognitive mechanisms of conceptual integration. *Cognitive Linguistics* 11 (3/4):335-345.

Grady, J., Oakley, T., and Coulson, S. 1999. Conceptual Blending and Metaphor. In *Metaphor in Cognitive Linguistics*, edited by R. Gibbs and G. Steen. Amsterdam: John Benjamins.Hervás, R., Pereira F. C., Gervás, P., Cardoso A. 2006. Cross-Domain Analogy in Automated Text Generation. In *3rd Joint Workshop on Computational Creativity at ECAI*.

Hutchins, E. 2005. Material Anchors for Conceptual Blends. *Journal of Pragmatics* 37:1555-1577.

Johnson-Laird, P. N. 2002. How Jazz Musicians Improvise. *Music Perception* 19 (3):415-442.

Jones, D.E. 2002. An Instinct for Dragons. Routledge.

Lakoff, G., and Johnson, M. 1980. *Metaphors We Live By*. Chicago: The University of Chicago Press.

Li, B., and Riedl, M. O. 2011a. Creative Gadget Design in Fictions: Generalized Planning in Analogical Spaces. In *8th ACM Conference of Cognition and Creativity*.

Li, B., and Riedl, M. O. 2011b. A Phone That Cures Your Flu: Generating Imaginary Gadgets in Fictions with Planning and Analogies. In *4th Workshop of Intelligent Narrative Technologies*. Palo Alto, CA: AAAI.

Li, B., Lee-Urban, S., Appling, D. S., Riedl, M. 2012. Automatically Learning to Tell Stories about Social Situations from the Crowd. In *the LREC 2012 Workshop on Computational Models of Narrative*.

Martinez, M., Besold, T., Abdel-Fattah, A., Kuehnberger, K.-U., Gust, H., Schmidt, M., and Krumnack, U. 2011. Towards a Domain-Independent Computational Framework for Theory Blending. In *The ACM Fall 2011 Symposium on Advances in Cognitive Systems*.

Nourot, P.M. 1998. Sociodramatic play: Pretending together. In *Play from Birth to Twelve and Beyond: Contexts, perspectives, and meanings.*

Ontañón, S., and Zhu, J. 2010. Story and Text Generation through Computational Analogy in the Riu System. In *7th AI and Interactive Digital Entertainment Conference*.

Pereira, F. C. 2007. *Creativity and AI: A Conceptual Blending Approach*. Berlin: Mouton de Gruyter.

Riedl, M. O., and Young, R. M. 2010. Narrative Planning: Balancing Plot and Character. *Journal of Artificial Intelligence Research* 39:217-268.

Rosch, E. 1978. Principles of Categorization. In *Cognition and Categorization*. Lawrence Erlbaum.

Thagard, P., and Stewart, T. C. 2011. The AHA! Experience: Creativity Through Emergent Binding in Neural Networks. *Cognitive Science* 35 (1):1-33.

Veale, T., and O'Donoghue, D. 2000. Computation and Blending. *Cognitive Linguistics* 11:253-281.

Weld, D. 1994. An Introduction to Least Commitment Planning. *AI Magazine* 15 (4):27-61.

Wolverton, M. 1994. Retrieving Semantically Distant Analogies. Ph.D. Dissertation, Stanford University.

Yamada, K., Taura, T., and Nagai, T. 2011. Design and Evaluation of Creative and Emotional Motion. In *8th ACM Conference on Creativity and Cognition*.

Zook, A. E., Riedl, M. O., and Magerko, B. S. 2011. Understanding Human Creativity for Computational Play. In *8th ACM Conference on Creativity and Cognition*.

A Creative Analogy Machine: Results and Challenges

Diarmuid P. O'Donoghue
Department of Computer Science,
National University of Ireland Maynooth,
Co. Kildare, Ireland.
diarmuid.odonoghue@nuim.ie

Mark T. Keane
Department of Computer Science and Informatics,
University College Dublin,
Ireland.
mark.keane@ucd.ie

Abstract

Are we any closer to creating an autonomous model of analogical reasoning that can generate new and creative analogical comparisons? A three-phase model of analogical reasoning is presented that encompasses the phases of *retrieval*, *mapping* and inference *validation*. The model of the retrieval phase maximizes its creativity by focusing on domain topology, combating the semantic locality suffered by other models. The *mapping* model builds on a standard model of the mapping phase, again making use of domain topology. A novel *validation* model helps ensure the quality of the inferences that are accepted by the model. We evaluated the ability of our tri-phase model to re-discover several *h-creative* analogies (Boden, 1992) from a background memory containing many potential source domains. The model successfully re-discovered all creative comparisons, even when given problem descriptions that more accurately reflect the original problem – rather than the standard (*post hoc*) representation of the analogy. Finally, some remaining challenges for a truly autonomous creative analogy machine are assessed.

Introduction

Analogy has a long and illustrious history within creativity, particularly within scientific and intellectual contexts (Brown, 2003). Many episodes of scientific creativity are driven by analogical comparisons (Dunbar and Blanchette, 2001), often involving image related analogies (Clement, 2008). Much progress has been made in cognitive science on modeling this analogical reasoning process (see below), prompting the following questions. Are we any closer to creating an autonomous model of the analogical reasoning that can generate new creative analogies? What progress has been made towards such a creative analogy model? What are the main challenges that lie ahead?

In this paper we envisage a creative process that can take any given target description and using a pre-stored collection of domain descriptions, identify potentially creative source domains with which to re-interpret the given problem. This paper explores and evaluates the potential for a model of analogy to act as a creativity engine.

While Boden (1992) argues that analogy is effectively the lowest form of creativity (*improbable*), we argue that analogical creativity should be seen a part of a cohesive human reasoning system. If the inferences mandated by an analogy contradicts a fundamental belief, especially one that has accrued many consequent implications, then resolving this contradiction might well involve the *"shock and amazement"* of transformational creativity. As such, it appears that analogies may drive creativity at any of Boden's levels of creativity. Our creativity model is domain independent and does not include a pragmatic component or domain context. So, as our model does not use domain-specific knowledge, arguably it cannot be easily cast as *improbable*, *exploratory* or *transformational* creativity (Boden, 1992).

The current work was driven by three main aims. Firstly, we wished to assess the creative potential of a three-phase model of analogy. Secondly, we wished to assess the impact of using differing knowledge bases upon the creative potential of our analogy model. Finally, we wished to assess the wider implications of analogical models for computational creativity. Is a three-phase model either necessary or sufficient to function as an engine of creativity? Can such a model re-discover analogies considered to be creative by people? Since people often overlook analogies (Gick and Holyoak, 1980) even when they are present, will such a model uncover many creative analogies or are creative analogies, in some way, different and rare?

We see the current model as being potentially useful in three distinct ways, but for now we do not commit to using it in one particular manner. Firstly, it could be used as a simple model of creativity, yielding creative interpretations for a presented problem. Secondly, it could be used as a tool to assist human creativity; suggesting source domains to people, to enable them to re-interpret a given target problem. Finally, it could be used as one possible model of how people analogize in a creative means.

The paper is structured as follows: first we describe the Kilaza[1] model for generating creative analogies, briefly illustrating its operation on the famous *atom:solar-system*

[1] Kilaza is not an acronym.

analogy. Then we present results that reflect the model's ability to re-generate some well-known *h-creative* analogies (Boden, 1992). Finally, the implications of these results are assessed and some remaining challenges are discussed.

Analogy as an Engine of Creativity

An analogy is a conceptual comparison between two collections of concepts, a *source* and *target* (Gentner, 1983), such that the source highlights particular aspects of the target, possibly suggesting some new inferences about it. In creative analogies, an productive source domain conjures up a new and revolutionary interpretations of the target domain, triggering novel inferences that help explain some previously incongruous phenomena or that help integrate some seemingly unrelated phenomena (Boden, 1992; Eysenck and Keane, 1995). Creative analogies differ from "ordinary" analogies primarily in the conceptual "distance" between the source and target domains (i.e., these two domains may never have been linked before) and the usefulness of the resulting comparison. Both creative and mundane analogies appear to use the same analogical reasoning process, as described in the following section, but different in their inputs and outputs.

Kekulé's is famous for his analogy between the carbon-chain and a snake biting its own tail. But this analogy could have been triggered by many alternative and more mundane source domains – from tying his own shoe-lace to buckling his belt. While many source domains could have generated the creative carbon-ring structure, Gick and Holyoak (1980) have shown most people (including Kekulé) frequently fail to notice many potential analogies. This highlights one potential advantage of a computational model, in that a model can tirelessly explore all potential analogies, returning only the most promising comparisons to a user for more detailed consideration. Thus, computational models could potentially act a tools helping people overcome one barrier; namely, their failure to perceive analogies when they are present.

Kilaza Analogical Creativity Engine

Keane (1994) presented a five-phase model of the analogical reasoning process, which recognises the distinct phases of *representation*, *retrieval*, *mapping*, *validation* and *induction*. While other authors describe slightly different subdivisions of this process, there is broad agreement on these phases. Our computational model encompasses the three central phases of analogy (see Figure 1). We highlight that Walls & Hadamard subdivide creativity into the phases of *preparation, incubation, illumination* and *verification* (Boden, 1992), which is reminiscent of several multi-phase models of analogy.

The heart of our creativity model is the central mapping phase and this borrows heavily from Keane and Brayshaw's (1988) IAM model (see also Keane, Ledgeway & Duff, 1994). Our model of the retrieval phase attempts to overcome the semantic bias suffered by many previous models, improving the diversity of the source domains that are returned. It was intended that this diversity might address the quality of *novelty* (Ritchie, 2001) associated with creativity, retrieving more "unexpected" and potentially creative sources. Finally, our model of the validation phase attempts to filter out invalid inferences, addressing the *quality* (Ritchie, 2001) factor associated with computational creativity.

Ritchie (2001) identifies the essential properties of creativity as being *directed*, *novel* and *useful*. We argue that our model is *directed* in that it focuses on re-interpreting some given target domain. Our model addresses the *novelty* property by its ability to retrieve potentially useful but semantically distant, even disconnected, source domains. Finally, the *useful* property is addressed through a validation process that imposes a quality measure on the inferences that are accepted by the model.

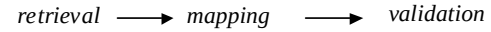
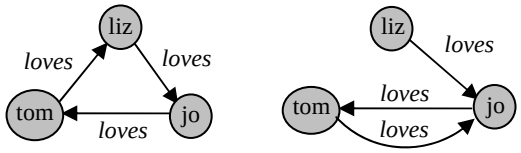


Figure 1: Kilaza is a three-phase model of Analogy

Analogical Retrieval Phase-Model

Existing models for analogical retrieval suffer from the limitations in the range of possible retrievals because their they either (i) focus exclusively on domain semantics (like MAC/FAC; Forbus, Gentner and Law, 1995) or (ii) focus primarily on domain semantics (like HRR; Plate, 1998). Other models -- such as ARCS (Thagard *et al*, 1990) and Rebuilder (Gomes *et al*, 2006) - supplement domain representations by elaboration from external sources (like WordNet) to widen the net to include more semantically non-identical sources. However, all of these approaches arguably over-constraint retrieval for the the proposes of creativity. We argue that a creative retrieval process must allow semantically distant and even semantically disconnected sources to be retrieved, ideally without overwhelming the subsequent phase-models with irrelevant domains.



(a) Love Triangle (b) Unrequited Love

Figure 2: Topology is a key characteristic in retrieving creative source domains

Gentner (1983) mentions two specific qualities are required of analogical comparisons: semantic similarity and structural similarity. The model presented in this paper performs retrieval based exclusively on structural similarity, performing retrieval based exclusively on the graph structure (or topology) of each domain description. This design decision was taken to overcome the semantic narrowness that constrains existing models, with the hope that

this would increase the possibility of retrieving surprising and creative source domains. As the example in Figure 2 illustrates, semantics and domain topology are often intertwined.

Each domain description is mapped onto a location in an *n*-dimensional structure space (Figure 3), where each dimension represents a particular topological quality of that domain. Structure space is somewhat akin to feature vectors (Yanner and Goel, 2006; Davies, Goel and Yanner, 2008). Image related analogies are often involved in creative comparisons (Clement, 2008) and a variety of image-based analogy models has been developed, focusing on specific topics such as; geometric proportional (IQ type) analogies (Evans, 1967; Bohan and O'Donoghue, 2000), geo-spatial comparisons (O'Donoghue *et al*, 2006), spatial representations of conceptual analogies (Davies et al, 2008; Yanner *et al*, 2008) and reasoning about sketch diagrams (Forbus *et al*, 2011). Our model performs a single retrieval process for each presented target, in contrast to the iterative retrieval and spreading activation phases employed by KDSA to retrieve semantically distant sources (Wolverton and Hayes-Roth, 1994).

Specific topological features used by our retrieval model include quantifying the number of objects and predicates (first order and higher order) and number of root predicates *etc*. Thus, the representation in Figure 4 might be mapped onto the location (4 0 2 2 0 0 1) in structure space – 4 object references, 0 high-order predicates, 2 unique first-order relations, 2 first-order relations and 2 root predicates *etc*. The distinction between unique and non-unique relations, for example, distinguishes between domains repeatedly using a small number of relations and domains that typically have one instance of each relation in its description. One advantage of this scheme is that the distance between domains is not impacted by the number of domains contained in memory so the retrieval system should scale reasonably well. For the retrieval results presented later in this paper a maximum retrieval distance of 10 is imposed – and only candidate source inside this threshold are considered.

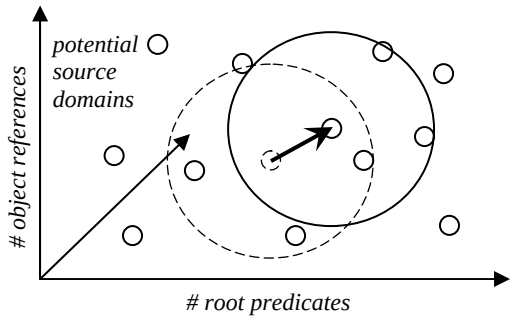


Figure 3: Displacing the Locus of Retrieval within a 3D representation of *n*-dimensional Structure Space. Only source domains within the displaced boundary are retrieved and passed to the remaining phases of analogy.

Topologically similar (i.e., homomorphic as well as isomorphic) domains are mapped onto similar locations within this topology-based structure space (O'Donoghue and Crean, 2002). To account for the inferences that were sought from any inspiring source domain, the locus of retrieval was slightly offset to account for this additional source domain material. Included in this offset is the desire for sources containing additional first-order relations and high-order relations. However, this offset has relatively little impact on the final results.

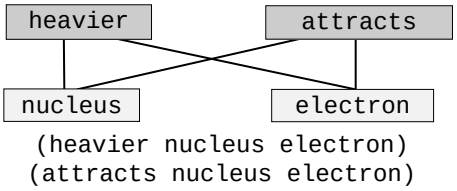


Figure 4: Simplified Model of Rutherford's Problem

Analogical Mapping Phase-Model

The model for the mapping phase is based on the Incremental Analogy Machine (IAM) model (Keane & Bradshaw, 1988; Keane et al, 1994). It consists of the three sub-processes of *root-selection*, *root-elaboration* and *inference generation*. Mapping proceeds as a sequence of root-selection and root-elaboration activities, gradually building up a single inter-domain mapping. Typically a domain description will consist of a small number of root predicates, each controlling a large number of (partly overlapping) lower-order predicates.

Root selection Root selection identifies "root predicates" within a representation, which are typically the controlling causal relations in that domain. Each root predicate lies at the root of a tree of predicates and each root is be seen as "controlling" the relations lower down the tree. In our implementation of IAM, the root-selection process examines the "order" of each predicate. Objects are defined as order zero and first-order relations that connect two objects are defined as order one. The order of a causal relation is defined as one plus the maximum order of its arguments. Mapping begins with the highest order relations and maps any unmapped low-order root-predicates last.

Root elaboration Root elaboration extends each root-mapping, placing the corresponding arguments of these relations in alignment. If these arguments are themselves relations, then their arguments are mapped in turn and so on until object arguments are mapped. Items are only added to the inter-domain mapping when they conform to the 1-to-1 mapping constraint (Gentner, 1983).

Inference Generation Each analogical comparison is passed to the inference generation sub-process. Analogical inferences are generated using the standard algorithm for pattern completion CWSG – Copy With Substitution and

Generation (Holyoak *et al*, 1994). In effect, additional information contained in the source domain is carried over to the target, creating a more cohesive understanding of that target problem.

Analogical Validation Phase-Model

The third part of our tri-phase model is focused on analogical validation. Validation attempts to ensure that the analogical inferences that are produced are correct and useful.

O'Donoghue (2007) discusses the accuracy of this validation process, using human raters to assess the goodness of inferences that were rated as either valid or invalid. However, this paper did not assess the models ability to discover creative analogies.

Phineas (Falkenhainer, 1990) is a multi-phase model of analogy that incorporates a post-mapping verification process. To achieve this Phineas incorporates a model of the target domain – qualitative physics simulation - illustrating the power of embedding an analogy model within a specific problem domain. However, this qualitative-simulation process effectively limits Phineas to reasoning only about physical and physics-related analogies.

The validation model presented in this paper is relatively simple, aimed at rejecting those predicates that are deemed invalid – rather than guaranteeing the validity of those inferences that are accepted. This approach helped maximise the creative potential of this model, by resisting the rejection of potentially plausible inferences. Of course, a more complex validation process could make use of problem-specific domain knowledge (where available). In the absence of such domain-specific knowledge verification and validation of the analogy could be carried using user feedback, employing Kilaza in a tool-like way.

The validation phase-model is composed to two main parts. The first performs validation by comparing the newly generated inference to predicates already stored somewhere in memory. The second mode of validation is more general and driven in part by the functionally relevant attributes that play a role in analogical inference (Keane, 1985).

Validation by Predicate Comparison The validation process compares newly inferred predicates (produced by CWSG) to the previous contents of memory. Inferences are firstly compared to predicates in memory, with both the agent and patient roles potentially being validated independently. This validation mechanism thus has access to the entire contents of memory, accessing predicates from any of the domains stored in that memory. This model of validation captures the advantages of simplicity and generality, but it does of course mean that dependencies between arguments are not captured. This limitation was deemed acceptable within the context of our desire for a creativity engine. While many simple inferences were validated by this mechanism, many creative inferences were not. This may be partly attributed to the relatively small number of predicates contained in memory and to the novelty associated with creative inferences. To address this shortcoming validation using functional attributes was introduced.

Validation with Functional Attributes Functional attributes specify necessary attribute requirements for each role of a predicate – being inspired by the *functionally relevant attributes* of Keane (1985). Functional attributes are intra-predicate constraints that ensure each predicate appears to be a plausible combination of a relation coupled with each of its arguments.

It should be pointed out that functional attributes have only been used with first-order predicates – those whose arguments are objects. Although validating higher-order (causal) relations might make use of the spatio-temporal contiguity associated with causality, but this cannot be relied upon (Pazzani, 1991) and is not enforced by our model. Thus our model treats all causal inferences as implicitly valid.

Functional attribute definitions connect each role of a predicate directly into an attribute hierarchy, whereby arguments filling those roles must conform to these attribute constraints. Kilaza stores functional attributes for both the agent and patient arguments of each relation independently. More general relations (*part-of, next-to*) typically have few functional attributes, whereas more specific relations (*hit, eat*) possess a greater number of attribute restrictions. For example the agent role of *hit* might require the hitter to be a physical object, whereas the agent of an *eat* relation might have to be a living organism or an animal. Relations that are more specific are seen to be more amenable to the validation process, while their more general counterparts are more difficult to validate accurately.

In addition, functional attributes have also been used to support a form of inference adaptation. This allows an inferred relation to be adapted to a semantically similar relation that better suits the arguments that pre-existed within in the target domain. Adaptation uses the functional attributes to conduct a local search of the taxonomy, to identify a more semantically suitable relation that better fits the given arguments.

Data Sets

Three datasets were used to conduct experiments using the described model. These are referred to as the Professions dataset, the Assorted dataset and an Alphanumeric dataset. The dataset contained a total of 158 domains and our creativity engine attempted to find creative source analogues for a given number of target problems. It was hoped that the differing natures of these collections would provide a reasonable grounds on which to evaluate the computational model – and to assess its potential to act as a creativity engine.

Professions Dataset consists of descriptions of fourteen professions, including *accountant, butcher, priest* and *scientist*. These are rather large domain descriptions created by Veale (1995) and range in size from 10 to 105 predicates (M=55.4, SD=29.3). One important feature of the Professions dataset is its reliance on many different instances of a small number of relational predicates, including *control, affect, depend,* and *part*. The domains range from using just 6 distinct relational predicates (ignoring

duplicates) to the most diverse domain that uses 15 (M=8.9, SD=2.2). Another important feature is that this dataset does not appear to use a set of clearly identifiable high-order relations (such as a *cause, result-in* or *inhibit*) between first-order predicates.

Assorted Dataset consists of a large number of smaller and more varied domain descriptions, including many of the frequently referenced domains in the analogy literature; such as the *solar-system*, *atom*, *heat-flow* and *water-flow* domains. It also includes an assortment of other domains describing *golf*, *soccer* and *story-telling*. The 81 domains of the Assorted dataset use 108 distinct (*ie* non-repeated) relations. Each of these domains contains between 1 and 15 predicates (M=4.16, SD = 2.9). The average number of distinct relational predicates in each domain is M=3.48, indicating that most relational predicates are used just once in each of the Assorted domains.

Alphanumeric Dataset One final dataset contained 62 semantically constrained domains. However, these domains contained a great deal of topological diversity. It was hoped that this mixture of topologies might support some novel comparisons and inferences and provided a counterpoint to the semantic richness of the other domains.

Example: p-Creative Re-Discovery of Rutherford's Analogy

Before presenting detailed results, we will first see how Kilaza can re-discover Rutherford's famous *solar-system:atom* analogy. We highlight that this is a test for the p-creativity (Boden 1992) of our model – though not necessarily a model how Ernest Rutherford actually conducted his own reasoning.

The traditional representation of this analogy (Figure 5) is heavily based on a *post hoc* description of the domains involved. These descriptions and are heavily influenced by the analogy itself. We shall first look at the traditional representation of this domain, before examining how our model can also deal with more realistic version of how Rutherford might have thought of the target problem *before* arriving at his famous comparison.

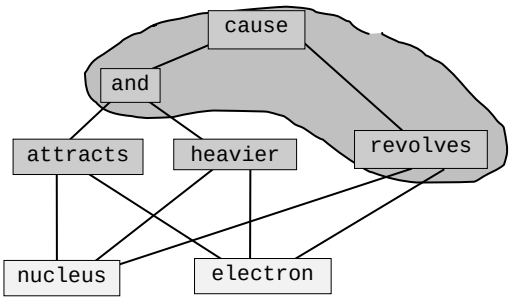


Figure 5: Traditional representation of Rutherford's Solution

First, the semantically impoverished target problem (Figure 4) is mapped onto its location in structure space. We highlight that the "locus of retrieval" is slightly displaced from the targets original location to account for the additional information that one expects to be found in a useful source domain. In this instance the desired source was retrieved at a distance of just over 6 "units" in structure space. The desired source (the *solar-system* domain) and all other candidate sources near the locus of retrieval were passed in turn to the mapping and validation phases of the model.

In total 10 other candidate source domains that were retrieved also generated inferences, most yielding only one inference each. Three domains generated more than one candidate inference – but all three were different versions of the *solar-system* domain. We point out that our semantic "free" retrieval process can also trigger identification of the same source, even if it was represented in a number of alternate ways (O'Donoghue, 2007). Our mapping model successfully generated the correct inter-domain mapping and CWSG generated the desired inferences without adaptation.

Representation Issues in *de novo* Discovery of p-creative analogies

We argue that the traditional presentation of Rutherford's analogy is a simplified pedagogical device (Figure 5). This description of the target problem effectively removes much of the complexity of the real discovery task as encountered by Rutherford. The description of the target problem uses terminology specifically designed to accentuate the semantic (and structural) similarity that is the *result* of Rutherford's comparison – and should not be treated as an input when re-creating this creative episode.

This distinction between the problem domain as it would have existed *before* the creative analogy and its subsequent representation *after* discovering that analogy is a serious problem - one that is easily overlooked. Any model that attempts to re-discover known creative analogies must address the original problem, not just the representation that accentuates the desired similarity. Differences in domain terminology and topology are central to the distinction between elaborating a given analogy, and the much more difficult task of generating a novel *h-creative* (or *p-creative*) analogy (Boden, 1992).

We argue that generating Rutherford's analogy using the representation in Figure 6 is a far better test of a models creative ability, than the normal *post hoc* representation in Figure 5. Terminological differences are particularly prevalent in distant between-domains analogies as the first-order relationships describing the problem domains originate in different disciplines. When modeling analogical creativity, we must expect to encounter these differences in terminology, and our models of retrieval, mapping and validation must be able to overcome these problems.

Ernest Rutherford would most likely have thought of the target relation between the nucleus and electron as `electromagnetic-attraction`, and not the more generic `attracts` relation. The corresponding relationship between source's sun and planet is `gravitation`. It is only after he found the analogy (which involved mapping

`electromagnetic-attraction` with `gravitation`) that these relationships can be generalized to a common super-class like `attracts` (Gentner, 1983).

We point out that our model can operate successfully on either the simplified or more realistic domain descriptions. This is primarily the result of our retrieval and mapping models using domain topology, rather than using identicality (or similarity) between the predicates in both domains.

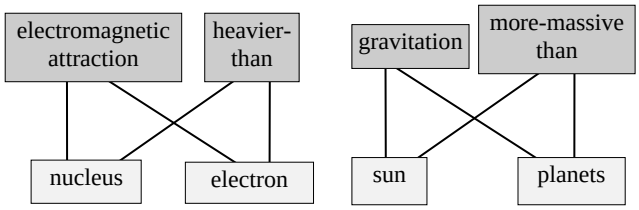


Figure 6: More realistic representation of Rutherford's Analogy

Results of Individual Phase Models

We shall first briefly examine the performance of the *retrieval* and *validation* models in isolation, before looking at their combined performance in the next section. We shall briefly examine the results of the *mapping* model, but our focus will remain on the inferences that it produced. Results were produced from a memory containing the three previously described datasets.

Retrieval Results Retrieval was performed in structure space. The distance between domains in structure space varied from 2.645 to 230 (M= 80, SD=57.3), with a large number of domains being given a unique structural index in this space. A small number of locations contained multiple domains – these mostly involved small domains of just a few predicates from the Assorted dataset.

Retrieval and Mapping A broad tendency was identified between structure-based retrieval and the size of the resulting inter-domain mapping, although the correlation was low. A range effect was identified between structure space and the size of the resulting mapping, indicating that larger distances between domains in structure space tend to produce smaller inter-domain mappings. This indicates a weak connection between structure-based retrieval and the size of any resulting mappings.

Validation Results Although the validation model was very simplistic, it proved surprisingly effective. For example with the inferences generated on the Professions dataset, the average (human) rating awarded to predicates that Kilaza categorized as valid was M=2.62 (SD=2.09), while the average rating awarded to the invalid predicates was M=1.57 (SD=1.23). As ratings were given between 1 and 7 with 7 representing clearly valid inferences, this indicates that many of the generated inferences were of rather poor quality.

Adaptation Results In addition, 24 inferences were passed to the adaptation process and 20 of these were adapted. While we *cannot* realistically assess if these adapted inferences matched what was "intended" by our analogy model, we did assess the validity of these inferences using two human raters.

When we look at human ratings for the 20 adapted predicates before and after adaptation, we see that the average ratings were increased by the adaptation process - from 1.57 (SD=1.23) to 2.57 (SD=1.70). The average ratings of the adapted predicates was broadly in line with the predicates from Kilaza's valid category above (M=2.62, SD=2.09). Before adaptation, 18 of the 20 (90%) predicates were given rated as invalid and after adaptation just 12 (60%) were rated as *invalid*. Thus, adaptation has a distinct influence on improving the ratings of the rejected inferences.

It may well be argued that this adaptation process is itself somewhat creative – identifying new relations that better fit the available target arguments. In contrast to the top-down nature of the creative analogy approach, predicate adaption is a very much a bottom-up process that is motivated by the detection of a potential analogical comparison.

Creativity Test Results

To assess the creative potential of our model, we assess its performance at the *p-creative* task of re-creating some well-known *h-creative* analogies (Boden, 1992). These include some of the famous examples of creative analogical comparisons including the Rutherford's *solar-system:atom* analogy, the *heat-flow:water-flow* and the *tumour:fortress* analogies. Our descriptions are based on the standard representation of these domains as found in the analogy literature.

Creative Retrieval We now examine the performance of our model on the creative retrieval task. We presented our model with the target domain of each of 10 creative analogies, together with a memory of 158 source domains. From this memory of 158 potential sources, the retrieval model selected a number of these domains as candidate sources. Only the selected candidate sources were passed to the mapping and validation phase-models. Evaluating only the selected source domains was necessary in order to avoid an exhaustive search through all possible analogical comparisons. While computationally feasibly in this instance, an exhaustive search would be impractical on a larger collection of domains.

Before looking at the results, we point out that many comparisons did not generate a viable inter-domain mapping. Furthermore, most analogies did not generate any valid inferences. The following results ignore these unproductive comparisons and we focus only on the productive analogies.

All of the desired creative sources were among the candidate sources that were retrieved by the model. This gives

our retrieval model a *recall* value of 100% on this creative retrieval task. While a large number of other candidate sources were also retrieved, this was still a pleasantly surprising result. The distance within structure space between the target and the creative sources ranged from 3.1 to 7.9, suggesting that structure based retrieval was reasonably accurate in locating candidate sources.

The *precision* of the retrieval processing is summarised in Figure 7. As can be seen, precision was above 0.2 for two problems showing that few other sources were located near the structural index of those targets. However, precision was much lower for most problems, indicating that the desired source was merely one of a larger number of candidate sources that had to be explored.

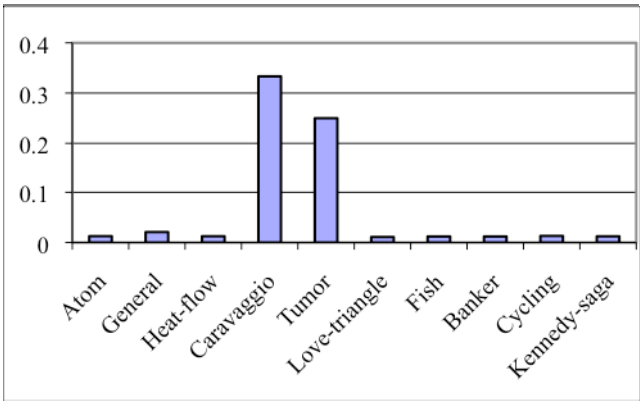


Figure 7 –*Precision of retrieval for 10 Creative Analogies*

Creative Inferences Next we summarise the inferences that were generated by each of these comparisons (Table 1). These results implicitly encompass a productive inter-domain mapping between the target and each candidate source in turn. Kilaza generated and validated the correct inferences for 9 (70%) of the creative analogies. The *cycling:driving* analogy correctly generated no inferences.

Target	Correct Inferences	Validated Inferences
Atom: Solar-System	y	4
Atom-Falkenhainer: Solar-System-Falkenhainer	y	3
General: Surgeon	y	4
Heat-flow: Water-Flow	y	4
Leadbelly : Caravaggio	y	4
Love-triangle: Triangle-Directed	y	0
Requited-love: Love Triangle	y	3
Fish : Bird	y	4
Vampire : Banker	y	3
Cycling: Driving	n	0

Table 1 – Number of Inferences generated by different analogies

One of these analogies also required one inference to be adapted. The `bird:fish` analogy generated the inference (`flies-through fish water`), which was correctly adapted to (`swim fish water`).

Conclusion

We presented a three-phase model of analogy, adapting it to function as a tool for discovering creative analogies. This model encompasses the three central phases of analogy, namely *retrieval*, *mapping* and *validation*. We argue that a model encompassing these three core phases of analogy is the minimum required to be considered a model of analogical creativity.

Our *retrieval* model overcomes the semantic bias of previous retrieval models, helping retrieve new and surprising source domains. This helps to improve the *novelty* of the source domains identified by our creativity engine. Our model of the post-mapping *validation* phase attempts to filter out any clearly invalid inferences, thereby improving the *quality* of the analogies identified as being creative. We note that novelty and quality are two attributes strongly associated with creativity (Ritchie, 2001).

Our three-phase model of analogy successfully rediscovered 10 examples of creative analogies, including the *heat-flow:water-flow* and *solar-system-atom* analogies. In doing so, the model retrieved the correct source from a large memory of potential sources. It then developed the correct mapping and successfully validated (and adapted) the resulting inferences. We point out that these analogical comparisons, if produced by a human analogizer, would be considered creative.

Our focus on creative analogies rather that the more normal (or pedagogical) analogies had a far-reaching impact on the model. Terminological differences are particularly prevalent in creative between-domains analogies, as the first-order relations describing each domain originate in different disciplines. When modeling analogical creativity, we must expect to encounter these differences and cannot rely heavily on the presence of identical relations. Our model successfully created Rutherford's famous *solar-system:atom* analogy, even when the target was represented in a more realistic and challenging from. Our model shows that very significant progress has been made towards an autonomous creativity machine, re-discovering many creative analogies.

We briefly outline three remaining challenges to analogical creativity, beginning with the issue of knowledge representation. Our results illustrate a trade-off between the *specifity* and the *generality* of domain descriptions. Overly specific representations make comparisons more difficult to discover, but overly general representations appear too profligate and can overwhelm the validation (and subsequent) processes. Perhaps multiple representations of each domain might offer a useful avenue for progress. Multiple representations might also help explain why exerts are more fluent in their use of analogy within their own domains (Dunbar and Blanchette, 2001). Our model does not currently include an explicit re-representation process

highlighting "*tiered identicality*" (Gentner and Kurtz, 2006).

It seems that the greatest challenge to computational analogizing might lie with the post-mapping phases. Challenges include assessing analogical inferences for validity, evaluating the significance of an analogy and considering the implications of creative comparisons. Surprisingly little attention has been given to this phase – partly because of its ultimate dependency on the target problem domain. Phineas (Falkenhainer, 1990) and also Rebuilder (Gomes *et al*, 2006) showed that integration of the analogy and case-based reasoning within the target domain can have very positive effects. While tight integration of all target domains into an analogy model seems most unlikely, Kilaza has show that a generic validation model can play a part improving the quality of the inferences that are accepted.

Overall, the results presented in this paper highlight that a three-phase model of analogical reasoning can operate successfully as a model of analogical creativity. Our results highlight the improbability of finding a suitable source domain to re-interpret a given target in a creative manner. Extending this model will necessitate a tighter integration of the analogy process with other facets of intelligence.

References

Boden, M. 1992. *The Creative Mind*. London: Abacus.

Bohan, A. O'Donoghue, D.P. 2000. A Model for Geometric Analogies using Attribute Matching, *Proc AICS*, Galway, Ireland, pp 110-119.

Brown, T.L. 2003. *Making Truth: Metaphor in Science*. University of Illinois Press, New York, USA.

Clement, J. 2008. *Creative Model Construction in Scientists and Students: The Role of Imagery, Analogy, and Mental Simulation*. Dordrecht: Springer.

Davies, J. Goel, A.K. Yaner P.W. 2008. Proteus:Visuospatial analogy in problem-solving, *Knowledge-Based Systems* 21(7): 636-654.

Dunbar, K. and Blanchette, I. 2001. The in vivo/in vitro approach to cognition: The case of analogy, *Trends in Cognitive Sciences*, 5(8): 334–339.

Falkenhainer, B. 1990 *A Unified Approach to Explanation and Theory Formation*, in Shrager, J. & Langley, P. (eds.) Computational Models of Scientific Discovery and Theory Formation, Morgan Kaufman: CA. pp 157-196.

Forbus, K. Gentner, D. Law K. 1995. MAC/FAC: A Model of Similarity-based Retrieval, *Cognitive Science*, 19(2): 141-205.

Forbus, K. Usher, J. Lovett, A Lockwood, K. Wetzel J. 2011. CogSketch: Sketch Understanding for Cognitive Science Research and for Education, *Topics in Cognitive Science*, 3(4): 648–666.

Gentner, D. Kurtz, J. 2006. Relations, Objects, and the Composition of Analogies, *Cognitive Science*, 30(4): 609–642.

Gentner, D. 1983. Structure-mapping: A theoretical framework for analogy, *Cognitive Science*, 7(2): 155–170.

Gomes P, Seco N, Pereira FC, Paiva P, Carreiro P, Ferreira JL, Bento C. 2006. The importance of retrieval in creative design analogies, *Knowledge-Based Systems*, 19(7): 480–488.

Holyoak K. J. Novick L. Melz E. 1994. *Component Processes in Analogical Transfer: Mapping, Pattern Completion and Adaptation*, in Analogy, Metaphor and Reminding, Eds. Barnden and Holyoak, Ablex, Norwood, NJ.

Keane, M.T. 1985. On Drawing Analogies When Solving Problem, *British Journal of Psychology*, 76: 449-458.

Keane, M.T., Bradshaw, M. 1988. The Incremental Analogical machine: In D. Sleeman (Ed.) *3rd European Working Session on Machine Learning*, Kaufmann CA: 53–62.

Keane, M.T., Ledgeway, T. Duff, S. (1994). Constraints on analogical mapping, *Cognitive Science*, 18: 387-438.

O'Donoghue, D.P. 2007. Statistical Evaluation of Process-Centric Computational Creativity: *4th International Joint Workshop on Computational Creativity (IJWCC)*, Goldsmiths, University of London, 17-19 June.

O'Donoghue, D.P, Bohan, A. Keane M.T, 2006. Seeing Things: Inventive Reasoning with Geometric Analogies and Topographic Maps, *New Generation Computing*, Ohmsha Ltd. and Springer 24(3): 267-288.

O'Donoghue, D.P. Crean, B. 2002. Searching for Serendipitous Analogies, *ECAI - Workshop on Creative Systems*, Lyon, France, 21-26 July.

Pazzani, M. 1991. A Computational Theory of Learning Causal Relationships, *Cognitive Science*, 15(3): 401-424.

Plate T. 1998. *Structured Operations with Distributed Vector Representations*, in Advances in Analogy Research, New Bulgarian University, Sofia, Bulgaria.

Ritchie G. 2007. Some Empirical Criteria for Attributing Creativity to a Computer Program, *Minds & Machines*, 17: 67–99.

Ritchie G. 2001. Assessing Creativity, *Proc. AISB Symposium on AI and Creativity*, York, March.

Thagard, P. Holyoak K. J. Nelson, G. Gochfeld, D. 1990. Analogue Retrieval by Constraint Satisfaction, *Artificial Intelligence*, 46: 259-10.

Veale, T. 1995. *Metaphor, Memory and Meaning*, PhD Thesis, Trinity College, Dublin, Ireland.

Wolverton M. Hayes-Roth, B. 1994. Retrieving Semantically Distant Analogies with Knowledge-Directed Spreading Activation, *Proceedings AAAI*, pp 56-61.

Yaner, P.W. Goel A.K. 2006 Visual analogy: Viewing analogical retrieval and mapping as constraint satisfaction problems, *Applied Intelligence*, 25(1): 91-105.

Automated Generation of Cross-Domain Analogies via Evolutionary Computation

Atılım Güneş Baydin[1,2], **Ramon López de Mántaras**[1], **Santiago Ontañón**[3]
[1]Artificial Intelligence Research Institute, IIIA - CSIC
Campus Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain
[2]Departament d'Enginyeria de la Informació i de les Comunicacions
Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain
[3]Department of Computer Science, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA
gunesbaydin@iiia.csic.es, mantaras@iiia.csic.es, santi@cs.drexel.edu

Abstract

Analogy plays an important role in creativity, and is extensively used in science as well as art. In this paper we introduce a technique for the automated generation of cross-domain analogies based on a novel evolutionary algorithm (EA). Unlike existing work in computational analogy-making restricted to creating analogies between two given cases, our approach, for a given case, is capable of creating an analogy along with the novel analogous case itself. Our algorithm is based on the concept of "memes", which are units of culture, or knowledge, undergoing variation and selection under a fitness measure, and represents evolving pieces of knowledge as semantic networks. Using a fitness function based on Gentner's structure mapping theory of analogies, we demonstrate the feasibility of spontaneously generating semantic networks that are analogous to a given base network.

Introduction

In simplest terms, analogy is the transfer of information from a known subject (the *analogue* or *base*) onto another particular subject (the *target*), on the basis of similarity. The cognitive process of analogy is considered at the heart of many defining aspects of human intellectual capacity, including problem solving, perception, memory, and creativity (Holyoak and Thagard 1996); and it has been even argued, by Hofstadter (2001), that analogy is "the core of cognition".

Analogy-making ability is extensively linked with creative thought (Hofstadter 1995; Holyoak and Thagard 1996; Ward, Smith, and Vaid 2001; Boden 2004) and plays a fundamental role in discoveries and changes of knowledge in arts as well as science, with key examples such as Johannes Kepler's explanation of the laws of heliocentric planetary motion with an analogy to light radiating from the Sun[1] (Gentner and Markman 1997); or Ernest Rutherford's analogy between the atom and the Solar System[2] (Falkenhainer, Forbus, and Gentner 1989). Boden (2004; 2009) classifies analogy as a form of *combinational creativity*, noting that it works by producing unfamiliar combinations of familiar ideas.

[1]Kepler argued, as light can travel undetectably on its way between the source and destination, and yet illuminate the destination, so can motive force be undetectable on its way from the Sun to planet, yet affect planet's motion.

[2]The Rutherford–Bohr model of the atom considers electrons to circle the nucleus in orbits like planets around the Sun, with electrostatic forces providing attraction, rather than gravity.

In this paper, we present a technique for the automated generation of cross-domain analogies using evolutionary computation. Existing research on computational analogy is virtually restricted to the discovery and assessment of analogies between a given pair of base case A and target case B (French 2002) (An exception is the Kilaza model by O'Donoghue (2004)). On the other hand, given a base case A, the approach that we present here is capable of creating a novel analogous case B itself, along with the analogical mapping between A and B. This capability of open-ended creation of novel analogous cases is, to our knowledge, the first of its kind and makes our approach highly relevant from a computational creativity perspective. It replicates the psychological observation that an analogy is not always simply "recognized" between an original case and a retrieved analogous case, but the analogous case can sometimes be created together with the analogy (Clement 1988).

As the core of our approach, we introduce a novel evolutionary algorithm (EA) based on the concept of "meme" (Dawkins 1989), where the individuals forming the population represent units of culture, or knowledge, that are undergoing variation, transmission, and selection. We represent individuals as simple semantic networks that are directed graphs of concepts and binary relations (Sowa 1991). These go through variation by memetic versions of EA crossover and mutation, which we adapt to work on semantic networks, utilizing the commonsense knowledge bases of ConceptNet (Havasi, Speer, and Alonso 2007) and WordNet (Fellbaum 1998). Defining a memetic fitness measure using analogical similarity from Gentner's psychological structure mapping theory (Gentner and Markman 1997), we demonstrate the feasibility of generating semantic networks that are analogous to a given base network.

In this introductory work, we focus on the evolution of analogies using a memetic fitness function promoting analogies. But it is of note that considering different possible fitness measures, the proposed representation and algorithm can serve as a generic tool for the generation of pieces of knowledge with any desired property that is a quantifiable function of the represented knowledge. Our algorithm can also act as a computational model for experimenting with memetic theories of knowledge, such as evolutionary epistemology and cultural selection theory.

After a review of existing research in analogy, evolution, and creativity, the paper introduces details of our algorithm. We then present results and discussion of using the fitness function based on analogical similarity, and conclude with future work and potential applications in creativity.

Background

Analogy

Analogical reasoning has been actively studied from both cognitive and computational perspectives. The dominant school of research in the field, advanced by Gentner (Falkenhainer, Forbus, and Gentner 1989; Gentner and Markman 1997), describes analogy as a structural matching, in which elements from a base domain are mapped to (or aligned with) those in a target domain via structural similarities of their relations. This approach named *structure mapping theory*, with its computational implementation, the *Structure Mapping Engine* (SME) (Falkenhainer, Forbus, and Gentner 1989), has been cited as the most influential work to date on the modeling of analogy-making (French 2002). Alternative approaches in the field include the coherence based view developed by Holyoak and Thagard (Thagard et al. 1990; Holyoak and Thagard 1996), in which analogy is considered as a constraint satisfaction problem involving structure, semantic similarity, and purpose; and the view of Hofstadter (1995) of analogy as a kind of high-level perception, where one situation is perceived as another one. Veale and Keane (1997) extend the work in analogical reasoning to the more specific case of metaphors, which describe the understanding of one kind of thing in terms of another. A highly related cognitive theory is the *conceptual blending* idea developed by Fauconnier and Turner (2002), which involves connecting several existing concepts to create new meaning, operating below the level of consciousness as a fundamental mechanism of cognition. An implementation of this idea is given by Pereira (2007) as a computational model of abstract thought, creativity, and language.

According to whether the base and target cases belong to the same or different domains, there are two types of analogy: *intra-domain*, confined to surface similarities within the same domain; and *cross-domain*, using deep structural similarities between semantically distant information. While much of the research in artificial intelligence has been restricted to intra-domain analogies (e.g. case-based reasoning), studies in psychology have been more concerned with cross-domain analogical experiments (Thagard et al. 1990).

Evolutionary and Memetic Algorithms

Generalizing the mechanisms of the evolutionary process that has given rise to the diversity of life on earth, the approach of *Universal Darwinism* uses a simple progression of variation, natural selection, and heredity to explain a wide variety of phenomena; and it extends the domain of this process to systems outside biology, including economics, psychology, physics, and even culture (Dennett 1995). In terms of application, the metaheuristic optimization method of evolutionary algorithms (EA) provides an implementation of this idea, established as a solid technique with diverse problems in engineering as well as natural and social sciences (Coello Coello, Lamont, and Van Veldhuizen 2007).

In an analogy with the unit of heredity in biological evolution, the gene, the concept of *meme* was introduced by Dawkins (1989) as a unit of idea or information in cultural evolution, hosted, altered, and reproduced in individuals' minds, forming the basis of the field of memetics[3].

[3]Quoting Dawkins (Dawkins 1989): *"Examples of memes are tunes, ideas, catch-phrases, clothes fashions, ways of making pots or of building arches. Just as genes propagate themselves in the gene pool by leaping from body to body via sperms or eggs, so memes propagate themselves in the meme pool by leaping from brain to brain..."*

Within evolutionary computation, the recently maturing field of memetic algorithms (MA) has experienced increasing interest as a method for solving many hard optimization problems (Moscato, Cotta, and Mendes 2004). The existing formulation of MA is essentially a hybrid approach, combining classical EA with local search, where the population-based global sampling of EA in each generation is followed by an individual learning step mimicking cultural evolution, performed by each candidate solution. For this reason, this approach has been often referred to under different names besides MA, such as "hybrid EA" or "Lamarckian EA". To date, MA has been successfully applied to a wide variety of problem domains such as NP-hard optimization problems, engineering, machine learning, and robotics.

The potential of an evolutionary approach to creativity has been noted from cultural and practical viewpoints (Gabora 1997; Boden 2009). EA techniques have been shown to emulate creativity in engineering, such as genetic programming (GP) introduced by Koza (2003) as being capable of "routinely producing inventive and creative results"[4]; as well as in visual art, design, and music (Romero and Machado 2008). In psychology, there are studies providing support to an evolutionary view of creativity, such as the behavioral analysis by Simonton (2003) inferring that scientific creativity constitutes a form of constrained stochastic behavior.

The Algorithm

Our approach is based on a *meme pool* comprising individuals represented as semantic networks, subject to variation and selection under a fitness measure. We position our algorithm as *a novel memetic algorithm*, because (1) it is the units of culture, or information, that are undergoing variation, transmission, and selection, very close to the original sense of "memetics" as it was introduced by Dawkins; and (2) this is unlike the existing sense of the word in current MA as an hybridization of individual learning and EA. This algorithm is intended as a new tool focused exclusively on the memetic evolution of knowledge itself, which can find use in knowledge-based systems, reasoning, and creativity.

Our algorithm proceeds similar to a conventional EA cycle (Algorithm 1), with a relatively small set of parameters. We implement semantic networks as linked-list data structures of concept and relation objects. The descriptions of representation, fitness evaluation, variation, and selection steps are presented in the following sections. Parameters affecting each step of the algorithm are given in Table 1.

Algorithm 1 Outline of the algorithm

```
1: procedure MEMETICALGORITHM
2:     P(t = 0) ← INITIALIZE(Pop_size, C_max, R_min, T)
3:     repeat
4:         φ(t) ← EVALUATEFITNESSES(P(t))
5:         S(t) ← SELECTION(P(t), φ(t), S_size, S_prob)
6:         V(t) ← VARIATION(S(t), P_c, P_m, T)
7:         P(t + 1) ← V(t)
8:         t ← t + 1
9:     until stop criterion
10: end procedure
```

[4]Striking examples of demonstrated GP creativity include replication of historically important discoveries in engineering, such as the reinvention of negative feedback circuits originally conceived by Harold Black in 1920s.

Representation

The algorithm is centered on the use of semantic networks (Sowa 1991) for encoding evolving memotypes. An important characteristic of a semantic network is whether it is definitional or assertional: in definitional networks the emphasis is on taxonomic relations (e.g. $IsA(bird, animal)$[5]) describing a subsumption hierarchy that is true by definition; in assertional networks, relations describe instantiations that are contingently true (e.g. $AtLocation(human, city)$). In this study we combine the two approaches for increased expressivity. As such, semantic networks provide a simple yet powerful means to represent the "memes" of Dawkins as data structures that are algorithmically manipulatable, allowing a procedural implementation of memetic evolution.

In terms of representation, our approach is similar to several existing graph-based encodings of individuals in EA. The most notable is genetic programming (GP) (Koza et al. 2003), where candidate solutions are computer programs represented in a tree hierarchy. Montes and Wyatt (2004) present a detailed overview of graph-based EA techniques besides GP, which include parallel distributed genetic programming (PDGP), genetic network programming (GNP), evolutionary graph generation, and neural programming.

Using a graph-based representation makes the design of variation operators specific to graphs necessary. In works such as GNP, this is facilitated by using a string-based encoding of node types and connectivity, permitting operators very close to their counterparts in conventional EA; and in PDGP, operations are simplified by making nodes occupy points in a fixed-size two-dimensional grid. What is common with GP related algorithms is that the output of each node in the graph can constitute an input to another node. In comparison, the range of connections that can form a semantic network of a given set of concepts is limited by commonsense knowledge, i.e. the relations have to make sense to be useful (e.g. $IsA(bird, animal)$ is meaningful while $Causes(bird, table)$ is not). To address this issue, we introduce new crossover and mutation operations for memetic variation, making use of commonsense reasoning (Mueller 2006) and adapted to work on semantic networks.

Commonsense Knowledge Bases Commonsense reasoning refers to the type of reasoning involved in everyday thinking, based on *commonsense knowledge* that an ordinary person is expected to know, or "the knowledge of how the world works" (Mueller 2006). Knowledge bases such as the *ConceptNet*[6] project of MIT Media Lab (Havasi, Speer, and Alonso 2007) and *Cyc*[7] maintained by Cycorp company are set up to assemble and classify commonsense information. The lexical database *WordNet*[8] maintained by the Cognitive Science Laboratory at Princeton University also has characteristics of a commonsense knowledge base, via synonym, hypernym[9], and hyponym[10] relations (Fellbaum 1998).

In our implementation we make use of ConceptNet version 4 and WordNet version 3 to process commonsense knowledge, where ConceptNet contributes around 560,000 definitional and assertional relations involving 320,000 concepts and WordNet contributes definitional relations involving around 117,000 synsets[11]. The hypernym and hyponym relations among noun synsets in WordNet provide a reliable collection of IsA relations. In contrast, the variety of assertions in ConceptNet, contributed by volunteers across the world, makes it more prone to noise. We address this by ignoring all assertions with a reliability score (determined by contributors' voting) below a set minimum R_{min} (Table 1).

Initialization

At the start of each run of the algorithm, the population of size Pop_{size} is initialized with individuals created by *random semantic network generation* (Algorithm 1). This is achieved by starting from a network comprising only one concept randomly picked from commonsense knowledge bases and running a semantic network expansion algorithm that (1) randomly picks a concept in the given network (e.g. $human$); (2) compiles a list of relations—from commonsense knowledge bases—that the picked concept can be involved in (e.g. $\{CapableOf(human, think), Desires(human, eat), \cdots\}$) (3) appends to the network a relation randomly picked from this list, together with the other involved concept; and (4) repeats this until a given number of concepts has been appended or a set timeout T has been reached (covering situations where there are not enough relations). Note that even if grown in a random manner, the resulting network itself is totally meaningful and consistent because it is a combination of rational information from commonsense knowledge bases.

The initialization algorithm depends upon the parameters of C_{max}, the maximum number of initial concepts, and R_{min}, the minimum ConceptNet relation score (Table 1).

Fitness Measure

Since the individuals in our approach represent knowledge, or memes, the fitness for evolutionary selection is defined as a function of the represented knowledge. For the automated generation of analogies through evolution, we introduce a memetic fitness based on analogical similarity with a given semantic network, utilizing the Structure Mapping Engine (SME) (Falkenhainer, Forbus, and Gentner 1989; Gentner and Markman 1997). Taking the analogical matching score from SME as the fitness, our algorithm can evolve collections of information that are analogous to a given one.

In SME, an analogy is a one-to-one mapping from the base domain into the target domain, which correspond, in our fitness measure, to the semantic network supplied at the start and the individual networks whose fitnesses are evaluated by the function. The mapping is guided by the structure of relations between concepts in the two domains, ignoring the semantics of the concepts themselves; and is based on the systematicity principle, where connected knowledge is preferred over independent facts and is assigned a higher matching score. As an example, the Rutherford–Bohr atom and Solar System analogy (Gentner and Markman 1997) would involve a mapping from sun and $planet$ in the first domain to $nucleus$ and $electron$ in the second domain. The labels and structure of relations in the two domains (e.g. $\{Attracts(sun, planet), Orbits(planet, sun),$

[5] Here we adopt the notation $IsA(bird, animal)$ to mean that the concepts $bird$ and $animal$ are connected by the directed relation IsA, i.e. "bird is an animal."

[6] `http://conceptnet.media.mit.edu`

[7] `http://www.cyc.com`

[8] `http://wordnet.princeton.edu`

[9] Y is a *hypernym* of X if every X is a (kind of) Y ($IsA(dog, canine)$).

[10] Y is a *hyponym* of X if every Y is a (kind of) X.

[11] A *synset* is a set of synonyms that are interchangeable without changing the truth value of any propositions in which they are embedded.

$\cdots\}$ and $\{Attracts(nucleus, electron), Orbits(electron, nucleus), \cdots\}$) define and constrain the possible mappings between concepts that can be formed by SME.

We make use of our own implementation of SME based on the original description by Falkenhainer et al. (1989) and adapt it to the simple concept–relation structure of semantic networks, by mapping the predicate calculus constructs of *entities* into *concepts*, *relations* to *relations*, *attributes* to IsA *relations*, and excluding *functions*.

Selection

After assigning fitness values to all individuals in the current generation, these are replaced with offspring generated by variation operators on parents. The parents are probabilistically selected from the population according to their fitness, with reselection allowed. While individuals with a higher fitness have a better chance of being selected, even those with low fitness have a chance to produce offspring, however small. In our experiments we employ tournament selection (Coello Coello, Lamont, and Van Veldhuizen 2007), meaning that for each selection, a "tournament" is held among a few randomly chosen individuals, and the more fit individual of each successive pair is the winner according to a winning probability (Table 1).

In each cycle of algorithm, crossover is applied to parents selected from the population until $Pop_{size} \times P_c$ offspring are created (Table 1). Mutation is applied to $Pop_{size} \times P_m$ selected individuals, supplying the remaining part of the next generation (i.e. $P_c + P_m = 1$). We also employ elitism, by replacing a randomly picked offspring in next generation with the individual with the current best fitness.

Variation Operators

In contrast with existing graph-based evolutionary approaches that we have mentioned, our representation does not permit arbitrary connections between different nodes and requires variation operators that should be based on information provided by commonsense knowledge bases. This means that any variation operation on the individuals should: (1) preserve the structure within boundaries set by commonsense knowledge; and (2) ensure that even vertices and edges randomly introduced into a semantic network connect to existing ones through meaningful and consistent relations[12].

Here we present commonsense crossover and mutation operators specific to semantic networks.

Commonsense Crossover In classical EA, features representing individuals are commonly encoded as linear strings and the crossover operation simulating genetic recombination is simply a cutting and merging of this one dimensional object from two parents. In graph-based approaches such as GP, subgraphs can be freely exchanged between parent graphs (Koza et al. 2003; Montes and Wyatt 2004). Here, as mentioned, the requirement that a semantic network has to make sense imposes significant constraints on the nature of recombination.

We introduce two types of *commonsense crossover* that are tried in sequence by the variation algorithm. The first type attempts a sub-graph interchange between two selected parents similar to common crossover in standard GP; and where this is not feasible due to the commonsense structure of relations forming the parents, the second type falls back to a combination of both parents into a new offspring.

Type I (subgraph crossover): A pair of concepts, one from each parent, that are *interchangeable*[13] are selected as *crossover concepts*, picked randomly out of all possible such pairs. For instance, in Figure 1, $bird$ and $airplane$ are interchangeable, since they can replace each other in the relations $CapableOf(\cdot, fly)$ and $AtLocation(\cdot, air)$. In each parent, a subgraph is formed, containing: (1) the crossover concept; (2) the set of all relations, and associated concepts, that are not common with the other crossover concept (In Figure 1 (a), $HasA(bird, feather)$ and $AtLocation(bird, forest)$; and in (b) $HasA(airplane, propeller)$, $MadeOf(airplane, metal)$, and $UsedFor(airplane, travel)$); and (3) the set of all relations and concepts connected to these (In Figure 1 (a) $PartOf(feather, wing)$ and $PartOf(tree, forest)$; and in (b) $MadeOf(propeller, metal)$), excluding the ones that are also one of those common with the other crossover concept (the concept fly in Figure 1 (a), because of the relation $CapableOf(\cdot, fly)$). This, in effect, forms a subgraph of information specific to the crossover concept, which is insertable into the other parent. Any relations between the subgraph and the rest of the network not going through the crossover concept are severed (e.g. $UsedFor(wing, fly)$ in Figure 1 (a)). The two offspring are formed by exchanging these subgraphs between the parent networks (Figure 1 (c) and (d)).

Type II (graph merging crossover): A concept from each parent that is *attachable*[14] to the other parent is selected as a *crossover concept*. The two parents are merged into an offspring by attaching a concept in one parent to another concept in the other parent, picked randomly out of all possible attachments ($CreatedBy(art, human)$ in Figure 2. Another possibility is $Desires(human, joy)$.). The second offspring is formed randomly the same way. In the case that no attachable concepts are found, the parents are merged as two separate clusters within the same semantic network.

Commonsense Mutation We introduce several types of *commonsense mutation* operators that modify a parent by means of information from commonsense knowledge bases. For each mutation to be performed, the type is picked at random with uniform probability. If the selected type of mutation is not feasible due to the commonsense structure of the parent, another type is again picked. In the case that a set timeout of T trials has been reached without any operation, the parent is returned as it is.

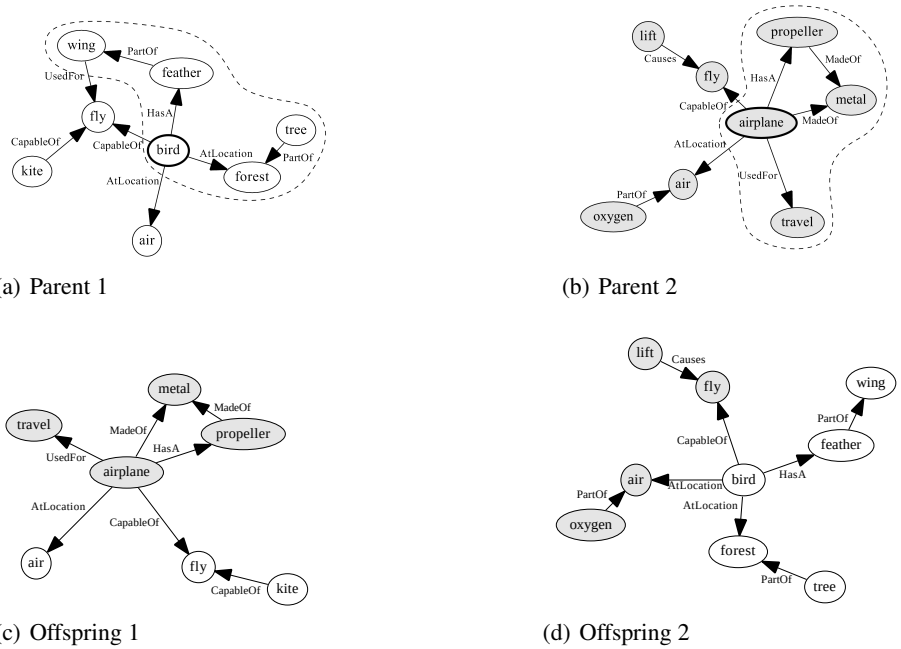
Type I (concept attachment): A new concept randomly picked from the set of concepts *attachable* to the parent is attached through a new relation to one of existing concepts (Figure 3 (a) and (b)).

Type IIa (relation addition): A new relation connecting two existing concepts in the parent is added, possibly connecting unconnected clusters within the same network (Figure 3 (c) and (d)).

[12] It should be noted that we depend on the meaningfulness and consistency (i.e. compatibility of relations with others involving the same concepts) of information in the commonsense knowledge bases, which should be ensured during their maintenance.

[13] We define two concepts from different semantic networks as *interchangeable* if both can replace the other in all, or part, of the relations the other is involved in, queried from commonsense knowledge bases.

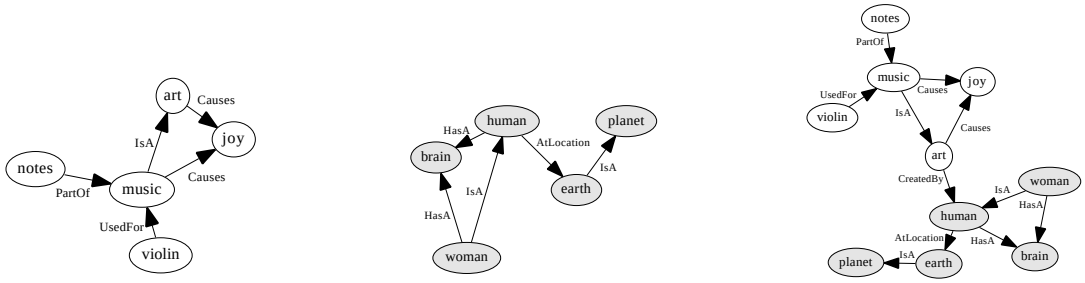
[14] We define a distinct concept as *attachable* to a semantic network if at least one commonsense relation connecting the concept to any of the concepts in the network can be discovered from commonsense knowledge bases.



(a) Parent 1 (b) Parent 2

(c) Offspring 1 (d) Offspring 2

Figure 1: Commonsense crossover type I (subgraph crossover), centered on the concepts of $bird$ for parent 1 and $airplane$ for parent 2.



(a) Parent 1 (b) Parent 2 (c) Offspring

Figure 2: Commonsense crossover type II (graph merging crossover), merging by the relation $CreatedBy(art, human)$. If no concepts attachable through commonsense relations are encountered, the offspring is formed by merging the parent networks as two separate clusters within the same semantic network.

Type IIb (relation deletion): A randomly picked relation in the parent is deleted, possibly leaving unconnected clusters within the same network (Figure 3 (e) and (f)).

Type IIIa (concept addition): A randomly picked new concept is added to the parent as a new cluster (Figure 3 (g) and (h)).

Type IIIb (concept deletion): A randomly picked concept is deleted with all the relations it is involved in, possibly leaving unconnected clusters within the same network (Figure 3 (i) and (j)).

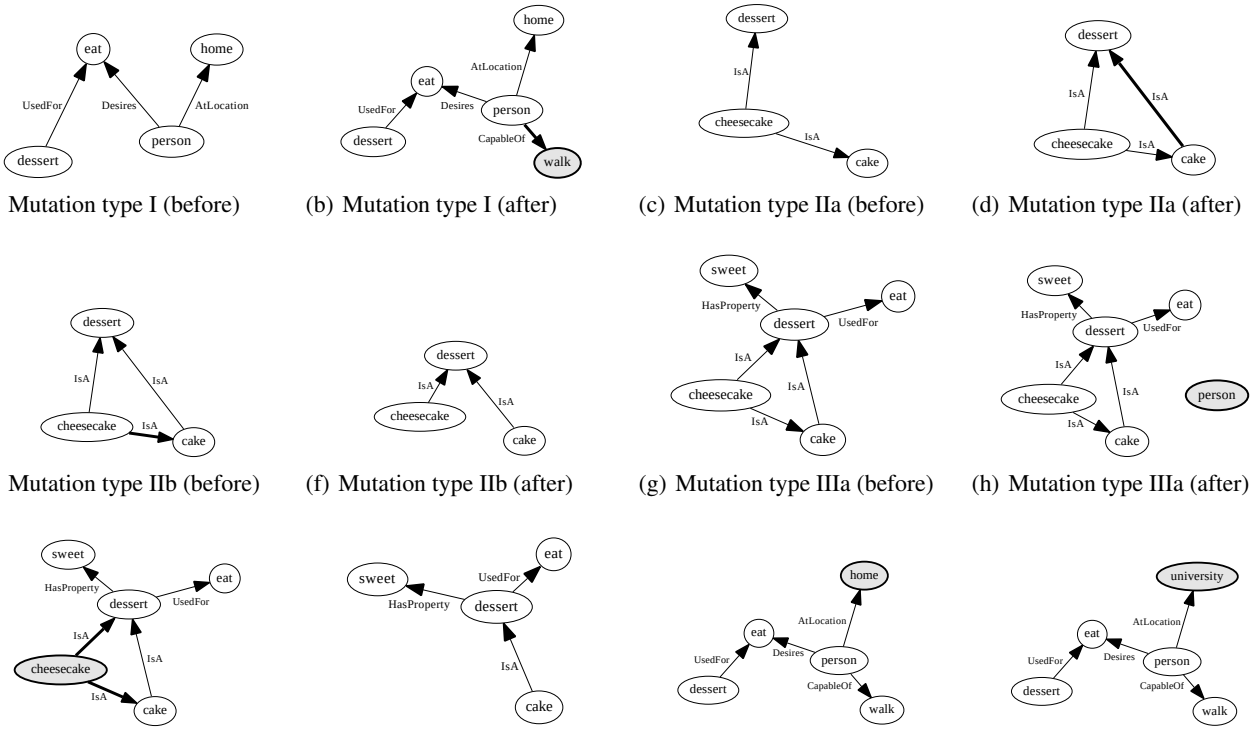
Type IV (concept replacement): A concept in the parent, randomly picked from the set of those with at least one *interchangeable* concept, is replaced with one (randomly picked) of its interchangeable concepts. Any relations left unsatisfied by the new concept are deleted (Figure 3 (k) and (l)).

Results and Discussion

In this introductory study, we adopt values for crossover and mutation probabilities similar to earlier studies in graph-based EA (Koza et al. 2003; Montes and Wyatt 2004) (Table 1). We use a crossover probability of $P_c = 0.85$, and a somewhat-above-average mutation rate of $P_m = 0.15$, accounting for the high tendency of mutation postulated in memetic literature[15]. In our experiments, we subject a population of $Pop_{size} = 200$ individuals to tournament selection with tournament size $S_{size} = 8$ and winning probability $S_{prob} = 0.8$.

Using this parameter set, we present the results from two runs of experiment: evolved analogies for a network de-

[15] See Gil-White (Gil-White 2008) for a review and discussion of mutation in memetics.



(a) Mutation type I (before) (b) Mutation type I (after) (c) Mutation type IIa (before) (d) Mutation type IIa (after)

(e) Mutation type IIb (before) (f) Mutation type IIb (after) (g) Mutation type IIIa (before) (h) Mutation type IIIa (after)

(i) Mutation type IIIb (before) (j) Mutation type IIIb (after) (k) Mutation type IV (before) (l) Mutation type IV (after)

Figure 3: Examples illustrating the types of commonsense mutation used in this study.

scribing some basic astronomical knowledge are shown in Figure 4 and for a network of familial relations in Figure 5. We show in Figure 6 (a) the progress of the best and average fitness in the population during the run that produced the results in Figure 4. The best and average size of semantic networks forming the individuals are shown in Figure 6 (b). We observe that evolution asymptotically reaches a fitness plateau after about 40 generations. This coincides roughly with the point where the size of the best individual (13–14) becomes comparable with that of the given base semantic network (11, in Figure 4), after which improvements in the one-to-one analogy become sparser and less feasible. We also note that, between generations 21–34, the best network size actually gets smaller, demonstrating the possibility of improvement in network configuration without adding further nodes. Our experiments demonstrate that the proposed algorithm is capable of spontaneously creating collections of knowledge analogous to the one given in a base semantic network, with very good performance. In most cases, our implementation was able to reach extensive analogies within 50 generations and reasonable computational time.

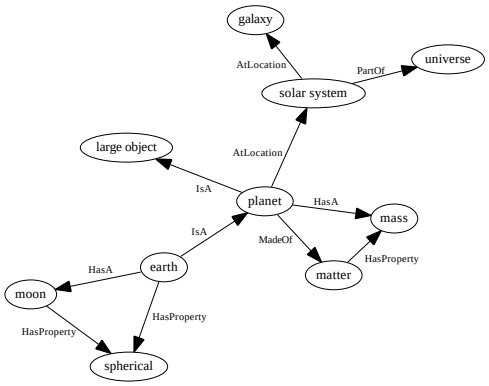
From an analogical reasoning viewpoint, the algorithm achieves the generation of diverse novel cases analogous to a given case. Compared with the Kilaza model of O'Donoghue (2004) for finding novel analogous cases, which works by evaluating possible analogies to a given target case from a collection of candidate source domains that are assumed to be available, our approach is capable of open-ended and spontaneous creation of analogous cases from the ground up, replicating an essential mode of creative behavior observed in psychology (Clement 1988).

An important result is that, even if the use of commonsense knowledge in our algorithm was prompted by concerns that are practical in nature (i.e. restrictions on the meaningfulness and consistency of memetic variation by the introduced crossover and mutation operators), it eventually serves to tackle a very fundamental and long-standing problem in computational creativity: as put forth by Boden (2009), "no current AI system has access to the rich and subtly structured stock of concepts that any normal adult human being has built up over a lifetime" and "what's missing, as compared with the human mind, is the rich store of world knowledge (including cultural knowledge) that's often involved." We believe that the inherent commonsense reasoning element in our approach provides a means to address this criticism of lack of world knowledge in computational approaches to creativity.

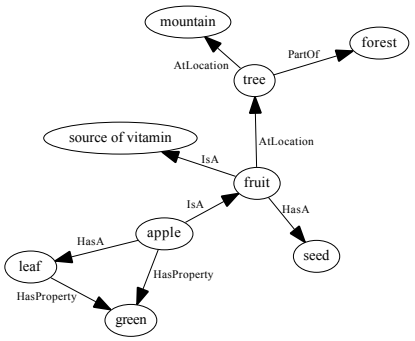
Conclusions and Future Work

We have presented a novel evolutionary algorithm that employs semantic networks as evolving individuals, paralleling the model of cultural evolution in the field of memetics. This algorithm, to our knowledge, is the first of its kind. The use of semantic networks provides a suitable basis for implementing variation and selection of memes as put forth by Dawkins (Dawkins 1989). We have introduced preliminary versions of variation operators that work on this representation, utilizing knowledge from commonsense knowledge bases. We have also contributed a memetic fitness measure based on the structure mapping theory from psychology.

Even if it is an intuitive fact that human culture and knowledge are evolving with time, existing models of cul-

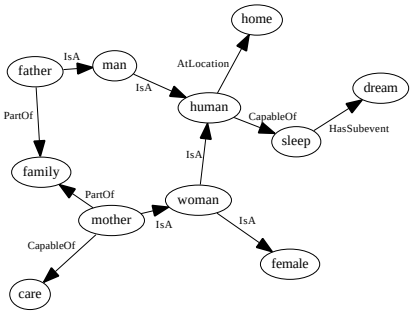


(a) Given semantic network, 10 concepts, 11 relations (base domain)

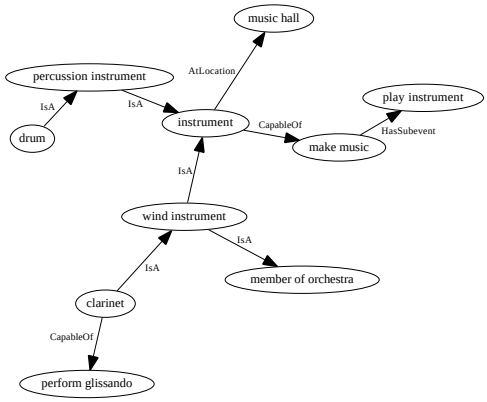


(b) Evolved individual, 9 concepts, 9 relations (target domain)

Figure 4: Experiment 1: The evolved individual is encountered after 35 generations, with fitness value 2.8. Concepts and relations of the individual not involved in the analogy are not shown here for clarity.



(a) Given semantic network, 11 concepts, 11 relations (base domain)



(b) Evolved individual, 10 concepts, 9 relations (target domain)

Figure 5: Experiment 2: The evolved individual is encountered after 42 generations, with fitness value 2.7. Concepts and relations of the individual not involved in the analogy are not shown here for clarity.

Table 1: Parameters used during experiments

	Parameter	Value
Evolution	Population size (Pop_{size})	200
	Crossover probability (P_c)	0.85
	Mutation probability (P_m)	0.15
Semantic networks	Max. initial concepts (C_{max})	5
	Min. relation score (R_{min})	2.0
	Timeout (T)	10
Selection	Type	Tournament
	Tournament size (S_{size})	8
	Tournament win prob. (S_{prob})	0.8
	Elitism	Employed

ture, in their current state, are too minimalistic and weak in their descriptions of individual creativity and novelty; and conversely, theories modeling individual creativity lack consideration of cultural transmission and replication (Gabora 1997). We believe that studies exploring creativity with evolutionary approaches have the potential for bridging this gap.

In future work, an interesting possibility is to start the random semantic network generation procedure with several given concepts, allowing the discovery of cases formed around a particular set of seed concepts. The simple fitness function used in this introductory study can be extended to take graph-theoretical properties of semantic networks into account, such as the number of nodes or edges, shortest path length, or the clustering coefficient. The research would also benefit from exploring different types of mutation and crossover, and grounding the design of such operators on existing theories of cultural transmission and variation, discussed in sociological theories of knowledge.

A direct and very interesting application of our approach would be to devise experiments with realistically formed fitness functions modeling selectionist theories of knowledge, which remain untested until this time. One such theory is the *evolutionary epistemology* of Campbell (Bickhard and Campbell 2003), describing the development of human knowledge and creativity through selectionist principles such as blind variation and selective retention (BVSR).

Acknowledgments

This work was supported by a JAE-Predoc fellowship from CSIC, and the research grants: 2009-SGR-1434 from the Generalitat de Catalunya, CSD2007-0022 from MICINN, and Next-CBR TIN2009-13692-C03-01 from MICINN.

References

Bickhard, M. H., and Campbell, D. T. 2003. Variations in variation and selection: the ubiquity of the variation-and-

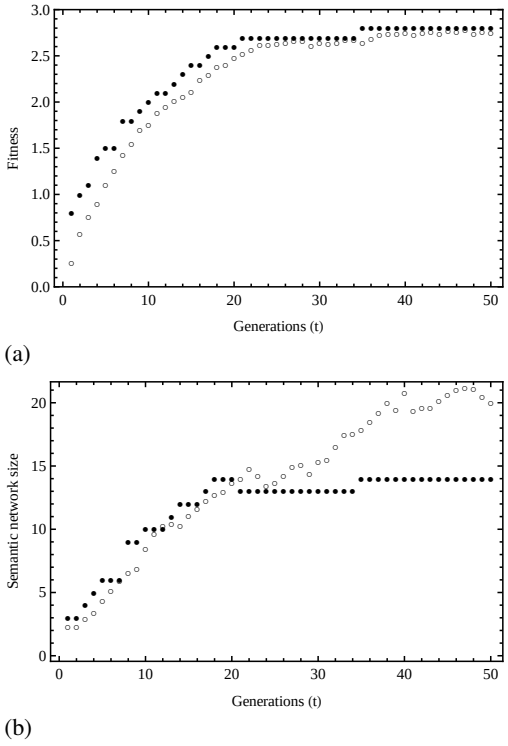


Figure 6: Evolution of (a) fitness and (b) semantic network size during the course of an experiment with parameters given in Table 1. Filled circles represent the best individual in a generation, empty circles represent population average. Network size is taken to be the number of relations (edges).

selective-retention ratchet in emergent ogranizational complexity. *Foundations of Science* 8:215–2182.

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. London: Routledge, second edition.

Boden, M. A. 2009. Computer models of creativity. *AI Magazine* 30(3):23–34.

Clement, J. 1988. Observed methods for generating analogies in scientific problem solving. *Cognitive Science* 12:563–586.

Coello Coello, C. A.; Lamont, G. B.; and Van Veldhuizen, D. A. 2007. *Evolutionary Algorithms for Solving Multi-Objective Problems*. Springer.

Dawkins, R. 1989. *The Selfish Gene*. Oxford University Press.

Dennett, D. C. 1995. *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. Simon & Schuster.

Falkenhainer, B.; Forbus, K. D.; and Gentner, D. 1989. The Structure-Mapping Engine: Algorithm and examples. *Artificial Intelligence* 41:1–63.

Fauconnier, G., and Mark, T. 2002. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. New York: Basic Books.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

French, R. M. 2002. The computational modeling of analogy-making. *Trends in Cognitive Sciences* 6(5):200–205.

Gabora, L. 1997. The origin and evolution of culture and creativity. *Journal of Memetics – Evolutionary Models of Transmission* 1.

Gentner, D., and Markman, A. B. 1997. Structure mapping in analogy and similarity. *American Psychologist* 52:45–56.

Gil-White, F. 2008. Let the meme be (a meme): insisting too much on the genetic analogy will turn it into a straightjacket. In Botz-Bornstein, T., ed., *Culture, Nature, Memes*. Newcastle upon Tyne: Cambridge Scholars.

Havasi, C.; Speer, R.; and Alonso, J. 2007. ConceptNet 3: a flexible, multilingual semantic network for common sense knowledge. In *Proceedings of Recent Advances in Natural Language Processing*.

Hofstadter, D. R. 1995. *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought*. New York: Basic Books.

Hofstadter, D. 2001. Analogy as the core of cognition. In Gentner, D.; Holyoak, K. J.; and Kokinov, B., eds., *Analogical Mind: Perspectives From Cognitive Science*. Cambridge, MA: MIT Press. 499–538.

Holyoak, K. J., and Thagard, P. 1996. *Mental Leaps: Analogy in Creative Thought*. Bradford Books.

Koza, J. R.; Keane, M. A.; Streeter, M. J.; Mydlowec, W.; Yu, J.; and Lanza, G. 2003. *Genetic Programming IV: Routine Human-Competitive Machine Intelligence*. Kluwer Academic Publishers.

Montes, H. A., and Wyatt, J. L. 2004. Graph representation for program evolution: An overview. Technical report, University of Birmingham School of Computer Science.

Moscato, P. A.; Cotta, C.; and Mendes, A. 2004. *Studies in Fuzziness and Soft Computing – New Optimization Techniques in Engineering*. New York: Springer. chapter Memetic Algorithms.

Mueller, E. T. 2006. *Commonsense Reasoning*. Morgan Kaufmann.

O'Donoghue, D. 2004. *Finding Novel Analogies*. Ph.D. Dissertation, University College Dublin, Department of Computer Science, Dublin, Ireland.

Pereira, F. C. 2007. *Creativity and Artificial Intelligence: A Conceptual Blending Approach*. New York: Mouton de Gruyter.

Romero, J. J., and Machado, P. 2008. *The Art of Artificial Evolution: A Handbook on Evolutionary Art and Music*. Springer.

Simonton, D. K. 2003. Scientific creativity as constrained stochastic behavior: The integration of product, person, and process perspectives. *Psychological Bulletin* 129(4):475–494.

Sowa, J. F. 1991. *Principles of Semantic Networks: Explorations in the Representation of Knowledge*. San Mateo: Mogran Kaufmann.

Thagard, P.; Holyoak, K. J.; Nelson, G.; and Gochfeld, D. 1990. Analog retrieval by constraint satisfaction. *Artificial Intelligence* 46:259–310.

Veale, T., and Keane, M. 1997. The competence of sub-optimal structure mapping on hard analogies. In *Proceedings of the 15th International Joint Conference on AI*. San Mateo, CA: Morgan Kauffman.

Ward, T. B.; Smith, S. M.; and Vaid, J. 2001. *Creative Thought: An Investigation of Conceptual Structures and Processes*. Washington, DC: American Psychological Association.

Cross-domain Literature Mining: Finding Bridging Concepts with CrossBee

Matjaž Juršič[1,2], Bojan Cestnik[3,1], Tanja Urbančič[4,1], Nada Lavrač[1,4]
[1] Jožef Stefan Institute, Ljubljana, Slovenia
[2] International Postgraduate School Jožef Stefan, Ljubljana, Slovenia
[3] Temida d.o.o., Ljubljana, Slovenia
[4] University of Nova Gorica, Nova Gorica, Slovenia
{matjaz.jursic, bojan.cestnik, tanja.urbancic, nada.lavrac}@ijs.si

Abstract

In literature-based creative knowledge discovery one of the challenging tasks is to identify interesting bridging terms or concepts which relate different domains. To find these bridging concepts, our cross-domain literature mining approach assumes that one first has to identify two seemingly unrelated domains of interest. Bridging terms, found in the intersection of these domains, are then ranked according to their potential to uncover useful, previously unexplored links between the two domains. Term ranking, based on voting of an ensemble of heuristics, is the main functionality of the CrossBee (Cross-Context Bisociation Explorer) system presented in this paper. The utility of the proposed approach is show-cased by exploring scientific papers on migraine and magnesium, which is a reference domain in literature mining.

Introduction

This paper[1] investigates the creative knowledge discovery process which has its grounds in Mednick's *associative* creativity theory (Mednick 1962) and Koestler's domain-crossing associations, called *bisociations* (Koestler 1964). Mednick defines creative thinking as the capacity of generating new combinations of distinct associative elements (concepts). He explains how thinking about the concepts that are not strictly related to the elements under investigation inspires unforeseen useful connections between these elements. On the other hand, according to Koestler, a bisociation is a result of creative processes of the mind when making completely new associations between concepts from domains that are usually considered separate. Consequently, discovering bisociations may considerably improve creative discovery processes. According to Koestler, through the history of science, this mechanism has been a crucial element of progressive insights and paradigm shifts.

The approach to creative knowledge discovery from text documents presented in this paper is based on identifying and exploring terms which have the potential to relate different domains of interest, i.e., two distinct domain literatures. While in general *literature* refers to any document corpus (articles, novels, stories, etc.), our approach to cross-domain literature mining focuses on the task of mining scientific papers in the so-called *closed discovery*[2] setting (Weeber et al., 2001) where two domains of interest, *A* and *C*, are identified by the expert prior to starting the knowledge discovery process, and the goal is to find interesting *bridging terms* that relate the two literatures.

Weeber et al. (2001) have followed the work of literature-based knowledge discovery in medical domains by Swanson (1986) who designed the so-called *ABC model* approach to investigate whether the phenomenon of interest *C* in the first domain is related to some phenomenon *A* in the other literature through some interconnecting phenomenon *B* addressed in both literatures. If the literature about *C* relates *C* with *B*, and the literature about *A* relates *A* with the same *B*, then combining these relations may suggest a relation between *C* and *A*. If closer inspection confirms that an uncovered relation between *C* and *A* is new, meaningful and interesting, this can be viewed as new evidence or considered as a new piece of knowledge.

Smalheiser and Swanson (1998) developed an online system ARROWSMITH which takes as input two sets of titles from disjoint domains *A* and *C* and lists terms that are common to literatures *A* and *C*; the resulting *bridging terms* (*b*-terms, forming set *B*) are further investigated for their potential to generate new scientific hypotheses (see an

[1] This work was supported by the European Commission under the 7th Framework Programme FP7-ICT-2007-C FET-Open project BISON-211898, and Slovenian Research Agency grant Knowledge Technologies (P2 0103).

[2] In contrast with *closed discovery*, *open discovery* leads the creative knowledge discovery process from a given starting domain towards a yet unknown second domain which at the end of this process turns out to be connected with the first one.

example in Figure 1). Investigation of pairs of documents might seem rather straightforward, like in the example documents titled "Migraine treatment with calcium channel blockers" (Anderson et al., 1986) and "Magnesium: nature's physiologic calcium blocker" (Iseri and French, 1984). However, it should be left to domain experts to check whether bridging term *calcium channel blocker* suggests a valid, new and interesting relation (in this case, the relation that migraine could be treated with magnesium). To this end, it is helpful not just to identify a set of candidate bridging terms *B* between literatures *A* and *C*, but also to provide an expert with an easy access to the documents to be checked and to support this laborious process by ranking bridging terms candidates in order to start the exploration by considering the most promising terms first.

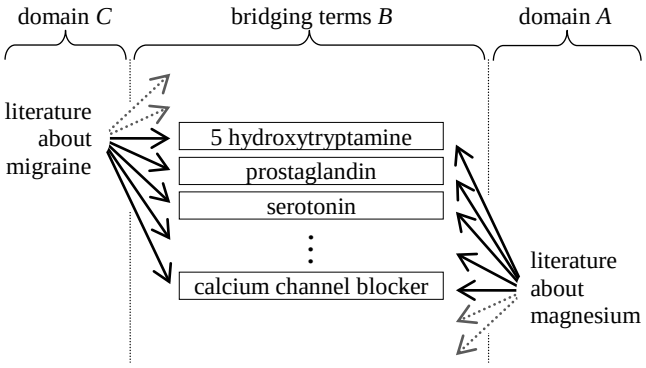


Figure 1. Gold standard cross-domain literature mining example: migraine (domain *C*) on the left, magnesium (domain *A*) on the right, and in between a selection of bridging terms *B* as identified by Swanson et al. (2006).

The approach presented in this paper is closely related to bridging terms identification in the RaJoLink system (Urbančič et al. 2007, Petrič et al. 2009). RaJoLink can be used to identify interesting scientific articles in the PubMed database, to compute different statistics, and to analyze the articles with the aim to discover new knowledge. The RaJoLink method involves three principal steps, Ra, Jo and Link, which have been named after the key elements of each step: Rare terms, Joint terms and Linking terms, respectively. In the Ra step, interesting rare terms in literature about phenomenon C under investigation are identified. In the Jo step, all available articles about the selected rare terms are inspected and interesting joint terms that appear in the intersection of the literatures about rare terms are identified as the candidates for A. This results in a candidate hypothesis that C is connected with A. To provide explanation for hypotheses generated in the Jo step, in the final Link step the method searches for b-terms, linking literatures A and C. Note that steps Ra and Jo implement the open discovery, while step Link corresponds to the closed discovery process of searching for b-terms when A and C are already known (as illustrated in Figure 1).

Focusing on the closed discovery process, the method proposed in this paper aims at finding bridging terms in documents of two given domains A and C, enabling the exploration of potentially interesting bisociative links between the given domains with the aid of an ensemble of new heuristics for bridging term discovery. Term ranking, based on voting of an ensemble of heuristics, is the main functionality of the new CrossBee (Cross Context Bisociation Exploration) system presented in this paper. To verify the utility of the proposed approach, CrossBee was tested on the problem of rediscovering links between migraine and magnesium literatures, first explored by Swanson (1986) and later by numerous other authors, including Weeber et al. (2001) and (Urbančič et al. 2009).

This paper is organized as follows. Section 2 presents and relates two creative knowledge discovery frameworks: Koestler's bisociative link discovery (Koestler 1964) and Swanson's *ABC* model of closed discovery in literature mining (Swanson 1986). It also relates our work to Boden's definition of creativity (Boden 1992) and Wigging's computational creativity definition (Wigging 2006). Section 3 presents the heuristics used for selecting the most promising bridging concepts (bridging terms or *b*-terms) in the intersection of two different sets of documents (two domains of interest), evaluated on the migraine-magnesium domain pair, explored originally in Swanson's research. It also presents an ensemble heuristic composed of six selected elementary heuristics. Section 4 presents the functionality and implementation of our system CrossBee for cross-context bridging term discovery. We conclude with a discussion and directions for further work.

Koestler's Bisociations, Cross-domain Literature Mining and Computational Creativity

Let us present some background on the mechanism of bisociative reasoning which is at the heart of creative, accidental discovery, referred to as *serendipity* by Roberts (1989). Bisociative discovery, studied in this work, is focused on finding unexpected terms/concepts linking different domains.

Scientific discovery requires creative thinking to connect seemingly unrelated information, for example, by using metaphors or analogies between concepts from different domains. These modes of thinking allow the mixing of conceptual categories or domains, which are normally separated. One of the functional bases for these modes is the idea of *bisociation*, coined by Artur Koestler (1964):

> "The pattern . . . is the perceiving of a situation or idea, L, in two self-consistent but habitually incompatible frames of reference, *M1* and *M2*. The event *L*, in which the two intersect, is made to vibrate simultaneously on two different wavelengths, as it were. While this unusual situation lasts, *L* is not merely linked to one associative context but bisociated with two."

Koestler investigated bisociation as the basis for human creativity in seemingly diverse human endeavors, such as humor, science, and arts.

In this paper we explore a specific pattern of bisociation (Berthold, 2012): terms, appearing in documents, which represent bisociative links between concepts of different domains, where the creative act is to find links which lead 'out-of-the-plane' in Koestler's terms, i.e., links which cross two or more different domains. According to Berthold (2012), we claim that two concepts are bisociated if (a) there is no direct, obvious evidence linking them, (b) one has to cross domains to find the link, and (c) this new link provides some novel insight into the problem domain. We explore an approach to bisociative cross-domain link discovery, based on the identification and ranking of terms which have the potential of acting as previously unexplored links between different predefined domains of expertise. It can be seen that–in terms of the Swanson's *ABC* model used in literature mining–this is an approach to closed knowledge discovery, where two domains of interest, *A* and *C*, are identified by the expert in advance. In terms of the Koestler's model, the two domains, *A* and *C*, correspond to the two habitually incompatible frames of reference, *M1* and *M2*. Moreover, the linking terms (called *bridging terms* or *b-terms* in this paper) that are common to literature *A* and *C*, explored by Smalheiser and Swanson (1998), clearly correspond to Koestler's notion of a situation or idea, *L*, which is not merely linked to one associative domain but bisociated with two domains *M1* and *M2*.

Since our work originates from Koestler's creative process definition, it naturally satisfies his notion of creativity. However, the concepts of *creativity* and *computational creativity* have several other definitions. We argue that our approach can be labeled as *creative* according to at least two other definitions, introduced by Boden (1992) and Wiggins (2006).

Boden (1992) defines creativity as "the ability to come up with ideas or artefacts that are new, surprising and valuable." Considering this definition, and given that the main output of our methodology is a ranked list of potentially interesting bridging terms/concepts, we argue that–although we do not produce new concepts–the ranking of potentially interesting bridging concepts itself may represent new, surprising and valuable ideas or artefacts. The proposed approach produces new term rankings, because–to the best of our knowledge–there are no similar methodologies available. The results are often also surprising, both because of their unlikeliness (as not commonly used terms may appear at the top of the ranked list) and their effect in subjective surprise (as noted by observing the expert using our system). The weakest claim we provide is the notion of value of the system as until now the developed approach did not yet produce any scientific breakthroughs; however, we already observed that it triggered novel insights by the expert who tested the early versions of our system. Therefore, we conclude that using Boden's definition, the level of our systems creativity is limited by the value of its results and only the reduced exploration time and the number of users will show how valuable the system is and how valuable its results really are.

Considering *computational creativity*, Wiggins (2006) proposes the following definition for which he states to be commonly adapted by the AI community: computational creativity refers to "performance of tasks (by a computer) which, if performed by a human, would be deemed creative." We argue that, although the ranking problem we solve is not something people usually do, our system can be considered creative according to this definition. Take an analogy with online search engines whose task is finding documents and ranking the search results. We believe that, if such rankings were performed by a human, this could be considered as a very creative process. The final results of our methodology–the insights which might arise from using our system–could also be considered scientifically creative, where the ultimate creative act will be performed by the experts using the system and not the system alone. We designed the methodology in a way to enable the expert to be more productive when generating such creative ideas. Therefore, we argue that this added effectiveness of the expert's creativity process originates from the system and its underlying methodology. Hence we believe our system possesses some elements of computational creativity proposed by Wiggins.

Bridging Term Detection Methodology

Creative thinking requires focusing on problems from new perspectives. In this paper we follow Koestler (1964), who investigated bisociation as the basis for human creativity, with a goal of developing a computational system with the ability to bridge different domains. Such relations between distinct domains can be revealed through bridging concepts (bridging terms, referred to as *b*-terms in this paper). Since this may lead to the generation of many possible ideas, the innovative generation of hypotheses as well as the support for facilitated exploration of alternatives are needed for creative cross-domain knowledge discovery.

Based on this assumption, we have developed and experimented with different heuristics for finding bridging terms in the context of closed knowledge discovery from two different domains of expertise. The intuition behind this research is that by developing appropriate heuristics for term evaluation and ranking, this will enable the user to inspect only the top-ranked terms which should result in a high probability to find observations that may lead to the discovery of new bridges between the literatures of different domains.

In summary, our research aim is to find cross-domain links by exploring the bridging terms in the intersection of two literatures that establish previously unknown links between literature *A* and literature *C*. In more detail, our method of *b*-term discovery is performed as follows.

1. Perform text preprocessing to encode input texts into the standard bag-of-words (BoW) representation. As in standard text preprocessing for text mining, this is performed through a number of steps:
    a. text tokenization (where a continuous character sequence is split into meaningful sub-tokens, i.e., individual words or terms),
    b. stop-word removal (removing predefined words from a language that usually carry no relevant information:. and, or, a, an, the, ...),
    c. stemming or lemmatization (the process that converts each word/token into its morphologically neutral form),
    d. n-gram construction (n-grams are terms defined as a concatenation of 1 to n words which appear consecutively in the text),
    e. bag-of-words (BoW) representation, i.e., a vector representation of a document, with value 1 (or word frequency-based weight) for words/terms appearing in the document, and value 0 for the rest of the corpus vocabulary.
2. Calculate the values of heuristics which favor *b*-terms over other terms.
3. Sort the intersecting terms according to the values of the best performing heuristics and present the top-ranked terms (hopefully representing the *b*-terms) to the expert during interactive exploration of the two domains.

The development of the best performing heuristics consisted of two phases:

1. Training: we proposed over 40 elementary heuristics, which vary from very simple term-frequency statistics to very elaborate combined measures. We then evaluated their quality on the migraine-magnesium gold standard domain investigated already by Swanson et al. (1988). Results of the evaluation were used to select some of the best performing and most complementary heuristics that were joined into a new ensemble heuristic. The ensemble heuristic proposed in this paper is generally more accurate and robust than any of the elementary heuristics used in its construction.
2. Testing: we independently evaluated the ensemble heuristic on a second dataset, autism-calcineurin documents investigated by (Petrič et al. 2009), to confirm its domain independence and its potential for b-term identification. Note that due to space restrictions, the description of testing of the system on the autism-calcineurin domain pair is out of the scope of this paper; the interested reader can find more information is provided in (Juršič et al. 2012).

Elementary Heuristics for b-term Detection

We have proposed over 40 elementary heuristics for *b*-term evaluation (Juršič et al. 2012), divided into four groups: frequency based, tf-idf based[3], similarity based, and outlier based heuristics. Most of these heuristics work fundamentally in a similar way: they manipulate the data present in the BoW document vector format to derive the term bi-sociation potential quality measure, named the *bisociation score*. The only exceptions are the outlier based heuristics which first detect outlier documents and then use the BoW vector information.

Instead of providing the entire list of heuristics whose performance we tested extensively, we only specify a subset of these which we actually selected to construct the ensemble heuristic. The selected heuristics are defined as follows.

Term to document frequency ratio: is a frequency based heuristic $freqRatio(t) = countTerm_{D_u}(t)/countDoc_{D_u}(t)$ defined as the ratio of the number of occurrences of term *t* in document set D_u (named term frequency in tf-idf related text preprocessing contexts), and the number of documents where term *t* appears in document set D_u (named document frequency in tf-idf related contexts).

Sum of term's importance in both domains: is a heuristic based on tf-idf metrics $tfidfDomnSum(t) = tfidf_{D_1}(t) + tfidf_{D_2}(t)$, defined as a sum of tf-idf value of term *t* in the centroid vector of document set D_1 plus term's tf-idf value in the centroid vector of document set D_2, where the centroid vector is defined as the sum of all document vectors and thus represents an average document of the given document collection.

Sum of term frequencies in three outlier sets: is an outlier based heuristic $outFreqSum(t) = countTerm_{D_{CS}}(t) + countTerm_{D_{RF}}(t) + countTerm_{D_{SVM}}(t)$ which computes the sum of term frequencies in three outlier sets, where the sets of outliers were identified by three classifiers (Sluban et al. 2012): Centroid Similarity (CS) classifier, Random Forest (RF) classifier, and Support Vector Machine (SVM) classifier.

Relative frequencies in outlier sets: focusses on outlier sets $outFreqRelCS(t) = countTerm_{D_{CS}}(t)/countTerm_{D_u}(t)$. Documents in the outliers set frequently embody new information that is often hard to explain in the context of existing knowledge. We concentrate on specific outliers–domain outliers–i.e., documents that tend to be more similar to the documents of the other domain than to those of their own domain. The procedure that we use to detect outlier documents first builds a classification model for each domain and then classifies all the documents using the trained classifier. The documents that are misclassified are declared as outlier documents, since according to the classification model they do not belong to their initial domain. The other two outlier based heuristics–*relative frequency in the RF outlier set (outFreqRelRF)* and *relative*

[3] tf-idf stands for Term Frequency Inverse Document Frequency word weight computation, used in text mining (Feldman and Sanger, 2007).

frequency in the SVM outlier set (*outFreqRelSVM*)–are defined in the same way as the *outFreqRelCS* heuristic.

We have defined also a supplementary *baseline heuristic*: $random(t) = randNum()$ which serves as a baseline for the others, as it returns a random number from interval (0,1) regardless of a term under investigation.

Evaluation of Elementary Heuristics

To test the proposed heuristics for *b*-term detection, we have evaluated them on the problem of detecting bisociative links between migraine and magnesium in the respective literatures. To this end, we replicated the early Swanson's migraine-magnesium experiment that represents a gold standard for literature-based discovery. The evaluation procedure used in this experiment differs from the original Swanson's method and the RaJoLink method in that a human expert was not involved.

Magnesium deficiency has been shown in several studies to cause migraine headaches (e.g., Swanson 1990; Thomas et al. 1992; Thomas et al. 2000; Demirkaya et al. 2001; Trauninger et al. 2002). In the literature-based closed discovery process Swanson managed to find more than 60 pairs of articles connecting the migraine domain with the magnesium deficiency via several bridging concepts. His closer inspection of the literature about migraine and the literature about magnesium showed that 11 pairs of documents, when put together, provided confirmation of a hypothesis that magnesium deficiency may cause migraine headaches (Swanson 1990). Some of the detected bridging terms are shown in Figure 1.

Similar to Swanson's original study of the migraine literature (Swanson 1988) we used titles as input to our closed discovery process. We performed the experiments on a subset of PubMed titles of articles that were published before 1988 (i.e., before Swanson's literature-based discovery of the migraine-magnesium relation) and were retrieved with the PubMed search engine. As a result we got 2,425 migraine and 5,633 magnesium titles of PubMed articles. These article titles were preprocessed with standard text mining techniques resulting in 13,525 distinct terms which were analyzed and scored by presented elementary heuristics. Each heuristic assigned a score to every term from the list. Afterwards we sorted all 7 lists (6 elementary heuristics and the baseline heuristic) and thus, created 7 ranked lists of terms. Among these 13,525 terms, there were also all 43 terms which Swanson (1988) marked as *b-terms* and which we hoped to propagate to the top of the ranked list using the designed heuristics methodology. The *b*-terms identified by Swanson, verified by the expert to provide new discoveries in the field, are used as a gold standard in the evaluation in this work.

We compared the heuristics based on their ROC (Receiver Operating Characteristic) curves and AUC (Area Under ROC) analysis. The idea underlying ROC curve construction is the following: go from the beginning of a ranked list and every time a *b*-term is seen, draw line up on the ROC canvas, otherwise draw line right. The ideal curve (when all *b*-terms are at the very beginning of a ranked list) would go straight up to the top followed by straight right section to the rightmost part of graph. Area under the ideal ROC curve is equal to 1 when both scales are normalized.

ROC analysis (see Figure 2) shows the performance of elementary heuristics on the migraine-magnesium gold standard dataset. Details on heuristics evaluation can be found in (Juršič et al. 2012), while the main observations and results are outlined below. It can be observed that some heuristics are really well constructed for the purpose of *b*-term discovery. We are especially satisfied with heuristics which have good performance at the start of the ranked list, e.g., heuristic *outFreqRelRF* places four *b*-terms already among the first 50 terms in its ranked list, while the *random* approach finds less than one *b*-term among its first 200 terms. On the other hand some heuristics do not perform so well at the start of the list, e.g., *outFreqSum* and *tfidfDomnSum* do not look promising at the first sight. However, we included them into the set of six heuristics on the basis of complementarity–so that they fit together well when used in the ensemble heuristics–providing not only better performance but also greater robustness of the ensemble.

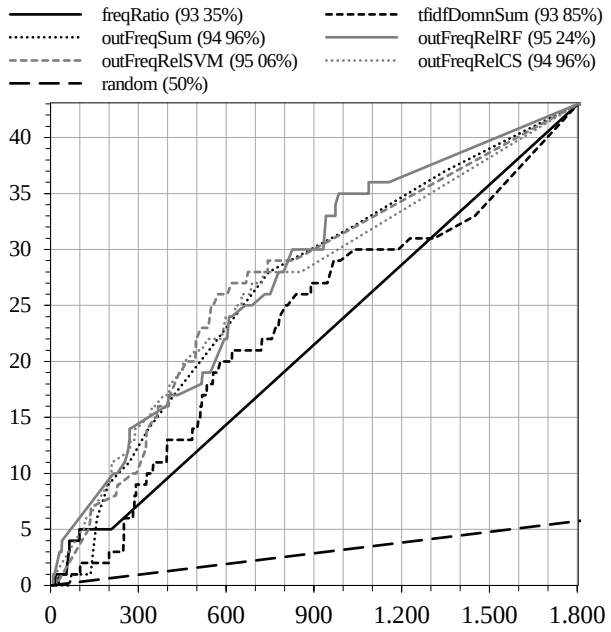


Figure 2. ROC curves representing the performance of elementary heuristics on the learning (migraine-magnesium) dataset.

The Ensemble Heuristic

The ensemble heuristic is a heuristic which combines results of the selected elementary heuristics (*outFreqRelRF*, *outFreqRelSVM*, *outFreqRelCS*, *outFreqSum*, *tfidfDomnSum*, and, *freqRatio*) into an aggregated result. In principle, the ensemble heuristic score is a sum of two parts: the

ensemble voting score and the *ensemble position score* and is computed as: $s_t = s_t^{vote} + s_t^{pos}$.

1. *Ensemble voting score* (s_t^{vote}) of term *t* is based on the number of times the term appears in the first third of the elementary heuristics ranked lists. Each selected base heuristic h_i gives one vote ($s_{t_j,h_i}^{vote} = 1$) to each term which is in the first third in its ranked list of terms and zero votes to all the other terms ($s_{t_j,h_i}^{vote} = 0$). Formally, the ensemble voting score of a term t_j that is at position p_j in the ranked list of n terms is computed as a sum of individual heuristics' voting scores:
$$s_{t_j}^{vote} = \sum_{i=1}^{k} s_{t_j,h_i}^{vote} = \sum_{i=1}^{k} \begin{cases} 1: p_j < n/3, \\ 0: otherwise \end{cases}.$$
Therefore, each term can get a score $s_{t_j}^{vote} \in \{0, 1, 2, \dots, k\}$, where k is the number of base heuristics used in the ensemble.
2. *Ensemble position score* (s_t^{pos}) of term *t* is based on an average position of the term in the elementary heuristics ranked lists. For each heuristic h_i, the term's position score s_{t_j,h_i}^{pos} is calculated as $(n - p_j)/n$, which result in position scores being in the interval $[0,1)$. For an ensemble of k heuristics, the ensemble position score is computed as an average of individual heuristics' position scores:
$$s_{t_j}^{pos} = \frac{1}{k}\sum_{i=1}^{k} s_{t_j,h_i}^{pos} = \frac{1}{k}\sum_{i=1}^{k} \frac{(n - p_j)}{n}.$$

Using the migraine-magnesium domain pair, we experimentally confirmed–through the ROC curve evaluation of different heuristics in terms of the quality of *b*-term retrieval–that the ensemble heuristic is the best measure for *b*-term detection and is able to retrieve *b*-terms approximately 7 times faster compared to the *random* approach. Besides testing on the migraine-magnesium dataset we evaluated the ensemble heuristic also on an independent autism-calcineurin dataset (Petrič et al. 2009) and confirmed the utility and domain independence of the proposed approach.

The CrossBee System

This section presents our system which helps the experts in searching for hidden links that connect two seemingly unrelated domains. We designed and implemented an online system named CrossBee (Cross-Context Bisociation Explorer)[4]. The system was first designed as an online implementation of the ensemble ranking methodology. To the core functionality we have however added other functionalities and content presentations which effectively turned CrossBee into a user-friendly tool for ranking and exploration of bisociative terms that have the potential for cross-context link discovery. This enables the user not only to spot but also to efficiently investigate terms that represent potential cross-domain links.

Below we describe a typical use-case and the extended system's functionality.

A Typical CrossBee Use Case

The most standard use case is the following. The user starts at the system's home page by inputting two sets of documents of interest and by tuning the parameters of the system. The minimal required user's input at this point is a file with the documents from two domains. The prescribed format of the input file is kept simple to enable all users, regardless of their computing skills, to prepare the files. Each line of the file contains exactly three tab-separated entries: (a) document identification number, (b) domain acronym, and (c) the document text. The other options available to the user include specifying the exact preprocessing options, specifying the base heuristics to be used in the ensemble, specifying outlier documents identified by an external outlier detection software, defining the already known *b*-terms, and others. When the user selects all the desired options he proceeds to the next step.

CrossBee then starts a computationally very intensive step in which it prepares all the data needed for the fast subsequent exploration phase. During this step the actual text preprocessing, base heuristics, ensemble, bisociation scores and rankings are computed in the way presented in the previous section. This step does not require any user intervention.

After computation, the user is presented with a ranked list of *b*-term candidates. The list provides the user with some additional information including the ensemble's individual base heuristics votes and term's domain occurrence statistics in both domains. The user then browses through the list and chooses the term(s) he believes to be promising for finding meaningful connections between the two domains.

At this point, the user can start inspecting the actual appearances of the selected term in both domains, using the efficient side-by-side document inspection as shown in Figure 3. In this way, he can verify whether his rationale behind selecting this term as a bridging term can be justified based on the contents of the inspected documents.

The most important result of the exploration procedure is a proof for a chosen term to be an actual bridge between the two domains, based on supporting facts from the documents. As experienced in sessions with the experts, the identified documents are an important result as well, as they usually turn out to be a valuable source of information providing a deeper insight into the discovered terms which indicate new cross-domain relations.

Extended CrossBee Functionality

Below we list the implemented functionalities of the CrossBee system.

[4] CrossBee is available at website: http://crossbee.ijs.si/.

- *Document focused exploration* empowers the user to filter and order the documents by various criteria. The user can find it more pleasing to start exploring the domains by reading documents and not browsing through the term lists. The ensemble ranking can be used to propose the user which documents to read by suggesting those with the highest proportion of highly ranked terms.
- *Detailed document view* provides a more detailed presentation of a single document including various term statistics and a similarity graph showing the similarity between this document and other documents from the dataset.
- *Methodology performance analysis* supports the evaluation of the methodology by providing various data which can be used to measure the quality of the results, e.g., data for plotting the ROC curves.
- *High-ranked term emphasis* marks the terms according to their bisociation score calculated by the ensemble heuristic. When using this feature all high-ranked terms are emphasized throughout the whole application making them easier to spot.
- *b-term emphasis* marks the terms defined as *b*-terms by the user.
- *Domain separation* is a simple but effective option which colors all the documents from the same domain with the same color, making an obvious distinction between the documents from the two domains.
- *UI customization* enables the user to decrease or increase the intensity of the following features: high-ranked term emphasis, *b*-term emphasis and domain separation. In cooperation with the experts, we discovered that some of them do like the emphasizing features while the others do not. Therefore, we introduced the UI customization where everybody can set the intensity of these features according to their preferences.

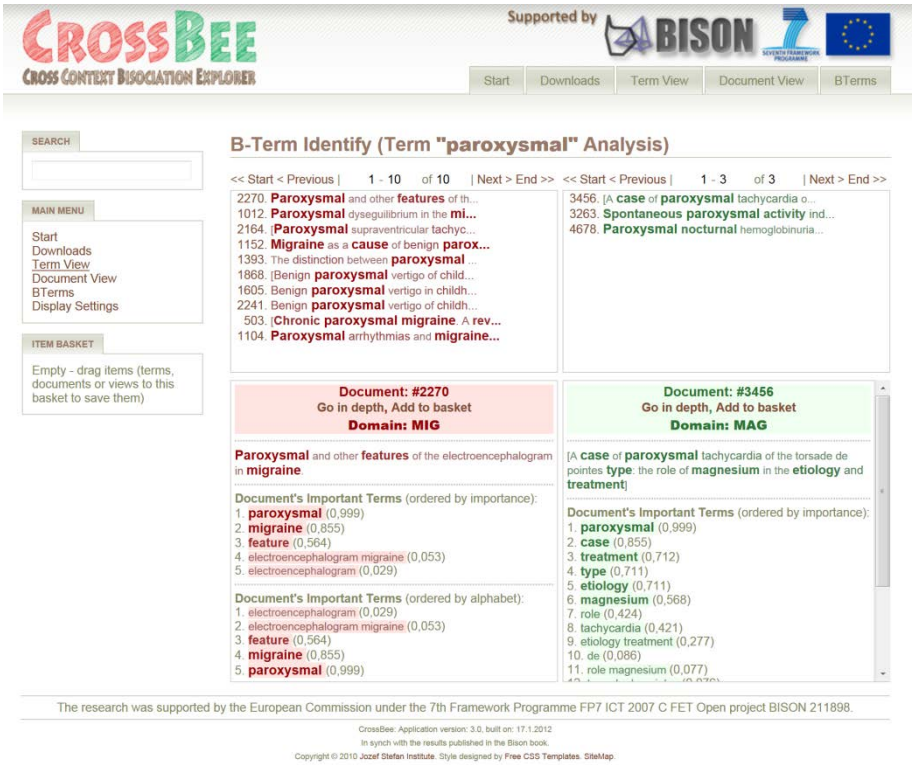


Figure 3. Illustration of the s*ide-by-side document inspection of potential cross-domain links* functionality of CrossBee, using an example from the migraine-magnesium dataset analysis, focusing on the analysis of the *paroxysmal* term.

Discussion and Further Work

Current literature-based approaches depend strictly on simple, associative information search. Commonly, a literature-based association is computed using measures of similarity or co-occurrence. Because of their 'hard-wired' underlying criteria of co-occurrence or similarity, association-based methods often fail to discover relevant information which is not related in obvious associative ways. Especially information related across separate domains is hard to identify with the conventional associative approaches. In such cases the domain-crossing connections, called bisociations (Berthold, 2012), can help generate creative and innovative discoveries.

There was previous research by Swanson (1986), Weeber et al. (2001), Petrič et al. (2009) and several other

authors investigating the means for finding novel interesting connections between disparate research findings which can be extracted from the published literature. They have shown that the analysis of implicit cross-context associations hidden in scientific literature can guide hypotheses formulation and lead to the discovery of new knowledge.

The methodology presented in this paper has the potential for improved computational creativity in supporting the expert in the task of cross-domain literature mining. The main novelty is an approach to ensemble-based bridging term ranking. The creative act of finding bridging terms is supported by the user-friendly CrossBee system for literature mining, implementing closed cross-domain link discovery. It has the potential to identify bridging concepts in the intersection of different domain literatures, as confirmed in the experiments in mining the literature on migraine and magnesium

In further work we will apply the CrossBee system to new domain pairs, focusing on the system's potential to lead to new scientific discoveries. In addition to linking to PubMed, we will explore also the ways to connect CrossBee to other document sources, including its connection to keyword search from documents on the web, Moreover, it would be interesting to explore the potential of CrossBee in media research, as well as linguistics where metaphors could potentially be discovered by cross-context text mining. One of the priorities of our work will be, however, to use CrossBee in collaboration with the experts from different fields (e.g. physicists and biologists) to address real life domain problems and to get valuable feedback from these targeted users.

References

Berthold, M., ed. 2012. *Bisociative Knowledge Discovery*. Springer 2012 (in press).

Boden, M. 1992. *The Creative Mind*. London: Abacus.

Demirkaya, S.; Vural, O.; Dora, B.; and Topcuoglu, M.A. 2001. Efficacy of intravenous magnesium sulfate in the treatment of acute migraine attacks. *Headache* 41(2): 171-177.

Feldman, R. and Sanger, J. 2007. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridge University Press.

Juršič, M.; Sluban, B.; Cestnik, B.; Grčar, M.; and Lavrač, N. 2012. Bridging concept identification for constructing information networks from text documents. In: Berthold, M.R. ed., *Bisociative Knowledge Discovery.* Springer LNAI 7250 (in press).

Koestler, A. 1964. *The Act of Creation*. New York: MacMillan.

Mednick, S.A. 1962. The associative basis of the creative process. *Psychol. Rev*. 69: 220-232.

Petrič, I.; Urbančič, T.; Cestnik, B.; and Macedoni-Lukšič, M. (2009) Literature mining method RaJoLink for uncovering relations between biomedical concepts. *J. Biomed. Inform.* 42(2): 219-227.

Roberts, R.M. 1989. *Serendipity: Accidental Discoveries in Science.* Wiley.

Sluban, B.; Juršič, M.; Cestnik, B.; and Lavrač, N. 2012. Exploring the power of outliers for cross-domain literature mining. In: Berthold, M.R. ed., *Bisociative Knowledge Discovery.* Springer LNAI 7250 (in press).

Smalheiser, N.R., and Swanson, D.R. 1998. Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Comput. Methods Programs Biomed.* 57(3): 149-153.

Swanson, D.R. 1986. Undiscovered public knowledge. *Library Quarterly* 56(2): 103-118.

Swanson, D.R. 1988. Migraine and magnesium: Eleven neglected connections. *Perspectives in Biology and Medicine* 31(4): 526–557

Swanson, D.R. 1990. Medical literature as a potential source of new knowledge. *Bull. Med. Libr. Assoc. 78(1)*: 29–37.

Swanson, D.R.; Smalheiser, N.R.; and Torvik, V.I. 2006. Ranking indirect connections in literature-based discovery: The role of Medical Subject Headings (MeSH). *J. Am. Soc. Inf. Sci. Tec.* 57(11): 1427-1439.

Thomas, J.; Millot, J.M.; Sebille, S.; Delabroise, A.M.; Thomas, E.; Manfait, M.; and Arnaud, M.J. 2000. Free and total magnesium in lymphocytes of migraine patients - effect of magnesium-rich mineral water intake. *Clin. Chim. Acta* 295(1-2): 63-75.

Thomas, J.; Thomas, E.; and Tomb, E. 1992. Serum and erythrocyte magnesium concentrations and migraine. *Magnes. Res.* 5(2): 127-130.

Trauninger, A.; Pfund, Z.; Koszegi, T.; and Czopf, J. 2002. Oral magnesium load test in patients with migraine. *Headache* 42(2): 114-119.

Urbančič, T.; Petrič, I.; Cestnik, B.; and Macedoni-Lukšič, M. 2007. Literature mining: towards better understanding of autism. In: Bellazzi, R.; Abu-Hanna, A.; and Hunter, J., eds., In *Proceedings of the 11th Conference on Artificial Intelligence in Medicine in Europe*, 217-226. Springer.

Urbančič, T.; Petrič, I.; and Cestnik, B. 2009. RaJoLink: A method for finding seeds of future discoveries in nowadays literature. In: Rauch, J., ed., *Foundations of Intelligent Systems*. LNAI, 5722. 129-138. Springer.

Weeber, M.; Vos, R.; Klein, H.; and de Jong-van den Berg, L.T.W. 2001. Using concepts in literature-based discovery: Simulating Swanson's Raynaud–fish oil and migraine–magnesium discoveries. *J. Am. Soc. Inf. Sci. Tech.* 52(7): 548-557.

Wiggins, G. A. 2006. A Preliminary Framework for Description, Analysis and Comparison of Creative Systems. *Journal of Knowledge Based Systems* 19(7): 449–458.

A closer look at creativity as search

Graeme Ritchie
Computing Science
University of Aberdeen
Aberdeen AB24 3UE
g.ritchie:abdn.ac.uk

Abstract

Several papers by Wiggins (building on ideas by Boden) have outlined a view of creative concept generation as a very general search process, but that formalisation has not been developed much in the past few years. Also, there are some aspects where clarification or spelling out of details would be useful. We present a re-formulation of the central ideas in Wiggins's framework, with slightly more rigorous statements of the definitions and a number of minor extensions. We also explain how this framework relates to some hitherto completely separate proposals by Ritchie.

Introduction

In recent years, there have been various formalisations of aspects of the computational creative process ((Pease, Winterstein, and Colton 2001), (Colton, Pease, and Ritchie 2001), (García et al. 2006), (Thornton 2007), (Colton, Charnley, and Pease 2011)). Hence there is a consensus among at least some established researchers that it is methodologically beneficial to have fully precise, detailed and formal accounts of any mechanisms being considered as 'creative'.

A prominent example is Wiggins's *creative systems framework* (CSF), presented in a number of papers (particularly Wiggins(2006a; 2006b)). That framework emphasises the notion of *search* as the central mechanism for simulating creativity, and outlines how a metalevel search could represent some phenomena sometimes discussed as 'transformational' creativity. Although these ideas are very helpful in clarifying the nature of creative computation, the published versions of the CSF are at best a preliminary sketch: some details are unspecified, some natural extensions are undeveloped, and there are some formal errors or infelicities. The current paper starts from the central ideas of the CSF, but redefines the formal mechanisms in a way which leaves fewer gaps, aspires to have fewer formal inconsistencies, and includes the description of more aspects of computationally creative processes. The central motivation for this is that, if we subscribe to a belief in the benefits of formal models (as noted above), then these models should not be left undeveloped, but should continue to be maintained, repaired, and extended as necessary.

It is important to realise that the underlying intuitive ideas – creation as the exploration of a 'conceptual space', and possible 'transformation' of that space – have been set out in numerous articles by Boden, with many illustrative examples from human creativity. Wiggins's contribution was to take those informal, broadbrush ideas and outline a formal framework which both captured the core notions and made sense computationally. The reader is referred to publications by Boden, Wiggins, and many others for more about the intuitive motivation; our aim here is to refine and extend Wiggins's proposals.

A summary of Wiggins's CSF

Although this paper is centrally concerned with formalisation, we start with a very brief informal overview of the ideas in the existing version of the CSF.

The framework posits a universal set of all *concepts*, a term which covers both abstract ideas (e.g. a mathematical theorem, a design for a better political system) and concrete artefacts (e.g. a painting, a poem). Within this wide-ranging set, there are particular types of idea/artefact (e.g. stories, paintings, poems), and what counts as a recognisable example of a story/painting/poem/etc. may be highly dependent on socio-cultural norms. For many such creative genres, it is not realistic for there to be a firm definition of acceptability, as the specific concept may be *vague* in the sense of (van Deemter 2010). That is, the extent to which a text is or is not a well-formed story (or other artistic category) is a matter of degree, rather than an all-or-nothing decision. Hence, within the CSF, the criterion for acceptability/recognisability is represented as a rating between 0 and 1; in effect, the set of examples of a genre is treated as a *fuzzy set*. As well as whether something falls within the definition of some artistic genre, there is the separate question of whether it is a *good* (high quality) instance (e.g. a profound poem, a beautiful painting). This is similarly a 'vague' notion, and again is represented within the CSF as a score between 0 and 1. This means that an artefact can be an acceptable instance (e.g. recognisably a story) without being of high quality (e.g. it may be a poor story); hence the need for two different ratings (mappings from concepts to values).

The inspiration for the CSF was the work of Boden(1998; 2003), in which creativity was described as occurring within a *conceptual space*, which could be *explored*, or – in more radical creativity – *transformed* into a new space. This view has some resemblance to the traditional ideas of search

within AI (Nilsson 1971), and Wiggins set out both to clarify the relationship between conventional AI search and creative computation, and to provide a formal framework for describing the latter. The idea is that a creative system starts from some set of concepts, and by a series of steps creates further concepts one after another, thus 'exploring the space'. Although the term 'creative' has connotations of 'construction', the CSF, following the practice in formalising AI search, regards 'new' concepts as not so much 'constructed' as 'reached'. That is, notionally all the possible concepts are elements of some universal set, but the creative system computes a route through that set to particular concepts, and those which have been thus reached represent 'discoveries' or 'inventions'.

The exploration of the space of concepts (the search method) is modelled by an operator which acts on a sequence of concepts (a list of the concepts the system has already processed), and yields as its result a new (presumably longer) list of concepts, which can then be processed in turn as the next cycle of search. Sequences are used because the search method, like an agenda-based AI search system, has to maintain some record of where it has reached within the space of available concepts, and what to work on next.

Wiggins points out that by separating the acceptability rating from the search method, we can describe the situation where different composers, each with a personal way of finding new artefacts (different search procedures) are working within the same style of music (a shared notion of acceptability). A creative agent should be able to recognise something as being a recognisable artefact, or a high quality artefact, without necessarily having a search method that would allow the agent to reach (create) that artefact.

More concretely, this is an intuitively plausible account of certain potentially creative programs. The MCGONAGALL poetry generator (Manurung, Ritchie, and Thompson 2012) uses an explicit search mechanism (a genetic algorithm) to find suitable candidate texts. Each text must at least be syntactically well-formed according to the system's linguistic grammar; this could be regarded as the acceptability mapping. During the search, items are scored for rhythmic suitability and proximity to an initial semantic message; this would be the quality mapping. At each stage, the system keeps an ordered list of the current candidates, from which each cycle in the search starts.

A small formal detail is that the CSF search operator takes as arguments the two fuzzy criteria (for acceptability and quality), and from there computes a way of going from an existing concept-sequence to a new concept-sequence. This means that the search method can be sensitive to these two criteria if necessary, or that we can describe two systems as having the same search strategy relative to their own different definitions of validity and quality.

For Wiggins, the mappings (the two fuzzy sets and the search mechanism) are to be represented as expressions in some symbolic language, translatable to actual mappings.

Hence, in the CSF, an *exploratory creative system* consists of the following seven components:

(i) the universal set of concepts
(ii) the language for expressing the relevant mappings
(iii) a symbolic representation of the acceptability mapping
(iv) a symbolic representation of the quality mappings
(v) a symbolic representation of the search mechanism
(vi) an interpreter for expressions like (iii) and (iv)
(vii) an interpreter for expressions like (v)

That constitutes the *object level* of the creative system, which searches through concepts in the domain (e.g. melodies). Wiggins also proposes that there can be a *metalevel*, which searches through possible object level systems to find an interestingly different 'conceptual space', thus modelling Boden's idea of a 'transforming' the space. The metalevel in CSF is structured in the same way as the object level (i.e. the seven parts as set out above), except that its set for exploration (i.e. its universal set) contains expressions in the symbolic language used at the object level. In this way, the metalevel searches through expressions describing object-level systems, assessing these descriptions for acceptability and for quality (using the metalevel's criteria for these two measures).

Relative to the published accounts of the CSF, the revisions or extensions made here are:

- The symbolic language for expressing the various mappings is given a much less central role.
- The way in which the metalevel defines the object level is explicitly stated. In particular, the notion of 'transformation' of an (object-level) space is defined.
- Some minor inconsistencies in definitions are removed.
- We outline how search methods within the CSF can be compared at the metalevel.
- The CSF is related to a proposal for formal assessment of creative systems (Ritchie 2001; 2007).

The object level

The structure of an object level system

Wiggins posits the existence of one universal set, $\mathcal{U}$, the set of all concepts, but then defines a creative system as a tuple, one component of which is the universal set. If the set is truly universal across all systems, it should not need to be mentioned as a defining component of a specific system. On the other hand, it would be useful to be able to allow different creative systems to consider only specific subsets of this hugely general set. The compromise here is to accept the existence of the wholly universal set, but for the definition of each creative system to specify a *subset* of this universe; this could, in principle, be a non-proper subset. We will use $\mathcal{P}$ (mnemonic for 'possibilities') for these subsets in our definitions, below. The idea is that $\mathcal{U}$ is universal enough to contain concepts for every type of artefact that might ever be conceived: it includes poems, stories, sculptures, jokes, paintings, theorems, architectural plans, designs for food mixers, etc. On the other hand $\mathcal{P}$ represents the whole range of concepts within some narrower sphere, such as two-dimensional arrays of coloured pixels (which could act as a 'universal' set for the creation of visual art), or finite sequences of words and punctuation (a possible 'universal' set for various textual artistic forms).

Notation: For any sets A, B, B^A denotes the set of mappings from A to B. In particular, for any set X, $[0,1]^X$ denotes the set of mappings from X to values between 0 and 1 inclusive. Since a fuzzy set is defined by a mapping from possible elements to values betwen 0 and 1, our fuzzy sets of 'acceptable' elements and of 'valuable' elements will be stated in this way; that is, as members of $[0,1]^{\mathcal{P}}$. We also take $tuples(X)$ to denote the set of finite tuples (of any length) of elements of the set X.

Definition 1: An *exploratory creative system* comprises:

(i) a subset $\mathcal{P}$ of $\mathcal{U}$ (possible concepts within this type or genre)

(ii) $\mathcal{N} \in [0,1]^{\mathcal{P}}$, the *acceptability mapping* (mnemonically, this describes **n***orms*)

(iii) $\mathcal{V} \in [0,1]^{\mathcal{P}}$, the *value mapping* (mnemonically, this describes **v***alue*)

(iv) a mapping $\mathcal{Q}$ (the *exploration scheme*) from $[0,1]^{\mathcal{P}} \times [0,1]^{\mathcal{P}}$ to the set of mappings from $tuples(\mathcal{P})$ to $tuples(\mathcal{P})$ (mnemonically, this describes a **q***uest* for creations – 's' for 'search' is used elsewhere).

The four components in our definition are direct counterparts of those in Wiggins's CSF, but we have chosen different symbols for the components of a system, to avoid confusion; the relationships to Wiggins's notation are: $\mathcal{N} \cong [\![\mathcal{R}]\!]$, $\mathcal{V} \cong [\![\mathcal{E}]\!]$, $\mathcal{Q}(\mathcal{N},\mathcal{V}) \cong \langle\langle \mathcal{R}, \mathcal{T}, \mathcal{E} \rangle\rangle$. The intuitive meanings of the components are the same: $\mathcal{P}$ is the set of possible concepts (e.g. arrays of pixels, sequences of words), $\mathcal{N}$ defines the fuzzy set of acceptable instances of whatever domain/genre is being explored, $\mathcal{V}$ indicates how 'good' an instance is, and $\mathcal{Q}$ defines how to explore the space.

The component $\mathcal{Q}$, the search method, may need some explanation. Directly following Wiggins's proposals, $\mathcal{Q}$ is applied to a particular $\mathcal{N}$ and $\mathcal{V}$, and from that produces a mapping which takes sequences of concepts into sequences of concepts; hence, $\mathcal{N}$ and $\mathcal{V}$ could in principle influence $\mathcal{Q}$, or could be ignored. It might seem odd to describe $\mathcal{Q}$ as taking these two parameters, when the only possible values for the parameters seem to be fixed as $\mathcal{N}$ and $\mathcal{V}$ – why not just 'compile in' these two values, as they are specified in the same 4-tuple package as $\mathcal{Q}$? At present, this level of parameterisation has no real advantage, but it leaves open the possibility, as the framework is elaborated, of considering a 'transformed' version of an exploratory creative system in which $\mathcal{Q}$ is unchanged, but one or both of $\mathcal{N}$, $\mathcal{V}$ are altered, with automatic consequences for the operation of $\mathcal{Q}$.

As in the original Wiggins formulation, $\mathcal{Q}(\mathcal{N}, \mathcal{V})$ maps from *sequences* of concepts to *sequences* of concepts, providing an agenda-like exploration of the set of possibilities, with the sequence representing the current search state

As noted earlier, the CSF includes a symbolic language in which mappings are expressed as rules, which are then interpreted into mappings. Here, we abstract away from the use of a language, and define a creative system using mappings. The advantage of this is that it states the essential relations within a creative system without regard for representational issues. In a later section, we show how the symbolic language can be incorporated, directly reflecting the Wiggins approach.

Characterising the conceptual space

As noted above, the basic definition of an exploratory creative system contains a fuzzy set, $\mathcal{N}$, of concepts, which are – intuitively – those concepts which conform to the current norms of the domain. In Wiggins's formulation, this fuzzy set is *not* regarded as modelling Boden's 'conceptual space'. Instead, Wiggins stipulates that conceptual spaces are ordinary subsets (*not* fuzzy) of the universal set $\mathcal{U}$, and that the conceptual space $\mathcal{C}$ (of a given creative system) consists of all concepts mapped by his $[\![R]\!]$ (our counterpart is $\mathcal{N}$) to values greater than or equal to 0.5. Similarly, Wiggins defines the *valued set* as those concepts which his $[\![\mathcal{E}]\!]$ (our equivalent is $\mathcal{V}$) maps to values greater than or equal to 0.5. That is, although the CSF allows both these sets to have graded membership (0 to 1), Wiggins immediately simplifies them to non-graded sets by imposing a threshold.

Within our formulation, the counterpart to Wiggins's non-fuzzy definitions would be as follows:

Definition 2: Given an exploratory creative system $S = (\mathcal{P}, \mathcal{N}, \mathcal{V}, \mathcal{Q})$, we define, for any $\alpha \in [0,1]$ and $X \subseteq \mathcal{P}$:

(i) $\mathcal{N}_\alpha(X) = \{c \in X \mid \mathcal{N}(c) > \alpha\}$ (the set of concepts which reach the threshold α in their 'normality').

(ii) $\mathcal{V}_\alpha(X) = \{c \in X \mid \mathcal{V}(c) > \alpha\}$ (the set of concepts which reach the threshold α in their 'quality').

(iii) the *fuzzy conceptual space* of S is $\mathcal{N}$ (this just confirms the status of $\mathcal{N}$ as outlined earlier).

(iv) the *conceptual space* of S is $\mathcal{N}_{0.5}(\mathcal{P})$ (this is for backwards compatibility with Wiggins's 0.5 threshold).

(v) the *fuzzy valued set* of S is $\mathcal{V}$ (this just confirms the status of $\mathcal{V}$ as outlined earlier)..

(vi) the *valued set* of S is $\mathcal{V}_{0.5}(\mathcal{P})$ (this is for backwards compatibility with Wiggins's 0.5 threshold).

Searching

In the CSF, the searching process begins from an initial set of concepts. This is to allow for the situation where the creative system starts from some given concept set, representing the status quo. It is also useful later, when considering metalevel search. In Wiggins's definitions, exploration always starts from the single totally unspecified concept, $\top$, representing a situation in which the system has no known concepts already. Here, we generalise this slightly.

If the sequences on which $\mathcal{Q}$ operates are to be like an agenda in conventional AI search, then those sequences should contain only items which have been produced from previous steps of the search (i.e. applications of $\mathcal{Q}$). This means that the initial agenda has to include every item (concept) which could ever participate in discoveries but is not produced by an application of $\mathcal{Q}$.

Wiggins defines 'reachable' concepts using indefinitely many applications of the search operator. There is a minor slip in his definition, in that repeated applications of $\langle\langle \mathcal{R}, \mathcal{T}, \mathcal{E} \rangle\rangle$ (which corresponds to our $\mathcal{Q}$) will compute *sequences* (tuples) of concepts, not individual concepts. This is easily remedied (and we can add an intermediate version, for limited search). First we need a minor definition of all the items appearing within a set of tuples:

Definition 3: Given any set of tuples A, we define $elements(A) \equiv$
$\{x \mid \exists \langle y_1, \ldots, y_n \rangle \in A, \exists i\ 1 \leq i \leq n : x = y_i\}$

The Wiggins formalisation defines all the concepts which can be reached with any amount of search, i.e. without limit. Although this case is of theoretical interest, in practice any system will search only to some finite depth, and so we also define the notion of 'reachable in a fixed number of steps', relying on the fact that a single application of the search mapping $\mathcal{Q}$ corresponds to one step in the search process.

Definition 4: Given an exploratory creative system $(\mathcal{P}, \mathcal{N}, \mathcal{V}, \mathcal{Q})$, and a set $B \subset \mathcal{P}$ of concepts (B is the starting set of concepts for the search):

(i) the set *reachable from B in m steps* is
$\bigcup_{n=0}^{m} elements(\mathcal{Q}(\mathcal{N}, \mathcal{V})^n(B))$
(i.e. any number of repeated applications of $\mathcal{Q}$, up to m; this describes search up to some depth.)

(ii) the set *reachable from B* is
$\bigcup_{n=0}^{\infty} elements(\mathcal{Q}(\mathcal{N}, \mathcal{V})^n(B))$
(i.e. any number of repeated applications of $\mathcal{Q}$; this allows any depth of search.)

(iii) the set of *reachable concepts* is the set reachable from $\{\top\}$. (This matches Wiggins's notion, where all search starts from a single unspecified concept).

(iv) the set of *concepts reachable in m steps* is the set reachable from $\{\top\}$ in m steps. (This is a bounded search variant of Wiggins's 'start from nothing' definition.)

In considering the behaviour of a creative system, it is important to know which of its final output (i.e. creations) were provided to it initially and which were computed by the system itself. We can define these thus:

Definition 5: Given an exploratory creative system $(\mathcal{P}, \mathcal{N}, \mathcal{V}, \mathcal{Q})$, a subset B of $\mathcal{P}$, and a set of concepts K reachable from B, the *discoveries in K* is the set of concepts in $K - B$.

Wiggins defines the set of valued concepts as being those concepts reachable from the undefined concept, $\top$, which exceed a particular threshold value (0.5) when his 'value' mapping (our $\mathcal{V}$) is applied. That definition can be restated in the terminology here.

Definition 6: Given an exploratory creative system $(\mathcal{P}, \mathcal{N}, \mathcal{V}, \mathcal{Q})$, where RC is the set of reachable concepts, and a value $\alpha \in [0, 1]$:

(i) the *α-valued set of reachable concepts* is $\mathcal{V}_\alpha(RC)$.

(ii) the *valued set of reachable concepts* is the 0.5-valued set of reachable concepts (i.e. $\mathcal{V}_{0.5}(RC)$); this mirrors Wiggins's definition.

The metalevel

Structure of the metalevel

For the metalevel, the first matter to be clarified concerns the set of items used for exploration. In the Wiggins papers, this is the set of possible expressions in a symbolic language $\mathcal{L}$. An expression in $\mathcal{L}$ is defined earlier as defining a rule-set representing either $\mathcal{R}$ (acceptability rules), $\mathcal{E}$ (value rules) or $\mathcal{T}$ (search rules), with different interpreters applying depending on which of these is intended. If the metalevel is considering single $\mathcal{L}$-expressions, how does one such expression represent an entire object level system, which contains all three of $\mathcal{R}$, $\mathcal{E}$, and $\mathcal{T}$? Within the language-based formalisation, a possible response would be to say that $\mathcal{L}$ must contain notation which allows one expression to represent *three* rule sets. If following this path, Wiggins's definitions of the language interpreters would also have to be patched. As we are separating out the language aspect, we have a different solution.

Here, the object level space is defined by the mappings $\mathcal{N}$, $\mathcal{V}$, $\mathcal{Q}$, so it seems reasonable to have the metalevel searching through triples (N, V, Q) where N, V, Q, are possible values for $\mathcal{N}$, $\mathcal{V}$, $\mathcal{Q}$ respectively (and hence are elements of the appropriate sets such as $[0, 1]^{\mathcal{P}}$). The exploration set at the metalevel will be the set of such triples; for brevity here, we will call this set of triples $ECS(\mathcal{P})$ (as it is the set of possible *exploratory creative systems* for $\mathcal{P}$).

Given a subset $\mathcal{P}$ of $\mathcal{U}$, a *metalevel creative system* for $\mathcal{P}$ is an exploratory creative system made up of:

(i) $ECS(\mathcal{P})$ (i.e. this is the metalevel's set to explore)

(ii) an element $\mathcal{N}^{meta}$ of $[0, 1]^{ECS(\mathcal{P})}$; this rates potential object-level systems as to how 'normal' they are, thus providing a (fuzzy) set of 'acceptable' triples (N,V,Q).

(iii) an element $\mathcal{V}^{meta}$ of $[0, 1]^{ECS(\mathcal{P})}$; this rates potential object-level systems as to their 'quality', thus providing a (fuzzy) set of 'valuable' triples (N,V,Q).

(iv) a mapping $\mathcal{Q}^{meta}$ from $[0, 1]^{ECS(\mathcal{P})} \times [0, 1]^{ECS(\mathcal{P})}$ to the set of mappings from $tuples(ECS(\mathcal{P}))$ to $tuples(ECS(\mathcal{P}))$; this is structured like the search device $\mathcal{Q}$ in an object-level creative system, but operates on elements of $ECS(\mathcal{P})$ instead of $\mathcal{P}$.

That is, the structure at the metalevel is exactly parallel to the structure at the object level, as in the Wiggins version. A metalevel has information about what an object level creative system should look like ($\mathcal{N}^{meta}$) and what would count as a 'good' object level system ($\mathcal{V}^{meta}$). It also contains a way of searching through potential object level systems ($\mathcal{Q}^{meta}$).

Characterising an object level system

Given a definition of the components of a metalevel, it is essential to then define exactly how the parts of the metalevel characterise an object level system. This is not discussed in detail by Wiggins, but he indicates that the metalevel is to operate (in terms of search, etc.) exactly as an object level creative system.

An object level creative system will ascribe various characteristics to a set of concepts. Each concept will be: rated (by $\mathcal{N}$) as to how acceptable it is as an member of the conceptual space in question, rated (by $\mathcal{V}$) as to its value/quality, and defined (by $\mathcal{Q}$) as either reachable or not. As noted earlier, Wiggins proposes that the ratings by (his equivalents of) $\mathcal{N}$ and $\mathcal{V}$ are turned into non-fuzzy sets using a threshold. However, even then the object level does not characterise a

single object, or a unique set of systems: it defines three independent sets, via $\mathcal{N}$, $\mathcal{V}$ and $\mathcal{Q}$. Since the metalevel has exactly the same structure as an object level system, the items which it explores (for Wiggins, expressions in a symbolic language $\mathcal{L}$) are presumably similarly allocated to 3 sets: the recognisable, the valued and the reachable (and for Wiggins, reachability is always relative to $\top$, not some specified starting set of items). Hence, the metalevel is assigning (potential) object level systems to these three categories. What the metalevel does not do is characterise a single object level system, or even a unique set of systems. This means that we do not, from the published papers, have a definition of how one object level system is a transformation of another, or how a computation at the metalevel will yield a new object level system – all that the metalevel provides is this tripartite classification. We will remedy this by defining how a metalevel can define (or transform) a specific object level system.

In the next few definitions, we assume two exploratory creative systems $S_{obj} = (\mathcal{P}, \mathcal{N}, \mathcal{V}, \mathcal{Q})$ and $S'_{obj} = (\mathcal{P}, \mathcal{N}', \mathcal{V}', \mathcal{Q}')$, and a metalevel system $\mathcal{S}^{meta} = (ECS(\mathcal{P}), \mathcal{N}^{meta}, \mathcal{V}^{meta}, \mathcal{Q}^{meta})$ for $\mathcal{P}$. Where the relation '$\neq$' is used here, this allows for the two items in question to have elements in common.

Definition 7:

(i) S'_{obj} is *a revision of* S_{obj} *using* $\mathcal{S}^{meta}$ if S'_{obj} is in the set reachable from $\{S_{obj}\}$ within $\mathcal{S}^{meta}$, and $S'_{obj} \neq S_{obj}$.

(ii) for any $\alpha \in [0, 1]$, S'_{obj} is *α-valued with respect to* $\mathcal{S}^{meta}$ if $\mathcal{V}^{meta}((\mathcal{N}', \mathcal{V}', \mathcal{Q}')) \geq \alpha$.

(iii) S'_{obj} is *a transformation of* S_{obj} *using* $\mathcal{S}^{meta}$ if S'_{obj} is a revision of S_{obj} using $\mathcal{S}^{meta}$ and also $\mathcal{N} \neq \mathcal{N}'$.

Here we have taken a 'transformation' to be a revision in which the definition of the conceptual space (acceptable set) changes, as indicated by the condition '$\mathcal{N} \neq \mathcal{N}'$'. As this could be true even if $\mathcal{N}$ and $\mathcal{N}'$ differ only on one element, proponents of transformation as a form of *radical* change might wish to enhance this definition.

Relationship to the original CSF

To clarify the amendments we have made to the formalisation, we can compare it with the original version in the papers by Wiggins. As mentioned earlier, the original CSF includes a symbolic language in which the components (the counterparts of our $\mathcal{N}$, $\mathcal{V}$, $\mathcal{Q}$) are expressed. We can add this to our framework by defining a symbolic-language version of an exploratory creative system, with appropriate links to the definitions given above. In order to mimic Wiggins's definitions, we first have to clarify certain aspects which are unclear in the published papers. Sometimes the mapping $\mathcal{N}$ (or what corresponds to this in Wiggins's framework) is represented as a *single* expression in a symbolic language, and sometimes it is said to be a *set of* expressions. Either of these accounts could be made to work, if applied consistently. Here, we have opted for the single expression version, with the observation that the symbolic language could contain connective symbols such as 'conjunction', 'disjunction' or other logical operators, thereby getting the effect of a *set* of rules in one syntactic expression.

Wiggins's version does not separate clearly the definition of the language from the specific language expressions used in a particular creative system. We have tried to draw this distinction in the next two definitions.

Definition 8: Given a set of concepts $\mathcal{P}$, a *creative systems language* for $\mathcal{P}$ is a tuple $(\mathcal{A}, \mathcal{L}_{\mathcal{R}}, \mathcal{L}_{\mathcal{T}}, [\![.]\!], \langle\!\langle . \rangle\!\rangle)$ where:

(i) $\mathcal{A}$ is a set of symbols, the alphabet.

(ii) $\mathcal{L}_{\mathcal{R}}$ and $\mathcal{L}_{\mathcal{T}}$ are languages over $\mathcal{A}$ (only 2 are needed because the language $\mathcal{L}_{\mathcal{R}}$ can be used for both the 'acceptability' rules and the 'value' rules, since these both describe fuzzy sets of concepts).

(iii) $[\![.]\!]$ is a mapping from $\mathcal{L}_{\mathcal{R}}$ to $[0, 1]^{\mathcal{P}}$ (this is the interpreter which takes an expression in the symbolic language and returns a mapping; that mapping is then a fuzzy set of concepts).

(iv) $\langle\!\langle . \rangle\!\rangle$ is a mapping from $\mathcal{L}_{\mathcal{R}} \times \mathcal{L}_{\mathcal{T}} \times \mathcal{L}_{\mathcal{R}}$ to $tuples(\mathcal{P})^{tuples(\mathcal{P})}$ (this is the interpreter which turns symbolic expressions specifying a search method into an actual mapping to carry out the search).

The above definition (which is closely modelled on Wiggins's proposals) provides the symbolic mechanisms, separately from any particular creative system which might use these representations. The next two definitions state how these mechanisms can be used to define a specific system.

Definition 9: Given a set of concepts $\mathcal{P}$ and a creative systems language $(\mathcal{A}, \mathcal{L}_{\mathcal{R}}, \mathcal{L}_{\mathcal{T}}, [\![.]\!], \langle\!\langle . \rangle\!\rangle)$ for $\mathcal{P}$, then a *symbolically represented exploratory creative system* for $\mathcal{P}$ consists of a tuple $(\mathcal{W}_{\mathcal{R}}, \mathcal{W}_{\mathcal{E}}, \mathcal{W}_{\mathcal{T}})$ where:

(i) $\mathcal{W}_{\mathcal{R}} \in \mathcal{L}_{\mathcal{R}}$; the norms or acceptability rules.

(ii) $\mathcal{W}_{\mathcal{E}} \in \mathcal{L}_{\mathcal{R}}$; the rules assigning value to items.

(iii) $\mathcal{W}_{\mathcal{T}} \in \mathcal{L}_{\mathcal{T}}$; rules which define the search method.

Definition 10: Given a set of concepts $\mathcal{P}$, a creative systems language $(\mathcal{A}, \mathcal{L}_{\mathcal{R}}, \mathcal{L}_{\mathcal{T}}, [\![.]\!], \langle\!\langle . \rangle\!\rangle)$, and a symbolically represented exploratory creative system $SE = (\mathcal{W}_{\mathcal{R}}, \mathcal{W}_{\mathcal{E}}, \mathcal{W}_{\mathcal{T}})$, then the *exploratory creative system associated with* SE is the tuple $S = (\mathcal{P}, \mathcal{N}, \mathcal{V}, \mathcal{Q})$ where

(i) $\mathcal{N} = [\![\mathcal{W}_{\mathcal{R}}]\!]$; i.e. the meaning of this rule expression is the normality mapping.

(ii) $\mathcal{V} = [\![\mathcal{W}_{\mathcal{E}}]\!]$; i.e. the meaning of this rule expression is the value mapping.

(iii) $\mathcal{Q}(\mathcal{N}, \mathcal{V}) = \langle\!\langle \mathcal{W}_{\mathcal{R}}, \mathcal{W}_{\mathcal{T}}, \mathcal{W}_{\mathcal{E}} \rangle\!\rangle$; i.e. the meanings of these rule expressions give the search mapping.

(This directly mirrors the arrangement of Wiggins's CSF.)

Using his formalisation, (Wiggins 2006a) provides a number of definitions of specific behaviours that a creative system could display, in terms of what concepts are valued, which can be reached, etc. These terms can all be defined within the formalisation given here, as follows, assuming an exploratory creative system $S = (\mathcal{P}, \mathcal{N}, \mathcal{V}, \mathcal{Q})$, and using some of the terminology already defined above:

Hopeless uninspiration: The valued set of concepts is empty. That is, there are no concepts anywhere within the universal set that meet the 'quality' threshold.

Conceptual uninspiration: $\mathcal{V}_{0.5}(\mathcal{N}_{0.5}(\mathcal{P})) = \emptyset$. That is, there are no concepts within the acceptable ('normal') set of concepts that meet the 'quality' threshold.

Generative uninspiration: The valued set of reachable concepts is empty. That is, there are no concepts which the search mechanism can reach which meet the quality threshold.

Aberration: Where $\mathcal{B}$ consists of exactly those elements in the reachable set which are not in $\mathcal{N}_{0.5}(\mathcal{P})$, *aberration* occurs if $\mathcal{B} \neq \emptyset$. That is, aberration is when the search mechanism goes outside the 'normal' set of concepts. *Perfect aberration* is where $\mathcal{V}_{0.5}(\mathcal{B}) = \mathcal{B}$ (i.e. all the non-normal concepts meet the quality threshold); *productive aberration* is where $\mathcal{V}_{0.5}(\mathcal{B}) \neq \emptyset$ and $\mathcal{V}_{0.5}(\mathcal{B}) \neq \mathcal{B}$ (i.e. just some of the non-normal concepts meet the quality threshold); *pointless aberration* is where $\mathcal{V}_{0.5}(\mathcal{B}) = \emptyset$ (i.e. no non-normal concepts meet the quality threshold).

Evaluating search methods

Ventura's analysis

Ventura(2011) gives an analysis of the limitations of uninformed search strategies in a creative context. His definitions and results are general enough that they should be applicable to the framework here, although there is one small formal point that needs to be stipulated first. Ventura (implicitly) makes the following assumption:

> **One concept per step:** Each formal 'step' in the search process corresponds to the generation of exactly one concept/artefact.

That is, Ventura's analysis does not allow for intermediate computational steps behind the scenes which do not directly correspond to the generation of an artefact. The Wiggins definitions (and our reformulations) do not demand this restriction, but it is a plausible constraint, and could be formalised thus:

> Given an exploratory creative system $(\mathcal{P}, \mathcal{N}, \mathcal{V}, \mathcal{Q})$, $\mathcal{Q}$ is a *one-concept-per-step* scheme if, whenever $z' = \mathcal{Q}(\mathcal{N}, \mathcal{V})(z)$, $\exists c' \in \mathcal{P}$ such that:
>
> (i) c' is an element of z';
>
> (ii) c' is not an element of z;
>
> (iii) for every element c of z' where $c' \neq c$, c is an element of z.

This perspective could be taken even further, by revising our definition of an exploratory creative system to include a set $\mathcal{OP}$ of *operators*, which are mappings from tuples of concepts to concepts; that is, each operator is a member of $\mathcal{P}^{\mathcal{P}^k}$ for some integer k. Then we would stipulate that each step in a search meets the constraint that it corresponds to the invocation of one operator:

> If two concept-sequences $\langle c_1, \ldots, c_n\rangle, \langle d_1, \ldots, d_m\rangle$ are such that $\mathcal{Q}(\mathcal{N}, \mathcal{V})(\langle c_1, \ldots, c_n\rangle) = \langle d_1, \ldots, d_m\rangle$, then:
>
> there must be an operator $p \in \mathcal{OP}$, and concepts $\langle e_1, \ldots, e_k\rangle$ (where $k \leq n$) such that for every $1 \leq i \leq k, e_i = c_j$ for some j, and $p(e_1, \ldots, e_k) = d_r$ for some $d_r \in \{d_1, \ldots, d_m\}$, and $d_r \notin \{c_1, \ldots, c_n\}$, and $\forall d_i \in \{d_1, \ldots, d_m\}$, either $i = r$ or $d_i \in \{c_1, \ldots, c_n\}$.

Ventura's analysis provides one possible formalisation of the intuitive notion of a search strategy being 'better' (or 'best'). It posits a set of target elements (concepts, in our terminology) and considers the probability of the search reaching one of these elements. In a footnote, Ventura also offers a definition where the desirability of elements is a function to $[0, 1]$ (cf. our $\mathcal{V}$), and computes the probability of reaching an element with a maximal value.

It is arguable that in the area of creative systems, there is less emphasis on finding specific target items (or even finding a maximal-scoring item) and more on generally reaching acceptable or highly valued concepts (a direction in which Ventura's footnote moves). Our formalisation allows an alternative perspective on the assessment or comparison of search methods; see the next subsection.

Comparing search

Our formulation of the CSF allows the comparison of two search methods according to how the concepts they reach are rated by the related mappings $\mathcal{N}$ and $\mathcal{V}$, taking into account the depth of search involved. As we will want to apply certain definitions to various mappings (including $\mathcal{N}$ and $\mathcal{V}$), we will start with a general schematic definition in which the mapping $\mathcal{F}$ can be anything of the appropriate type (so $\mathcal{F}$ is not mnemonic for anything, being just a placeholder for now). Also, AGG will stand for either AVG (arithmetic mean) or MAX (maximum) of a function applied to a set.

Definition 11: Suppose there are two exploratory creative systems $(\mathcal{P}, \mathcal{N}, \mathcal{V}, \mathcal{Q}_1)$ and $(\mathcal{P}, \mathcal{N}, \mathcal{V}, \mathcal{Q}_2)$, with $S_i(k, k') =$ the set of concepts reachable in no fewer than k and no more than k' steps in these two systems ($i = 1, 2$). Also, $\mathcal{F} \in [0, 1]^{\mathcal{P}}$ (i.e. a rating of concepts, of some sort). Then

(i) $\mathcal{Q}_1$ *has higher AGG $\mathcal{F}$-values up to k' steps* than $\mathcal{Q}_2$ if $AGG(\mathcal{F}, S_1(0, k')) > AGG(\mathcal{F}, S_2(0, k'))$.

(ii) $\mathcal{Q}_1$ *has higher AGG $\mathcal{F}$-values beyond k steps* than $\mathcal{Q}_2$ if $AGG(\mathcal{F}, S_1(k, k')) > AGG(\mathcal{F}, S_2(k, k'))$ for any $k' > k$.

This compares two variants of a system in which only the search method $\mathcal{Q}$ is different. The above definitions will be applied, below, to specific values for $\mathcal{F}$.

Definition 12: Given a two exploratory creative systems $(\mathcal{P}, \mathcal{N}, \mathcal{V}, \mathcal{Q}_1)$ and $(\mathcal{P}, \mathcal{N}, \mathcal{V}, \mathcal{Q}_2)$:

(i) $\mathcal{Q}_1$ is *higher valued on average up to k steps* than $\mathcal{Q}_2$ if $\mathcal{Q}_1$ has higher AVG $\mathcal{V}$-values up to k steps than $\mathcal{Q}_2$.

(ii) $\mathcal{Q}_1$ is *more normal on average up to k steps* than $\mathcal{Q}_2$ if $\mathcal{Q}_1$ has higher AVG $\mathcal{N}$-values up to k steps than $\mathcal{Q}_2$.

(iii) $\mathcal{Q}_1$ *achieves higher value up to k steps* than $\mathcal{Q}_2$ if $\mathcal{Q}_1$ has higher MAX $\mathcal{V}$-values up to k steps than $\mathcal{Q}_2$.

(iv) $\mathcal{Q}_1$ *achieves greater conformity up to k steps* than $\mathcal{Q}_2$ if $\mathcal{Q}_1$ has higher MAX $\mathcal{N}$-values up to k steps than $\mathcal{Q}_2$.

For each of the 4 subparts of the above definition, we can frame a corresponding definition which says that there exists some depth k' after which one of the search methods $\mathcal{Q}_i$ gives a higher value than the other; e.g.:

$\mathcal{Q}_1$ is *higher valued eventually* than $\mathcal{Q}_2$ if there is some integer $k' > 0$ such that $\mathcal{Q}_1$ has higher AVG $\mathcal{V}$-values beyond k' steps than $\mathcal{Q}_2$.

Similar substitutions can be made in the other definitions.

In this way, we have several ways of describing one search method $\mathcal{Q}_1$ as being 'better' than another, $\mathcal{Q}_2$. Next, we consider how comparisons of search methods can be more detailed.

Descriptive criteria

Ritchie(2001; 2007) defines a set of formal criteria which can be used to describe aspects of a potentially creative system's behaviour. Central to these formal criteria are two *rating schemes* for assigning values in $[0, 1]$ to elements of the set of *basic items* (i.e. the set of possible artefacts). One rating scheme (*typ*) represents *typicality*, indicating the extent to which an item lies within the norm for this type of artefact. The other rating (*val*) is for *value*, and indicates the 'quality' of an item. Ritchie's *typ* and *val*, Wiggins's $\mathcal{R}$ and $\mathcal{E}$ and our $\mathcal{N}$ and $\mathcal{V}$ all appear to capture the same intuitive notions: that we can rate possible creations as to their membership of a concept set, and in terms of the quality of such creations.

There are some differences of nuance between Ritchie's constructs and those in the CSF, to which we will return later, but for the moment let us consider how the central ideas in some of Ritchie's criteria could be used within the CSF as stated here.

The first eight of Ritchie's criteria are stated in terms of a *result set*, R, which is the set of artefacts produced by the computer program, and the two ratings *typ* and *val*. There is not space here to reproduce them all, but Criterion 7 illustrates the general idea:

$ratio(V_{\gamma,1}(R) \cap T_{0,\beta}(R), T_{0,\beta}(R)) > \theta$,
for suitable β, γ, θ.

where $V_{\gamma,1}(R)$ is the set of elements of set R which are rated above γ by *val*, $T_{0,\beta}(R)$ is the set of elements of R which are rated below β by *typ*, and $ratio$ computes the ratio of the sizes of two sets. That is, this computes the proportion of the untypical items which are of good quality.

Criteria like this could be applied to a creative system $(\mathcal{P}, \mathcal{N}, \mathcal{V}, \mathcal{Q})$, using $\mathcal{N}$ to define $T_{i,j}$ and $\mathcal{V}$ to define $V_{i,j}$. There are various ways in which the result set R could be defined in terms of reachable concepts: all reachable concepts? concepts reachable from a starting set B? concepts reachable after some number of steps k? All of these are plausible models of a 'result set'. Hence there would be a few families of very similar formula, parameterised according to starting set or number of steps.

Ritchie(2007) emphasises that these criteria are not all measures of creative success, but can be used to 'profile' a (potentially) creative program by describing its behaviour in more detail. In the same way, they could give a more detailed picture of a creative system, in the CSF sense.

Ritchie also postulates an *inspiring set*, I, which are the existing artefacts upon which the design of the creative program was based. The remaining criteria (9 - 18 in Ritchie(2007)) make comparisons of different sorts between I and the result set R. There is no exact counterpart within the CSF, as there is no 'design' stage in the formalisation. However, the formal structure of Ritchie's criteria which involve I could be coerced into service within the CSF, by replacing I with some initial set of concepts B, from which search starts. That is, where Ritchie has a criterion such as:

$ratio(V_{\gamma,1}(R - I) \cap T_{\alpha,1}(R - I)), (R - I)) > \theta$, for suitable α, γ, θ

(informally, 'a high proportion of the novel results are both highly valued and very typical of the genre') we would have:

$ratio(V_{\gamma,1}(R - B) \cap T_{\alpha,1}(R - B)), (R - B)) > \theta$, for suitable α, γ, θ.

where R is the set of concepts reachable from B. (Again, there is a possible variant where a number of steps k is stipulated.)

Although this seems to indicate that we can simply port the Ritchie criteria into the CSF, there is one further issue to consider: there is a difference in the overall perspective of the two formal accounts. There are various viewpoints one could assume in devising a formal model in this area. For example, it would be possible to have an abstract declarative formalisation of the nature of the creative task, without including any details of how this task might be executed. Or a model might be proposed as describing (at some suitably abstract level) *how* a creative system operates. Ritchie's criteria arguably take a third viewpoint, in which one treats the program as a conveyor of input/output data, and attempts, from an external viewpoint, to say more precisely how it has performed. The *typ* and *val* mappings are certainly not proposed as components of the program or system. Instead, these are measures which might be applied by, for example, having humans judge the program's output.

In Wiggins's CSF, the formal definitions of $\mathcal{R}$ and $\mathcal{E}$ (the symbolic counterparts of our $\mathcal{N}$, $\mathcal{V}$) would be compatible either with a characterisation of the abstract nature of creativity, or with a model of a creative system. However, the terminology used, and the inclusion of the $\mathcal{T}$ ('agenda') mapping (our $\mathcal{Q}$) determine that this is a model (at a very abstract level) of the working of a creative system. Hence, the whole intent of the CSF is radically different from that of Ritchie's definitions, even though there is a clear parallel between the intuitive meanings of $\mathcal{N}$ and *typ*, $\mathcal{V}$ and *val*.

This means that what we have sketched above is *not* the direction application of Ritchie's criteria within the framework, but the definition of *counterparts* within the CSF, where the $T_{i,j}$ and $V_{i,j}$ mappings are defined using internal components ($\mathcal{N}$, $\mathcal{V}$) of the system, not external judgements. However, this means that if we are scrutinising the behaviour of a creative system, we can apply the conditions outlined above (i.e. the counterparts of Ritchie's criteria) in distinct ways:

Internal: How is the system performing, in its own terms? For this, we use $\mathcal{N}$ and $\mathcal{V}$ to define $T_{i,j}$ and $V_{i,j}$.

External: How is the system performing, in terms of independent measures such as human judgements? For this, we use the independent measures to define $T_{i,j}$ and $V_{i,j}$.

In both of these, we retain the notions of B (initial set) and R (reachable concepts) discussed above. Given that the set of reachable concepts is in effect the 'result set', if the inspiring set I is known, then using I instead of B, with an External perspective, is effectively the scenario in the Ritchie papers.

These various adaptations of the criteria to the CSF can be viewed alongside the definitions in our section **Comparing search** above, and provide a slightly finer grained and more detailed vocabulary for comparing search strategies.

Guiding the metalevel

We have already shown how the metalevel of a creative system can start from an existing object level system and search for a variant system, using a metalevel value function $\mathcal{V}^{meta}$. What should the content of $\mathcal{V}^{meta}$ be? Since the metalevel has access to the object level mappings $\mathcal{N}$ and $\mathcal{V}$, it would be possible for $\mathcal{V}^{meta}$ to be defined using composite criteria such as those we have outlined above, the counterparts of Ritchie's criteria. That is, the metalevel search could be guided by how candidate object-level systems performed according to these criteria. For this, the distinction between internally-parameterised and externally-parameterised versions of the criteria is important. Whereas the externally-parameterised version (using human judgements or other measures) is exactly appropriate for profiling or assessing the success of the object-level system, the internally-parameterised version (using $\mathcal{N}$ and $\mathcal{V}$) are the only ones that make sense within the creative (metalevel) system itself.

This glosses over the significant question of whether a real creative program would be implemented with the meta/object strata of the CSF, or whether the formal framework is only a way of describing, at some fairly abstract level, what creative systems do (or could do). It is possible that the actual use of structured criteria which compare ratings of initial sets and of reachable concepts would not be realistically applicable in implemented systems.

Summing up

We have presented a reformalisation of Wiggins's CSF, which:

- makes the use of a symbolic language an optional extra
- indicates how search strategies can be formally compared within the CSF
- clarifies the metalevel, defining some metalevel constructs in more detail and making explicit some formal comparisons.
- shows how Ritchie's criteria can be adapted, in a number of ways, to the CSF formalisation, thereby clarifying the relationship between these frameworks

In this way, we have extended the development of formal descriptive frameworks for creative systems.

Acknowledgments

The author is very grateful to Geraint Wiggins for detailed discussions of this material, and for comments on an draft of this paper.

References

Boden, M. A. 1998. Creativity and Artificial Intelligence. *Artificial Intelligence* 103:347–356.

Boden, M. A. 2003. *The Creative Mind*. London: Routledge, 2nd edition. First edition 1990.

Colton, S.; Charnley, J.; and Pease, A. 2011. Computational creativity theory: The FACE and IDEA descriptive models. In Ventura et al. (2011), 90–95.

Colton, S.; Pease, A.; and Ritchie, G. 2001. The effect of input knowledge on creativity. In Weber, R., and von Wangenheim, C. G., eds., *Case-Based Reasoning: Papers from the Workshop Programme at ICCBR 01*.

García, R.; Gervás, P.; Hervás, R.; and Pérez y Pérez, R. 2006. A framework for the E-R computational creativity model. In *5th Mexican International Conference on Artificial Intelligence (MICAI-06)*, volume 4293 of *LNAI*, 70–80. Springer Verlag.

Manurung, R.; Ritchie, G.; and Thompson, H. 2012. Using genetic algorithms to create meaningful poetic text. *Journal of Experimental & Theoretical Artificial Intelligence* 24(1):43–64.

Nilsson, N. J. 1971. *Problem-solving methods in artificial intelligence*. New York: McGraw-Hill.

Pease, A.; Winterstein, D.; and Colton, S. 2001. Evaluating machine creativity. In Weber, R., and von Wangenheim, C. G., eds., *Case-Based Reasoning: Papers from the Workshop Programme at ICCBR 01*, 129–137.

Ritchie, G. 2001. Assessing creativity. In *Proceedings of the AISB Symposium on Artificial Intelligence and Creativity in Arts and Science*, 3–11.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17(1):67–99.

Thornton, C. J. 2007. How thinking inside the box can become thinking outside the box. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*.

van Deemter, K. 2010. *Not Exactly: in praise of vagueness*. Oxford, UK: Oxford University Press.

Ventura, D.; Gervás, P.; D. Fox Harrell; Maher, M. L.; Pease, A.; and Wiggins, G., eds. 2011. *Proceedings of the 2nd International Conference on Computational Creativity*.

Ventura, D. 2011. No free lunch in the search for creativity. In Ventura et al. (2011), 108–110.

Wiggins, G. 2006a. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19:449–458.

Wiggins, G. A. 2006b. Searching for computational creativity. *New Generation Computing* 24(3):209–222.

Creative Search Trajectories and their Implications

Kyle E. Jennings
Department of Psychology
University of California, Davis
Davis, CA 95616 USA
kejennings@ucdavis.edu

Abstract

Creative search trajectories are chronologically organized intermediate products (such as sketches and drafts) from the creative process. We discuss what sorts of conclusions can be made when these trajectories show non-monotonic progress toward the final creation. We introduce several key distinctions that are often overlooked, and argue that two null hypothesis processes must be rejected before non-monotonicity can be claimed to support more complex processes. We show that these null hypotheses are in fact difficult to rule out definitively using the sorts of evidence that past research has offered.

The sketches, drafts, revisions, and rejected ideas that creators leave in their wake on the way toward great masterpieces offer glimpses of the mental processes that are responsible for their achievements. Ordered in time, these artifacts trace a trajectory through a mental space, and may be the signatures of the specific exploration strategies that differentiate great thinking from the mundane. Even the differences in creativity that can be observed among study participants may be explainable by tracing and analyzing the detailed steps taken with simple creative problems.

This approach—which we call trajectory analysis—borrows from the problem solving research of Newell and Simon (1972) and others. However, whereas most problem solving research uses problems that are concrete, objective, and formalizable in terms of specific states, operators, and goals, the creative problems faced by artists, writers, and scientists are less easily reduced to symbols, rules, and computational steps. Though there are examples where researchers have meticulously examined creative search trajectories using objective features (Weisberg 2004), most research has let subjective judgments stand in for complete formal details (Kozbelt 2006; Simonton 2007; Kozbelt and Serafin 2009; Damian and Simonton 2011). Despite taking different approaches, many of these results point to the conclusion that creative outcomes are not arrived at directly, but rather through the twists and turns of false starts and retraced steps. These observations have been taken as evidence that the creative process is tentative and experimental rather than deliberate and informed.

Given how complex it would be to devise comprehensive and objective descriptors of intermediate states in creative problems, it may not be practical to overcome all of the imperfections of subjective judgments. However, this does not mean that conclusions drawn about the creative process on the basis of these judgments do not need to be rigorously justified. While it seems intuitively clear that a process that is tentative, uncertain, or experimental would produce trajectories that do not move monotonically closer to the final solution (Simonton 2007; Damian and Simonton 2011) or that do not show monotonic improvement over time (Kozbelt 2006; Kozbelt and Serafin 2009), caution must be exercised before concluding that trajectories with these features could *only* have been produced by such processes.

This paper aims to clarify what can and cannot be concluded about the creative process on the basis of creative search trajectories. Though we are ultimately optimistic about the potential of this approach and advocate the view that creativity requires uncertainty and experimentation, we will show that existing evidence that creative search trajectories are non-monotonic is in fact compatible with very straightforward search processes, and that more care and precision must be used when analyzing search trajectories. We begin with a more detailed discussion of existing trajectory analysis approaches, and then describe in overview the distinctions that past work has overlooked. We then illustrate these distinctions by presenting simple but formally complete examples that demonstrate the need for caution when drawing the conclusion that non-monotonic search trajectories necessarily reflect something other than a straightforward process. We conclude by discussing the burden of proof placed on researchers who wish to infer the causes of non-monotonicity and by offering suggestions for future work.

Background

While there are many approaches that people have taken to using intermediate products to understand the creative process (Getzels and Csikszentmihalyi 1976; Finke, Ward, and Smith 1992; Hennessey 1994; Ruscio, Whitney, and Amabile 1998; Rostan 2010), this paper focuses specifically on techniques that characterize the nature of the changes between successive revisions of the work (Kozbelt 2006; Simonton 2007; Kozbelt and Serafin 2009; Damian and Simonton 2011). Common to these analyses is the prediction that creativity should be associated with non-monotonicity,

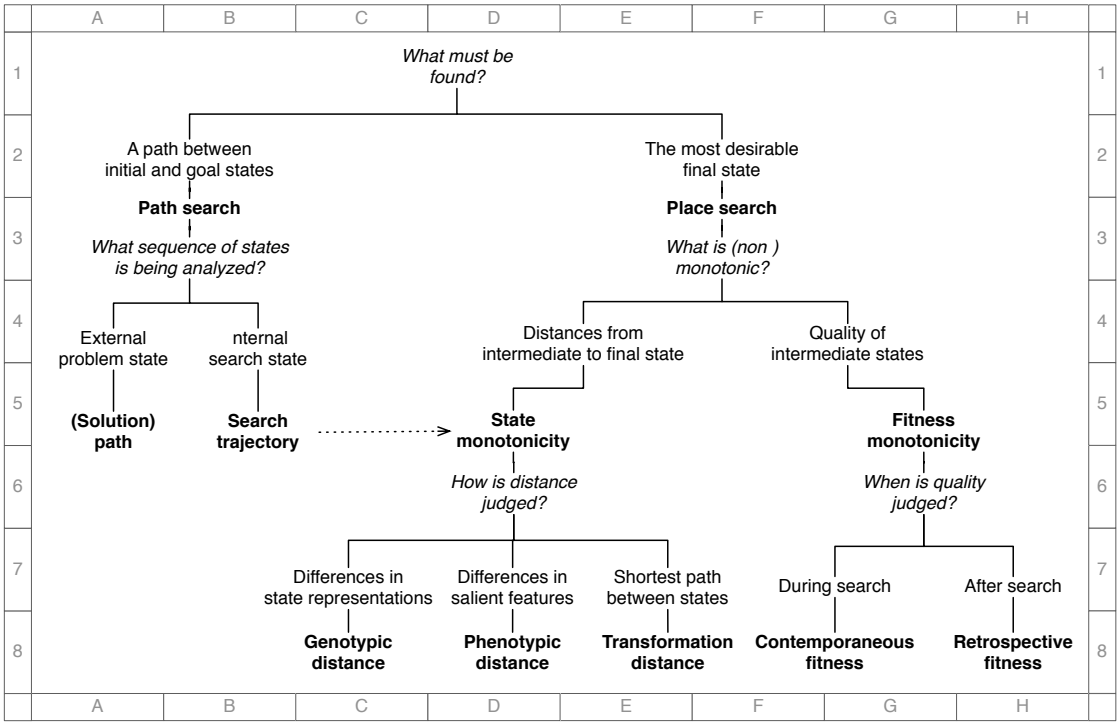


Figure 1: Depiction of the key questions, answers (italicized), and terms (bolded) that impact how search trajectories are analyzed.

which is perhaps best understood as the opposite of direct, incremental progress toward the final creation (other theories predict this, too, e.g., Perkins 2000). Beyond this commonality, the theories differ in their reasons that non-monotonicity should be expected, and indeed in how they operationalize monotonicity.

One approach to characterizing monotonicity is that taken by Kozbelt (2006) in his analysis of the sketches preceding Matisse's *Large Reclining Nude*. His approach focuses on the evaluation of each sketch, and in particular on whether the sketches become monotonically better with time. Given that Matisse's own evaluations are not available, Kozbelt has both artists and non-artists rate each sketch (presented in a random order) on 26 items, which are then analyzed in order to extract a latent quality dimension. The artists' ratings are found to be maximal at the final image, but beyond this they show non-monotonic variations over time (the contour of which is replicated in the non-artist sample). Similarly, Kozbelt and Serafin (2009) analyze intermediate sketches from drawings by non-eminent artists, and find that artwork that had been previously rated as more creative resulted from less monotonic trajectories. This is suggested to be due to an "interactive, hypothesis-testing dynamic" (Kozbelt and Serafin 2009, p. 358), though the specifics of this process are not articulated.

The other approach taken to characterizing monotonicity involves looking at the structure of the sketches themselves. The preponderance of this evidence stems from sketches Picasso left of *Guernica*. After an informal suggestion by Simonton (1999) that these sketches showed "false starts and wild experiments" (p. 197), Weisberg (2004) undertook a detailed analysis of the features shared between sketches and concluded that they were elaborations on a basic idea that itself had precedent in other work, including Picasso's own *Minotauromachy*. In response, Simonton (2007) undertook an alternative analysis in which various raters arranged the sketches in the order they would most logically have been generated, which reliably resulted in an order that did not match the actual temporal order. Later, Damian and Simonton (2011) had raters judge the similarity of the components from *Minotauromachy* to their counterparts in the *Guernica* sketches, and again found that the *Guernica* work did not get monotonically closer (or further) from the prototypes. Simonton claims this as evidence in support of the Blind Variation and Selective Retention (BVSR) theory of creativity (Simonton 2003; 2010), which holds that both desirable and undesirable variations will be generated during the creative process.

The Simonton work, in particular, has generated a great deal of recent controversy (Dasgupta 2011; Gabora 2011), much of it focused on the computational and algorithmic specifics of BVSR theory. In fact, neither Kozbelt's nor Simonton's analyses offered precise and detailed accounts of the processes that would lead to non-monotonicity, with Gabora (2011) suggesting that BVSR wouldn't even predict non-monotonicity. While Simonton has begun better formalizing BVSR theory (Simonton 2011; 2012), the fact is that there are basic problems for the trajectory analy-

sis approach that any theory of the creative process must overcome. Therefore, rather than specifically addressing the claims by Kozbelt and Simonton, this paper will instead try to address some basic inconsistencies regarding how monotonicity is conceptualized and operationalized, and will demonstrate that non-monotonicity can result from processes that are more straightforward than most theories suggest.

Overview of Distinctions

In this paper we introduce several distinctions that are essential when analyzing search trajectories. As depicted in Figure 1, these distinctions revolve around questions about what the search must find, what aspect of trajectory monotonicity is of interest, and how monotonicity is measured. Throughout the paper we will define and illustrate the terms that these questions delineate, using the coordinates in the margins to reference to the relevant portions of the figure.

The first and most important question is what the search must find (Figure 1, D1). Problem solving research describes problems by the set of states and the operators that move between them. The problem solver's goal is to find a sequence of operator applications that transforms the initial state into (one of) the goal state(s). Because the goal state is already known, the thing being searched for is the path itself, which is why we refer to these as *path searches* (Figure 1, AB23). (See also Jennings, Simonton, and Palmer 2011.)

Because the goal states are well known in path search, solution quality depends more on the quality of the path than on the specific goal state reached, with shorter (or less costly) paths being better. Though this situation aptly describes many problems (e.g., proving a theorem, inventing a process for synthesizing a given protein) there are other problems where the end points are not known in advance. For instance, in painting the artist seeks to depict a certain scene, theme, or emotion using brush and paint. In most cases we compare paintings not by the set of brushstrokes that led from an empty canvas to the completed image, but rather by that image itself. Thinking of these final images as places in a solution space, we refer to this as a *place* search (Figure 1, EF23).

Creativity is possible with both path and place search, and most real problems involve some element of each (e.g., choosing a place and then finding the path to it). Though we'll speak of these as distinct kinds of searches in this paper, we recognize that understanding joint path-place search is an essential task for future work.

Monotonicity in Path Searches

Our discussion begins with path search. For simplicity we'll consider the classical Towers of Hanoi problem. (Though this problem leaves little room for creativity, it nicely illustrate our key points.) As depicted in Figure 2, there are three disks of decreasing size that are initially stacked on the leftmost of three pegs. The problem is to move the disks to the rightmost peg by moving one disk at a time without placing larger disks on top of smaller disks. The figure shows the shortest sequence of states that solves the problem, which together constitute the path found in this path search.

The Towers of Hanoi is often used to illustrate the failure of an heuristic called difference reduction, which entails iteratively applying the operator that most reduces the discrepancy between the current state and the goal state (Anderson 1993). In fact, solving this problem requires selecting operators that temporarily make the current state less similar to the goal state, and for this reason it could be argued that even a simple problem like the Towers of Hanoi has a non-monotonic solution. By this logic, there is no controversy behind claiming that *creative* problems exhibit non-monotonicity. However, we will see that the Towers of Hanoi isn't inherently non-monotonic, at least in sense that matters when making inferences about the search process.

Let's begin by looking at the monotonicity of the sequence of states forming the shortest path in Figure 2. Though we'll ultimately conclude that these are not necessarily the states that we should be analyzing when making inferences about the search process, they conveniently illustrate the different ways that monotonicity can be judged. The difference reduction heuristic works by economically comparing the current and goal states. For example, we could compare states by looking directly at their representations. Each state in Figure 2 can be described as an ordered triple indicating which pegs the smallest, middle, and largest disks are on. (This is sufficiently descriptive since larger disks cannot be on top of smaller disks.) Thus, the starting state is $(1, 1, 1)$, the final state is $(3, 3, 3)$, and the intermediate states are $(3, 1, 1)$, $(3, 2, 1)$, etc. By analogy to genetics, we can think of the state encoding as being a genotype. We'll define *genotypic distance* to be the dissimilarity between the encodings of two states, which here we can define as the sum of absolute differences in state representations (Figure 1, C78). For instance, the third state, $(3, 2, 1)$ differs from the final state, $(3, 3, 3)$, by $|3-3|+|3-2|+|3-1| = 3$.

We could also compare states by looking at their salient features, which may or may not be directly related to the state encoding. Again by analogy to genetics, we'll call this *phenotypic distance* (Figure 1, D78). For the Towers problem, let's calculate phenotypic distance by counting the number of disks that are on different pegs. Thus, the first and final states differ by three, the second and final by two, and so on.

As Figure 2 shows, the best solution to the Towers problem is indeed non-monotonic in genotypic distance (D_G) and phenotypic distance (D_P), with the solution becoming less similar to the goal at various points, and difference reduction would indeed fail with either of these methods for choosing operators. However, the fact remains that the sequence of moves shown is minimal (the path is the shortest path possible). Defining the *transformation distance* (D_T) as the length of the shortest path between two states (Figure 1, E78), then we can see in Figure 2 that the path does get monotonically closer to the final state. On the one hand this is a useless insight since performing difference reduction with D_T just pushes the work of finding the solution into calculating D_T. On the other hand this reveals that the solution really isn't non-monotonic, in the sense that there

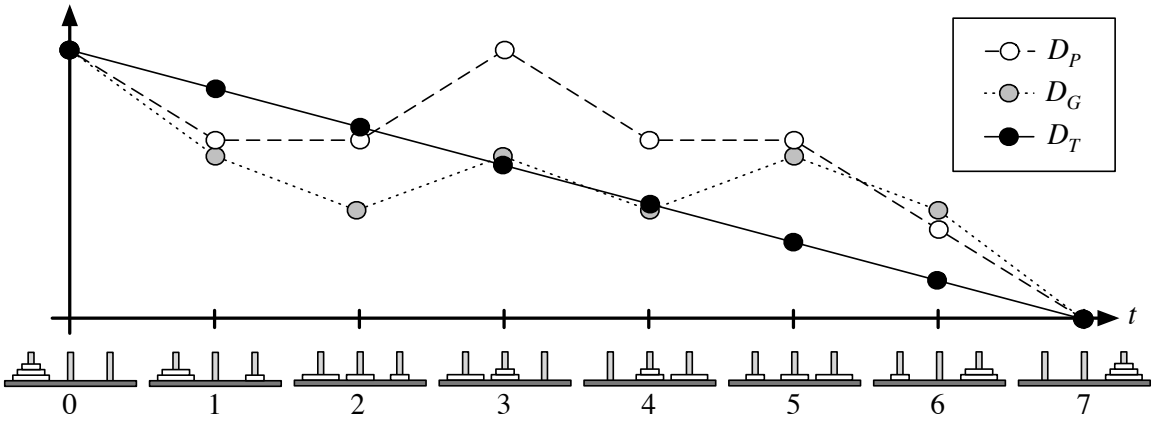


Figure 2: Illustration of non-monotonicity in genotypic distance, D_G, and phenotypic distance, D_P, but monotonicity in transformation distance, D_T, with the Towers of Hanoi problem. Distances are between the current state and the goal state, and have been normalized. The sequences of moves shown is optimal.

is no point when the path takes steps leading away from the goal state.

Having defined the various distance metrics that can apply to states, we need to ask whether we're in fact looking at the states that are relevant when making process inferences. Recall that in path search the solution is itself a sequence of states connected by operators. Thus, the states at the bottom of Figure 2 jointly constitute the solution path (Figure 1, A45). If the problem solver only represents the current state and the goal state, essentially treating each step in the solution as a new problem, then each of the individual states in Figure 2 may indeed describe the problem solver's internal state in each step, and we would conclude that the path is in fact monotonic since no state is ever revisited.

Suppose instead that the problem solver uses a technique like means-ends analysis (Newell and Simon 1961). In this case the internal search state would need to represent subgoals and paths with gaps. For example, the first subgoal in means-ends analysis would be to move the largest disk to the rightmost peg. This might be represented as:

$$(1,1,1) \rightarrow ??? \rightarrow (?,?,3) \rightarrow ??? \rightarrow (3,3,3)$$

The path would then be built recursively by filling in the steps before $(?,?,3)$ and then the steps after $(?,?,3)$. Whether this process is monotonic depends on, for example, whether the problem solver ever fails to achieve one subgoal and tries another one. Monotonicity would have to be evaluated according to the sequence of internal states (the search trajectory; Figure 1, B45), not the states of the problem itself (Figure 1, A45).

The importance of looking at internal states is clear when we realize that any search process that successfully finds a monotonic path may have explored longer paths that were ultimately edited before emitting the solution. For instance, a mathematician may prove several different lemmas as part of the proof of a larger theorem before realizing that they are all part of one general lemma and collapsing them accordingly. The final published proof would reflect this realization, but that does not imply that the mathematician's own thought processes followed the minimal path presented in the publication.

To summarize, when analyzing path searches, the relevant states to consider are the internal states, which will contain but may not be identical with the problem states. Non-monotonicity should be judged using transformation distance, as this is the most direct measure of whether the trajectory includes apparently wasted effort.

Monotonicity in Place Searches

We're now ready to shift our emphasis to place searches, where the major aim of the search is to find the most desirable final state (Figure 1, F23). Here we'll adopt a landscape metaphor, wherein states, $\vec{x}$, are thought of as being topographically organized according to the available operators. Each state's desirability is denoted $f(\vec{x})$, which we'll call it's fitness in keeping with genetics-inspired language used for genotypic and phenotypic distance.

Our discussion in this section considers search processes that maintain a single current state that is iteratively improved. Though these processes may consider several alternate states in each iteration, only one survives into the next iteration. (As with path-place searches, we leave place searches where multiple states are under simultaneous refinement for future work.) We're not going to attempt to show that any particular process fitting these parameters is most plausible. Instead, we discuss two processes that are relatively implausible and yet not straightforward to rule out with trajectory analysis. The first is a computationally implausible process that always finds the most efficient path between the initial and final state, which we call the *direct* process.[1] The second, which we call the *hill climbing* process, takes the psychologically implausible steps of evaluat-

[1] Practically speaking, finding evidence for the direct process suggests that the entire search occurred mentally, making trajectory analysis the wrong analytical approach for that problem.

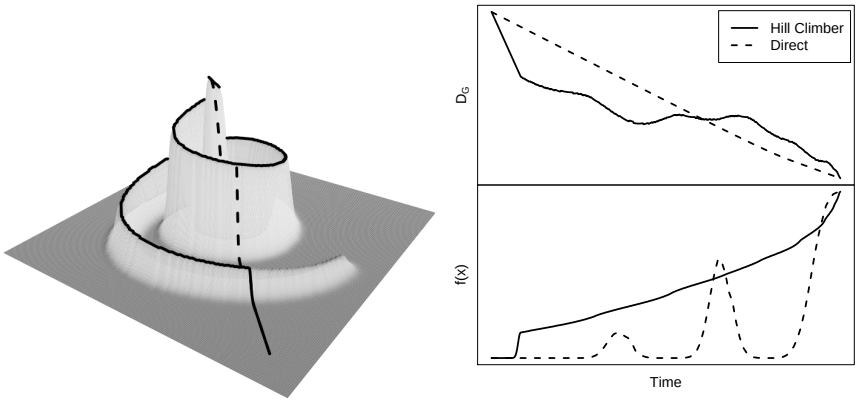


Figure 3: Illustration of (a) the hill-climbing process, which shows non-monotonicity in genotypic distance but monotonicity in fitness (solid lines), and (b) the direct process, which shows monotonicity in genotypic distance but non-monotonicity in fitness (dashed lines). Times and distances have been normalized.

ing *every* available move and *always* choosing the one that most improves fitness, stopping when no improvement is possible (regardless of how low the current state's fitness).

Whereas monotonicity in path search always referred to the states in the search trajectory, place search lets us look at the monotonicity of both the intermediate states and their fitnesses (Figure 1, F3). We'll call the monotonicity of the intermediate states *state monotonicity* (Figure 1, DE45). The same distinctions between genotypic, phenotypic, and transformation distance apply, and as before transformation distance is the most relevant form of monotonicity (though usually not the most convenient to calculate). We'll call the monotonicity of the fitness function over time *fitness monotonicity* (Figure 1, G45). As we'll see, different conclusions can result from considering fitness monotonicity as assessed during search (Figure 1, FG78) or after search (Figure 1, H78).

In the following we'll present visual examples of searches over two-dimensional grids, with the states in the x-z plane and fitness on the y-axis. In this way, the ideal endpoint for a place search is the highest point on the landscape. We'll allow single-unit moves in the up-down, left-right, and diagonal directions. Note that in this case D_G, defined as the Euclidean distance, is a good proxy for D_T.

Insufficiency of Either Monotonicity Alone

Now we can evaluate whether fitness or state trajectories are individually sufficient to rule out either the direct or hill climbing processes. For the landscape shown in Figure 3, a hill-climber that follows the fitness function, f, will trace a path like the solid line shown in the figure. This trajectory is non-monotonic in genotypic distance but monotonic in fitness. The direct process would form a trajectory that is monotonic in genotypic distance but non-monotonic in fitness. Therefore, finding one but not both of state or fitness non-monotonicity is not sufficient to rule out both null hypotheses processes.

Partially Observable States

The trajectory traced in Figure 3 is a spiral. While this is literally a roundabout path to take, it is at least free of cycles where the trajectory revisits previously-encountered states. A trajectory with cycles seems to clearly contradict both null hypotheses. Indeed, the presence of a cycle can rule out the direct process, as editing out the cycle will not prevent the path from reaching the same final state. A cycle is likewise impossible with hill climbing, *unless* the internal state is only partially observable (cf. Gabora 2011).[2]

Consider the landscape in Figure 4, where the complete state $\vec{x} = (x, \hat{x})$ has an observable part (x) and an unobservable part ($\hat{x}$). Projecting the full trajectory into x shows a cycle, but the cycle disappears with $(x, \hat{x})$.

Criteria Change

In a case almost identical to the previous example, suppose that people's criteria change during search. This is akin to having an unobserved portion of the state that affects the evaluation function. Here again, cycles could emerge in the observable state that are in fact monotonic if the evaluation function's hidden parameter were observable. (The cases are not quite identical since changes to this hidden parameter are dependent on some higher-level process, and hence aren't fully explainable by reference to f.) If one considers changes to evaluation criteria as separate from the main search process then it's still possible for this apparent non-monotonicity to occur with a hill climber.

A special problem can occur when criteria change occurs since the same state will be evaluated differently before and after the change. This could lead to retrospective evaluations of fitness over time showing non-monotonicity while in fact fitness was experienced as increasing monotonically during the search (see Figure 1, FH78). Suppose that the

[2]Cycles could also occur in the special case where all of the states in the cycle are indistinguishable in terms of fitness.

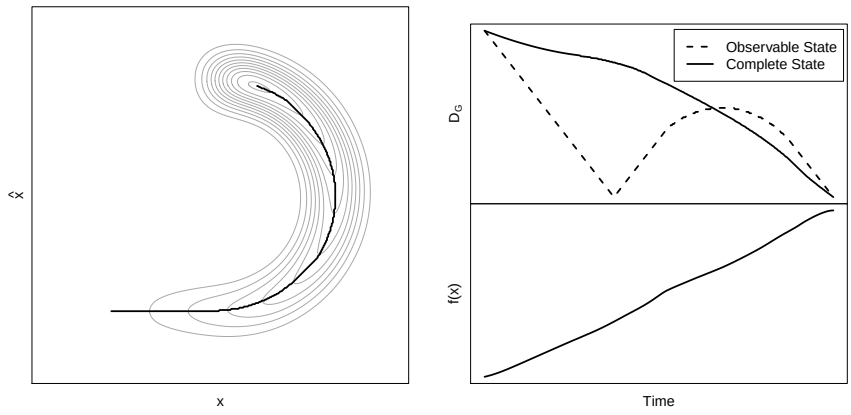


Figure 4: Illustration of an apparent cycle using a hill climbing process where the state is only partially observable. Projecting the trajectory into observable space (horizontal axis) shows a cycle, while the complete trajectory is non-cyclic.

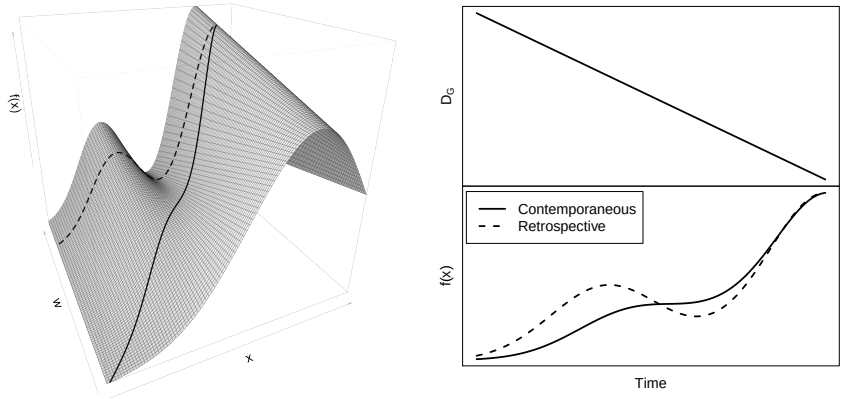


Figure 5: Illustration of a search over x while a criteria weight w is simultaneously changing. The path (solid line) starts in the foreground and proceeds to the background. As shown in the right graph, contemporaneous fitness is monotonic. However, retrospective evaluation of the traversed points would be non-monotonic, as shown by the dashed lines in both graphs.

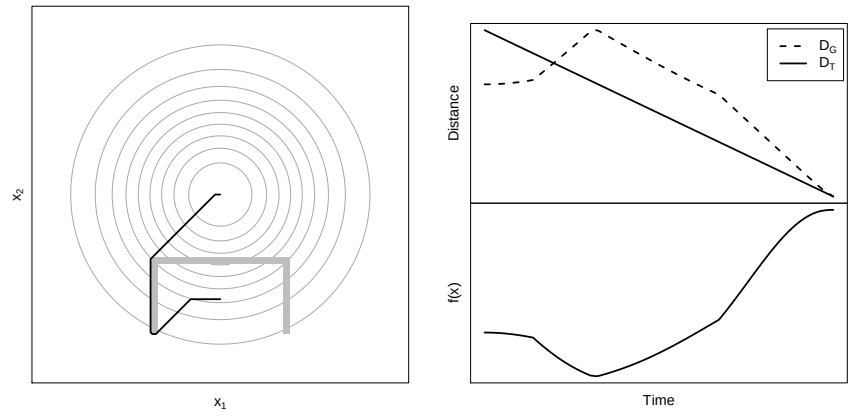


Figure 6: Illustration of a path that is non-monotonic with D_G but monotonic with D_T. The gray region in the left plot shows infeasible configurations that can be represented with the genotype but that cannot be realized.

state is a scalar x and that $f(x)$ depends on a weight parameter w such that $f(x) = w \cdot f_1(x) + (1 - w) \cdot f_2(x)$. Figure 5 demonstrates how retrospective fitness evaluation could show non-monotonicity even though contemporaneous fitness was monotonic.

Infeasible Regions

So far we have considered problems where D_G is a good proxy for D_T. However, there are problems where the state representation does a poor job reflecting the actual difficulty of transforming one state into another. Consider the landscape in Figure 6, where the gray regions are infeasible states. Though a hill climber would fail on this particular landscape, the direct process would find the path shown. As can be seen, the state trajectory appears non-monotonic with D_G but isn't when judged with D_T. The fitness trajectory is also non-monotonic, though that could occur with any direct process.

Phenotypic Distance

In real-world situations, both transformation distance and genotypic distance can be impractical to compute. In contrast, phenotypic distance, which is based on comparing the state's salient features, can be assessed fairly directly, such as by having several human raters make intuitive similarity judgments for pairs of intermediate products. However, there are two issues with this approach. From a psychological standpoint, it has long been known that human similarity judgments are not well-behaved metrics, meaning that they can be context-dependent, asymmetric, and intransitive (Tversky 1977). These properties could introduce non-monotonicity that was is not in the original stimulus (Gabora 2011).

Another problem with similarity judgments lies in the fact that genotype and phenotype may be non-monotonically related. Let the state be a scalar $x \in [0, 1]$ and suppose that this maps to a single salient property, $p(x) = \sin(2\pi x)$ (see top half of Figure 7). If a search proceeds linearly from $x = 0$ to $x = 1$, D_G (the absolute difference in state representations) will be monotonically decreasing but D_P (the absolute difference in the level of the property) will be non-monotonic (see bottom half of Figure 7). This is true regardless of what sort of landscape or search process is used.

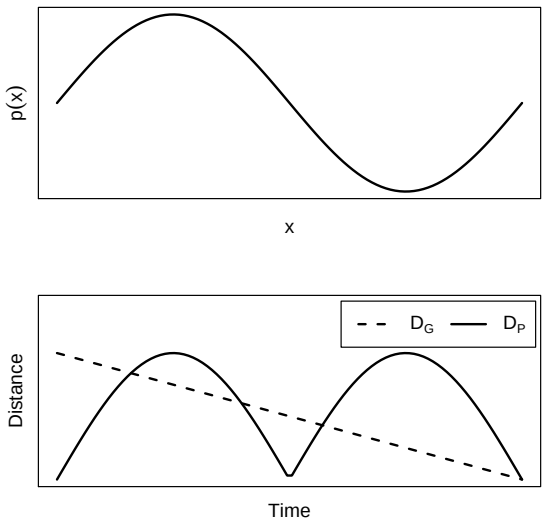


Figure 7: The top graph shows the relationship between the state x and its salient property, $p(x)$. The left graph shows monotonicity in D_G (and presumably D_T) as x goes from zero to one, but non-monotonicity in D_P.

Discussion

This paper has introduced a set of important distinctions that must be made when analyzing creative search trajectories, as summarized in Figure 1. We have argued that path searches and place searches must be understood differently. With path searches, we have claimed that the important measure of the efficiency of the search process is the transformation monotonicity of the problem solver's internal states, which may or may not be the same as the states of the problem itself. With place searches, we have argued that researchers claiming that people use complex search processes must first reject two null hypothesis processes, the direct and hill-climbing processes. We have shown how rejecting these processes entails demonstrating not just that there is non-monotonicity in both states and fitness, but also that:

- State non-monotonicity occurs with transformation distance, not just with the more convenient indices of genotypic and phenotypic distance
- The observed state non-monotonicity does not reflect movement on unobservable dimensions or the effects of criteria changes
- Fitness non-monotonicity assessed retrospectively would also have been non-monotonic when assessed contemporaneously

We have also illustrated how non-monotonicity may result from processes that are at either end of the spectrum from intelligent and informed (the direct process) and mechanical and uninformed (the hill climbing process). Regarding existing empirical work, we have shown that authors have differed on whether they've analyzed state (Simonton 2007; Damian and Simonton 2011) or fitness (Kozbelt 2006; Kozbelt and Serafin 2009) monotonicity. Though we don't claim that the counterexamples provided here disprove the claims made by these authors, we have raised important issues that future work must address.

We also acknowledge that our work has several significant limitations. First, for simplicity we have falsely dichotomized path and place search, even though we are quite sure that real creativity involves elements of each. Indeed, the joint operation of these searches may map nicely onto the problem finding/problem solving distinction that Getzels and Csikszentmihalyi (1976) introduced. Second, though we have alluded to criteria change as an important consideration when analyzing place searches, we have avoided discussing how and when this would occur and what could drive it. This decision was ultimately practical, since any

discussion of changes to criteria leads to the question of what guides criteria selection, and whether this itself can change—leading to an infinite regresss that may not be resolvable at the level of an individual creator (cf. Jennings 2010a). We have also treated criteria change as compatible with the hill climbing process, which we recognize may not be without controversy. Third, we recognize that we have not carefully considered the role of operators, and in particular how the discovery of new operators mid-search may affect conclusions about state monotonicity. Finally, we are fully aware that all of our counterexamples involve highly abstracted problems, and so ultimately they serve more to illustrate our points rather than provide an existence proof. We look forward to addressing these and other limitations in future work.

Our purpose here is not to cast aspersions on the trajectory analysis approach. Indeed, we are actively pursuing empirical techniques that rely upon trajectory analysis (Jennings 2010b; Jennings, Simonton, and Palmer 2011), although these techniques promise to reveal more about the underlying problem than can be obtained from a search trajectory alone. Beyond this, we continue to believe that trajectory monotonicity is an eminently practical way to study the creative process, particularly in cases where the available data are slim. The objections that we've raised in this paper do not seal the fate of this approach, but rather offer constructive critiques that should be addressed in future research.

References

Anderson, J. R. 1993. Problem solving and learning. *American Psychologist* 48(1):35–44.

Damian, R. I., and Simonton, D. K. 2011. From past to future art: The creative impact of Picasso's 1935 *Minotauromachy* on his 1937 *Guernica*. *Psychology of Aesthetics, Creativity, and the Arts*. Advance online publication. doi: 10.1037/a0023017.

Dasgupta, S. 2011. Contesting (Simonton's) blind variation, selective retention theory of creativity. *Creativity Research Journal* 23(2):166–182.

Finke, R. A.; Ward, T. B.; and Smith, S. M. 1992. *Creative Cognition: Theory, Research, and Applications*. Cambridge, MA: MIT Press.

Gabora, L. 2011. An analysis of the blind variation and selective retention (BVSR) theory of creativity. *Creativity Research Journal* 23(2):155–165.

Getzels, J. W., and Csikszentmihalyi, M. 1976. *The Creative Vision: A Longitudinal Study of Problem Finding in Art*. New York: John Wiley & Sons.

Hennessey, B. A. 1994. The consensual assessment technique: An examination of the relationship between ratings of product and process creativity. *Creativity Research Journal* 7(2):193–208.

Jennings, K. E.; Simonton, D. K.; and Palmer, S. E. 2011. Understanding exploratory creativity in a visual domain. In *Proceedings of the Eighth ACM Conference on Creativity and Cognition*.

Jennings, K. E. 2010a. Developing creativity: Artificial barriers in artificial intelligence. *Minds and Machines* 20:489–501.

Jennings, K. E. 2010b. Search strategies and the creative process. In *First International Conference on Computational Creativity*, 130–139.

Kozbelt, A., and Serafin, J. 2009. Dynamic evaluation of high- and low-creativity drawings by artist and nonartist raters. *Creativity Research Journal* 21(4):349–360.

Kozbelt, A. 2006. Dynamic evaluation of Matisse's 1935 *Large Reclining Nude*. *Empirical Studies of the Arts* 24(2).

Newell, A., and Simon, H. A. 1961. GPS, a program that simulates human thought. In Billing, H., ed., *Lernende automaten*. Munchen: R. Oldengourg. 109–124.

Newell, A., and Simon, H. A. 1972. *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Perkins, D. 2000. *The Eureka Effect: The Art and Logic of Breakthrough Thinking*. New York: W. W. Norton.

Rostan, S. M. 2010. Studio learning: Motivation, competence, and the development of young art students' talent and creativity. *Creativity Research Journal* 22(3):261–271.

Ruscio, J.; Whitney, D. M.; and Amabile, T. M. 1998. Looking inside the fishbowl of creativity: Verbal and behavioral predictors of creative performance. *Creativity Research Journal* 11(3):243–263.

Simonton, D. K. 1999. *Origins of genius*. New York: Oxford.

Simonton, D. K. 2003. Scientific creativity as constrained stochastic behavior: The integration of the product, person, and process perspectives. *Psychological Bulletin* 129(4):475–494.

Simonton, D. K. 2007. The creative imagination in Picasso's *Guernica*: Monotonic improvements or nonmonotonic variants? *Creativity Research Journal* 19:329–344.

Simonton, D. K. 2010. Creativity as blind-variation and selective-retention: Constrained combinatorial models of exceptional creativity. *Physics of Life Reviews* 7:156–179.

Simonton, D. K. 2011. Creativity and discovery as blind variation and selective retention: Multiple-variant definition and blind-sighted integration. *Psychology of Aesthetics, Creativity, and the Arts* 5(3):222–228.

Simonton, D. K. 2012. Creativity, problem solving, and solution set sightedness: Radically reformulating BVSR. *Journal of Creative Behavior* 46(1):48–65.

Tversky, A. 1977. Features of similarity. *Psychological Review* 84(4):327–352.

Weisberg, R. W. 2004. On structure in the creative process: A quantitative case-study of the creation of Picasso's *Guernica*. *Empirical Studies of the Arts* 22(1):23–54.

The Creative Computer as Romantic Hero? Computational Creativity Systems and Creative Personæ

Colin G. Johnson
School of Computing
University of Kent
Canterbury, UK
C.G.Johnson@kent.ac.uk

Abstract

A popular definition of computational creativity is that it consists in behaviour that would be regarded as creative if performed by humans. This raises the question of *which* humans, as there are many different styles of human creative behaviour. This paper unpacks a number of ways in which human artistic creativity can be characterised, compares them with the kinds of creative actions found in computational creativity, and explores some aspects of human creativity that are underrepresented in computational creativity systems.

Introduction

Computational creativity (CC) has been pursued for decades, with intense activity in recent years. One of the most common definitions of CC is "building software that exhibits behavior that would be deemed creative in humans" (Colton and others 2009). This paper explores this definition.

Creative behaviour is not monolithic; there are many different ways in which to validly be a creative person. It might be asked what *kind* of creative behaviour in humans we want to compare computer systems with. This paper is concerned with creative systems in artistic domains such as visual art, music and literature, not with so-called *everyday creativity*.

Creative Personæ

There are clear parallels between the idea of CC as exhibiting human-like behaviour and with Turing test-style definitions of intelligence. In both, the system is designated as acting in a creative/intelligent way if it can generate behaviour or product that would require, creative/intelligent action in humans to produce it, that a human observer would recognise.

Creative acts in artistic domains are not actions that most people will carry out regularly. So, care must be taken in selecting which humans are taken as exemplars. In traditional Turing style AI tests, the exemplar human is a member of the general population. In these tasks we will have to assume that the exemplar has some specific skills and knowledge.

One approach is to choose beginners in the domain as exemplars. This fits with the approach to CC system development that sees the development of systems that can do beginner tasks as the first step towards the development of more sophisticated systems. Another approach is to build systems to be compared with mature creative work, where the exemplars are mature artists. There is a danger with any of these exemplar-based systems (Pease and Colton 2011) that they encourage pastiche; but, if evaluators are primed sufficiently, perhaps this can be avoided.

McCormack (2005) notes that CC algorithms will be valued when they "produce art recognized by humans for its artistic contribution (as opposed to any purely technical fetish or fascination)". This seems, reasonable; yet, it might just be a temporary state. Once we have got beyond the point at which the products/activities of computational artistic creative systems are acknowledged as valid artworks, we might become interested in "biographical" aspects of them, and produce works that reflect on the origins of the work without this seeming like "technical fetish". This reflection on origins might become part of the depth of the works.

An important point is that not all creative people are creative in the same way. This paper will consider a number of dimensions of what will be termed *creative persona space*: an informally defined space representing broad attitudes/approaches. This paper considers three dimensions: the social vs. individualist dimension; the importance (or not) of ongoing tradition, development and "craft skills"; ideas of new and old media and the way in which technology is used in the artistic production itself.

Ritchie (2007) uses the *inspiring set* in evaluating CC: a set of human *works* as exemplars of what a successful CC system would generate. The intention here is similar, but with regard to the *creator* rather than the works: for a particular CC system, can exemplars of the creator that is being represented by the system be given?

Dimension 1: Socially Embedded vs. Individualistic

One dimension of difference between creative artists is between those that work as individuals and those that create work in a socially embedded context. No creative artist is entirely divorced from social context, but we focus on those that work directly with others in collaborative creation.

This is rare in the literary arts and uncommon in the visual arts; occasionally small groups will work consistently together over a long period of time (e.g. artistic duos such as Thomson & Craighead, the Chapman brothers and Cardiff & Miller), but literature, visual art and theatre/film writing are dominated by individual creators (comedy writing—

especially for TV—is a notable exception, as are the works of groups such as the *Dogme 95* filmmaking collective). In music, collaboration is more common. This reaches its peak in performance styles such as free improvisation, where a number of performers work together to collaboratively create a work without a preformed plan or an idea of leadership or direction. In some more commercial creative domains, such as advertising, group creative work is standard.

Most work in CC focuses on the individualistic concept of creativity: writing of stories, creation of jokes, composition of melodies, creation of pictures. There has been some interest (Cook and Colton 2011) in mapping out the various contributors to creativity in CC; this is explored futher below in the discussion of computer as medium. Whilst CC systems might create interactive works—for example in game level design (Togelius and others 2010)—it is rare that the system *continues* to be creatively active during interaction.

There are a number of potential reasons for this focus on the individual. These are readily criticisable and there is no reason to believe that all of them are believed by all practitioners, but listing them gives us an initial scoping out of the potential reasons:

- The work in CC is coming out of an AI tradition, which has focused on the idea of the simulation of the individual mind interacting with a task (though multi-agent systems are a counterexample).
- From an artistic perspective, there is a tradition of the "lone genius" in the romanticist tradition in art (Lovejoy 1948), which sees the role of the artist as developing their own authentic and original voice. This idea of the artist as *romantic hero* rejects the idea of collaboration and the development of an ongoing tradition, instead of which the great artist is seen as creating an individual body of work expressing their own personal world-view.
- Creativity might be seen as happening because of various interacting processes within the mind (see e.g. Koestler (1964)). However, people have a curious reluctance to admit hierarchical models of interacting networks, both in intelligence and creativity: creativity/intelligence might be seen as a product of interactions, but it is tempting to contain those interactions to one *level* in a complex system. It is difficult to conceive of a system where creativity is a product of interacting systems within the mind and *also* a network of interacting minds.
- There is a desire to be able to pin down exactly where the creativity is coming from. If a system is embedded in a complex social system with both human and computer agents, it is harder to point to a specific creative act by the computers. An easy criticism of such systems is that all of the creativity is coming from the human agents, and the seeming creativity of the computer agents is replicating, decorating or making trivial responses to or elaborations of the creative acts of the human agents.
- Individual creativity might be seen as the first stage in the development of more sophisticated interactive creative systems; therefore, until CC systems have demonstrated individual creativity, there is no point in tackling the "more complicated" task of group creativity.
- In practice, most creative work is individual; group creativity, in the arts, is confined to specialised areas. Therefore, CC systems are just emulating the world.

There are a number of such collaborative systems that have been created. Consider *Voyager* (Lewis 2000), where heuristics interact within a listening/responding musical system that improvises alongside human musicians; and Sanfilippo's LIES (Sanfilippo 2012), where sound processing systems are connected, the parameters of these interactions being adjusted by the user. Is this a CC system? It is easy to say that the creativity is coming from the user in the form of the parameter changes; but, those might be provoked by sounds from the system, what Blackwell and Young call "strong interactivity", which "depends on instigation and surprise as well as response." (Blackwell and Young 2005).

What might a CC system that was designed to work in a free, collaborative environment look like? Take musical improvisation as an example. One source of inspiration might be the broad guidelines that are given to beginners making a start in improvisation; whilst improvisation might be "free", it is not "anything goes" and there is a strong, often unarticulated, tradition about what is acceptable behaviour in such performances. This needs to be learned by participation and reflection, but guidelines can help to guide beginners so that they are not floundering totally.

Consider, the three guidelines put forward by Dave Smith: Listen, Don't waste sounds, and Develop a sense of social responsibility. How might a CC system attempt to work with these? "Listen." is both trivial (we need to have some means of getting input from the other performers) but, of course, is actually a very deep and complex guideline, especially considering that listening is an active process. Clearly, "listen" means that improvisation should take account of that listening. However, account-taking can easily become trivial, and bejust imitation; learning to *develop* material, and links between different heard sounds, is important to produce depth.

How might a CC system "listen" in this deeper sense? One characteristic of listening is that people frame listening with regard to what has been heard in the past, making subtle distinctions between some objectively very similar sounds, and grouping other sounds together that are objectively different (e.g. (Goto 1971)). One way to do this would be to run a system for a long time, and accumulate a set of listenings. This runs into the problem (Bown and McCormack 2010) of getting people to interact over a long period of time with a "naive" system that isn't providing engaging feedback.

An alternative would be to take inspiration from the idea of an *adult learner* who is new to free improvisation but already has a body of sonic knowledge from listening, speaking and playing an instrument. For example, a system might match listened phrases to a corpus of sonic information—a set of melodies, or a set of nature sounds like birdsong, or an artificially generated set such as sound contours derived from spoken text. Developments in audio information retrieval might mean that such information could be gained from web search. The system could then base responses on these matches, which would bring in a broader, allusive set of responses than those that just work with varying the input based on just that input alone.

Consider the second guideline: "Don't waste sounds.". Again, there is a naive interpretation of this: don't play all of the time. This is simplistic, but could provide the basis for an implementation; indeed, this is how many beginning improvisers might, with some success, interpret it.

However, it has more depth. "Waste" could include aspects of listening: don't waste the other sounds that are going on in the environment, whether by ruining them with your own sounds or by failing to exploit the sounds that have potential. There is a suggestion of depth of contribution: providing a contribution that is constantly striving for more depth and development and eschewing cliché,

Again, simple versions of these could be implemented in a CC system, particularly when combined with the deeper listening ideas above. For example, the agent could measure the complexity of the current activity and hold back in over-complex situations, the agent could hold a medium-term memory of what material is being worked on and wait until a particular piece of material is being "played out" before introducing new material, and so on.

Turning to the final point, how could a computational agent "develop a sense of social responsibility"? The "society" in question clearly includes the fellow players; also perhaps, the audience. One aspect of being socially responsible is recognising when behaviour is regarded by others as gauche or inappropriate. Could a computer agent measure this by the inference of affective states from fellow players (Picard 1997)? Clumsily, this could be done by direct feedback from the players to the agent. Another way in which humans learn social cues is by observing social interactions amongst others; could a creative agent build a model of such interactions within the group of improvisers and then use this model to develop its own interactions? Or, more simply, by analyse whether the material it performs is taken up by fellow players and use that as a proxy for the appropriateness of certain kinds of interaction?

Finally, how could such a system be evaluated? Evaluation of "individualist" creative systems is typically done via the impression that the system makes on the audience. However, a collaborative system could also be evaluated by the fellow players, and such players might use different criteria to that used by (or, indeed, perceptible by) an audience. For example, a creative improvisation system might be rated highly by fellow (human) improvisers if it provides a contextually-sensitive way of provoking the other improvisers to produce more engaging music.

Dimension 2: Tradition-Classicism vs. Romantic Individualism

Another aspect to the notion of the "romantic hero" is the contrast between the artist working in isolation as an individual genius and the artist working in a tradition. This is somewhat different to the above discussion: this section considers the contrast between the artist who pursues their own "individual voice" against the idea of one who is concerned with developing out from and contributing to the development of an ongoing tradition. By contrast, consider a *classicist* view which sees originality happening via the gradual development of a tradition, constantly underpinned by ideals of balance and proportion.

The idea of *romantic originality* was a significant shift in concept of artistic development within the European mainstream tradition, contrasting with earlier traditions about artistic creation being about skillful execution within a style, including reuse and redevelopment of material. The romantic artist creates work that reflects their own, individual, often troubled, engagement with the world (Butler 1981).

The relationship of CC to this dimension is complex. The idea of creativity expressed by Boden's model of *transformational creativity* reflects the classicist notion of a gradually developing tradition; the space is, after all, *transformed*, not *rejected*. But, perhaps too much stall shouldn't be placed on this—after all, all artistic work builds to some extent on previous work, and the most individualist romantic hero uses tools developed by previous generations. Indeed, individual initiative eventually tips over into eccentricity; witness the reviews collected by Slonimsky (2000), or the reception of "outsider art" (Rhodes 2000).

In areas where there is no ground truth, radical novelty is difficult to evaluate. A seemly incomprehensible piece could be madness or genius. In Boden's terms, if a transformation consisted of taking one space and replacing it by another, how would works in this new space be evaluated? For a more sober transformation, evaluation can start with our existing ideas of evaluation and push them a little; in a completely new space there is no corresponding grounding.

The view along this axis therefore gives a contrast to the previous one. Typical CC systems represent individual creativity rather than social creativity; but, they represent individual creativity within a tradition rather than that of the radical outsider.

Dimension 3: Old vs. New Media

Another distinction that can be made in artistic creativity is between so-called *old media* and *new media*. Defining new media is complex and contestable. Manovich presents an initial, naive understanding of new media as those cultural objects that essentially involve the use "of a computer for distribution and exhibition." (Manovich 2001).

He goes on to describe ways in which digital technology has influenced the *process* of creating cultural works. One example is where speedup faciliates a difference in kind rather than just a difference in degree. E.g., real-time rendering of 3D scenes makes interactive games possible as well as improving the production process of traditional animation. Furthermore, computer technologies provoke artists into exploring new creative areas: for example, the notion of *transcoding* (Manovich 2001), i.e. the ready exchange of data between different media formats, makes us think about creating works in which different media streams are created from the same source material, whether in a supportive or disruptive manner.

Computational Creativity: Old or New Media?

Are the products of CC systems *old media* or *new media*? Perhaps counterintuitively, the vast majority of CC systems are in an old media tradition. Whilst the computer is essential in CC, the role it plays is as *creator*; in new media, the computer is essential in the work *as it is presented*.

The work on computational creation of stories (Gervás 2009), poetry (Manurung and others 2000) and jokes (Ritchie 2009) is clearly in this vein: the aim of the vast majority of such systems is to create a work that is presented as words-on-the-page; whether these words are presented on paper or on the computer screen (or read out

loud) is not part of their essence. There are a few examples of literary creativity systems that *are* clearly within the new media tradition: for example, *nn* by Montfort (2007) is an example of the generation of interactive fiction.

In the case of CC systems for music, the landscape is more mixed. Systems such as *Voyager* produce sequences of notes to be performed by a synthesis system (sounding like a traditional instrument) or by a mechanical instrument. However, there are a number of examples that demonstrate how a creative music computing system could work in a new media fashion. For example, Magnus's *evolutionary music concrète* experiments (Magnus 2006) and Sanfilippo's LIES system discussed above show how creative computer systems can create electronic music that is concerned with sound manipulation rather than the production of notes.

CC works in a real new media tradition are rare. There is little CC work producing internet art (Greene 2004) or multimedia works. There are few works that use computationally-based means of organising material such as transcoding, or dynamic creation of work from a database (Manovich 2001), or whose aesthetic is a computational one such as the *database aesthetics* discussed by Vesna (2007).

One example is *Dance Evolution* by Dubbin & Stanley (2010), which uses a computer game engine as the basis for a system whereby characters learn to dance in time to music. The characters are stock video-game images of soldiers; it is not clear whether this was simply a use of the resources that were readily available within the engine, or whether the choice was deliberate. Regardless, this unusual choice of avatar provides a provocative image, reminiscent of various artists attempts to subvert the video-game culture by, for example, the performance of street theatre within MMORPG environments (Greene 2004).

The Creative Networked Computer

Despite CC arising during the "Internet age", the typical CC system produces creative output in a fixed medium (e.g. words, pixels on a screen, MIDI notes) that ignores the networked context of the computer. CC systems that draw on allusions to the world beyond are rare. Most storytelling systems (Gervás 2009) produce stories about a fixed set of ideas and characters. Most of the CC systems for visual arts work in an abstract medium.

Where they *do* provide external reference, this has typically been provided by the system designer directly. One way is that the designer builds into the system some understanding of the external world: for example, whilst the details of how people are drawn in Cohen's *AARON* system (Cohen 1995; Boden 1990) are created by the system, the basic idea of a person-shape is part of the system design. The second way this that the designer might place processes within the system that generate allusions to the world beyond, for example in *ecologically* inspired systems such as those discussed by Bown & McCormack (2010) where the interactions between components of the system are inspired by the kinds of interactions found in natural ecological systems. This could well suggest aspects of the natural world to viewers of the system, even if the presentation is very abstracted; but, the decision to make this allusion is that of the system designer.

There are exceptions. Krzeczkowska et al. (2010) present a system where the initial source material is drawn from current news stories, and keywords extracted are used in web searches for pictures, which are used as the source material for the creation of a collage using Colton's *Painting Fool* system (Colton 2012).

An important part of much art (particularly visual and conceptual art) is connotation: the depth of the work comes from ideas that are suggested, triggering connections that remain under conscious awareness, or revealed via "ah-ha" moments where the link or allusion is suddenly revealed (the idea of CC systems *framing information* as part of their creativity (Colton and others 2011) captures some of this).

Computer as Creator; Computer as Medium

CC researchers might eschew working in new media because of the potential for confusion between two different roles for the computer. For a CC system in a computer-based medium, the computer is playing two roles; that of creator, and that of the medium in which the work is realised. In theory, there is no reason why such a system should not be successful. However, in terms of *evaluation*, the creative computer working in new media is harder to evaluate. There is a question of the role that the "creative" computer played, versus the role played by the broader computational context.

McCormack (2005) has argued that to be taken seriously, CC systems need to create works that are not just examples of "technical fetish", but that are accepted in their own right. This may be why CC systems have avoided new media works: many new media works play with the idea of their technological groundedness in a self-referential and essential way, and CC system creators avoid building systems that work in such media to avoid accusations of "mere" technical obsession. But, awareness of origins need not reflect an over-obsession with a trivial part of the work; indeed, the *lack* of any sense of biographical depth is one of the shallowing aspects of many CC-produced works.

One new direction would be to build CC systems that explore and celebrate what computers are good at. One inspiration for this comes from John Cage's account of his development:

> After I had been studying with him for two years, Schoenberg said, "In order to write music, you must have a feeling for harmony." I explained to him that I had no feeling for harmony. He then said that I would always encounter an obstacle, that it would be as though I came to a wall through which I could not pass. I said, "In that case I will devote my life to beating my head against that wall." (John Cage, *Lecture on Indeterminacy* (Cage 1973))

A CC system could adopt a similar attitude, to acknowledge that computers might not be capable of "passing off" certain aspects of human creativity (e.g. creating fluent natural language text) but that certain new forms of creativity are facilitated that draw on the specific capabilities of computers and the computational infrastructure.

Let us consider two recent works. Kessel's 2011 *Photography In Abundance* where a day's uploads to *Flickr* were printed out and placed in a gallery—hundreds of thousands of photos. Such a work could not have been achieved without computational infrastructure; yet, is this *computational creativity*? Presumably not, as the creative decision about

what to do with the computational capability of Flickr was made by the human artist. Yet, the computational infrastructure seems to have played a stronger role here than that of medium—the contribution of the computer to this work is greater than the contribution of paint to a painting. Perhaps we need to get used to the idea of artists working in a richer medium, which "pushes back" in terms of creative contribution much more than a conventional medium.

The second example is my own 2011 piece *Blank*: nine printed panels containing image search hits for "blank". Most of the images are of empty objects: map outlines, blank signs, empty music paper. One image in the seventh panel consists of a photo of a number of gun cartridges—referencing "blanks" as a cartridge without a bullet.

Had this work been produced unaided by a human artist, people might ascribe it creative depth. Having set up the expectation of empty, neutral images, there is this single flash of violence in the middle of one panel, causing us to reinterpret the meaning of the emptyness. Of course, there is no *intention* to create this as such; it is "mere" serendipity. Yet, it is deeper than the readerly (Barthes 1970) interpretation of a purely aleatoric work; the images were "chosen" by a process that has a huge amount of infrastructure behind it. Is this any shallower than some purely human-produced work which has "come to mind" as the end result of the memory structures produced by the artist's life experiences?

References

Barthes, R. 1970. *S/Z*. Editions du Seuil.

Blackwell, T., and Young, M. 2005. Live algorithms. *AISB Quarterly* 122:7–9.

Boden, M. A. 1990. *The Creative Mind, Myths and Mechanisms*. Weidenfeld.

Bown, O., and McCormack, J. 2010. Taming nature: Tapping the creative potential of ecosystem models in the arts. *Digital Creativity* 21(4):215–231.

Butler, M. 1981. *Romantics, Rebels and Reactionaries: English Literature and its Background 1760-1830*. Oxford University Press.

Cage, J. 1973. *Silence*. Marion Boyars.

Cohen, H. 1995. The further exploits of AARON, painter. *Stanford Humanities Review* 4(2).

Colton, S., et al. 2009. Computational creativity: Coming of age. *AI Magazine* 30(3):11–14.

Colton, S., et al. 2011. Computational creativity theory: The face and idea models. In *Proceedings of the 2011 International Conference on Computational Creativity*.

Colton, S. 2012. The painting fool. http://www.thepaintingfool.com/ (visited April 2012).

Cook, M., and Colton, S. 2011. Automated collage generation—with more intent. In *Proceedings of the 2011 International Conference on Computational Creativity*.

Dubbin, G., and Stanley, K. O. 2010. Learning to dance through interactive evolution. In Di Chio, C., et al., eds., *Applications of Evolutionary Computation*, 331–340. Springer.

Gervás, P. 2009. Computational approaches to storytelling and creativity. *AI Magazine* 30(3):49–62.

Goto, H. 1971. Auditory perception by normal japanese adults of the sounds "l" and "r". *Neuropsychologia* 9(3):317–323.

Greene, R. 2004. *Intenet Art*. Thames and Hudon.

Koestler, A. 1964. *The Act of Creation*. Macmillan.

Krzeczkowska, A., et al. 2010. Automated collage generation—with intent. In Ventura, D., et al., eds., *Proceedings of the International Conference on Computational Creativity*, 36–40.

Lewis, G. E. 2000. Too many notes: Computers, complexity and culture in *Voyager*. *Leonardo Music Journal* 10:33–39.

Lovejoy, A. 1948. On the discrimination of romanticisms. In *Essays in the History of Ideas*. Johns Hopkins University Press.

Magnus, C. 2006. Evolutionary musique concrète. In Rothlauf, F., et al., eds., *Applications of Evolutionary Computing*, volume 3907 of *Lecture Notes in Computer Science*, 688–695. Springer Berlin / Heidelberg.

Manovich, L. 2001. *The Language of New Media*. MIT Press.

Manurung, H. M., et al. 2000. Towards a computational model of poetry generation. In *Proceedings of AISB Symposium on Creative and Cultural Aspects and Applications of AI and Cognitive Science*, 79–86.

McCormack, J. 2005. Open problems in evolutionary music and art. In Rothlauf, F., et al., eds., *Applications of Evolutionary Computing*, volume 3449 of *Lecture Notes in Computer Science*, 428–436. Springer Berlin / Heidelberg.

Montfort, N. 2007. Ordering events in interactive fiction narratives. In Magerko, B. S., and Reidl, M. O., eds., *Intelligent Narrative Technologies: Papers from the 2007 AAAI Fall Symposium*, 87–94. AAAI Press.

Pease, A., and Colton, S. 2011. 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*.

Picard, R. 1997. *Affective Computing*. MIT Press.

Rhodes, C. 2000. *Outsider Art: Spontaneous Alternatives*. Thames and Hudson.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17(1):67–99.

Ritchie, G. 2009. Can computers create humor? *AI Magazine* 30(3):71–81.

Sanfilippo, D. 2012. Dario Sanfilippo. http://dariosanfilippo.tumblr.com/ (visited April 2012).

Slonimsky, N. 2000. *Lexicon of Musical Invective*. W.W. Norton and Co.

Togelius, J., et al. 2010. Search-based procedural content generation. In Di Chio, C., et al., eds., *Applications of Evolutionary Computation*. Springer. 141–150.

Vesna, V., ed. 2007. *Database Aesthetics: Art in the Age of Information Overflow*. University of Minnesota Press.

Whence is Creativity?

Bipin Indurkhya (bipin@agh.edu.pl)
Department of Computer Science, AGH University of Science and Technology, Cracow, Poland
Cognitive Science Lab, International Institute of Information Technology, Hyderabad, India

Abstract

We start with a critical examination of the traditional view of creativity in which the creator is the major player. We analyze many different examples to point out that the origin of all different creativity scenarios is rooted in the viewer-artifact interaction. To recognize this explicitly, we propose an alternative formulation of creativity by putting the viewer in the driver's seat. We examine some implications of this formulation, especially for the role of computers in creativity, and argue that it captures the essence of creativity more accurately.

Introduction: Traditional View of Creativity

In a typical creativity scenario, there is a creator, a product and the audience. The creator creates the product, and the audience appreciates it. The creativity is almost always imputed to the creator. One could talk about a creative process, but that is again associated with the creator. The audience plays a role, and has been dubbed *the field* by Csikszentmihalyi (1996), but somewhat indirectly, and even then, one usually makes a distinction between a popular artist and a creative artist, which are not synonymous.

This framework is questioned in this paper by analyzing a number of creativity scenarios and tracing the root factor that allows us to dub them creative. It is not the first time that these issues are being raised, for they are well known in the literature. But it may be the first time we are bringing them all together to suggest that perhaps there is something fundamentally wrong with this view of creativity. We then propose an alternative formalism for creativity by looking at it from the audience's point of view. Though this may seem almost heretical at first sight, we argue that it provides a more accurate framework to address various issues surrounding creativity, including the role of computers therein.

Analysis of Some Creativity Scenarios

We present here several creativity scenarios and analyze them to identify the root cause as to why they are labeled creative.

Case of Creative Individuals

If we try to think of creative people, who comes to mind? Perhaps Einstein, Mozart, Michelangelo or Leonardo da Vinci. In the modern times, we might think of Steve Jobs. But what do we mean when we say that they are creative?

Perhaps music came naturally to Mozart. In a letter to his father on Nov. 8, 1777, he wrote: "I cannot write in verse, for I am no poet. I cannot arrange the parts of speech with such art as to produce effects of light and shade, for I am no painter. Even by signs and gestures I cannot express my thoughts and feelings, for I am no dancer. But I can do so by means of sounds, for I am a musician." However, what makes his work great is because of the way people have responded to his music over more than two centuries. (See also Kozbelt 2005; and Painter 2002.)

Steve Jobs has sprouted and nurtured many creative ideas, but again it is how people responded to the artifacts based on his ideas, like Mac, iPod, and iPhone, that is the key factor in his having become an icon of technological innovation and creativity. And one could easily dispute whether he was creative when it came to his clothes.

Einstein's brain was preserved after his death so that people can study it to get any clues about the biological basis for creativity. But here also it is the impact of his theory of relativity, and its eventual acceptance by the scientific community that was a key factor in him becoming an icon of scientific creativity of the twentieth century. Moreover, Einstein was also dogmatic at times, perhaps the most famous case being his rejection of Alexander Friedmann's expanding universe hypothesis (Singh 2004).

Needless to say, we are not trying to argue that these people were not creative, but merely to point out that they were creative some of the times, and in some areas; and, more importantly for the discussion here, that we determine when they were creative based on how the audience responded to their ideas or artifacts.

Creativity in Mentally Different Individuals

Take the case of Stephen Wiltshire, discussed in Sacks (1995). He has an amazing ability to draw a landscape from memory after seeing it only once. Though he is diagnosed with autism, his work is highly regarded both by critics and general population. He was awarded *Member of the Order of the British Empire* for services to art in 2006. So he is no doubt a very creative person, no matter which criterion one chooses to apply.

But let us think about it a minute. What do we mean by saying that he is creative? His work has a certain style, level of details that most people cannot reach, aesthetic appeal, and all that. As with Mozart, we can go further and say that perhaps this is the way he expresses himself naturally: just like you and I might describe what we did on our last summer vacation, he draws fantastic landscapes. The

landscapes are fantastic to us, his audience, and that is the crucial factor in his being recognized as a creative genius.

We can now throw in here examples of people with schizophrenia or brain damage, savants or manic-depressive people, and so on (Sawyer 2006). When these people produce work that is considered creative, it is exclusively the evaluation of the audience that is the key factor in this judgment. For many of them, this is their mode of being, and it could not have been otherwise. Often the intention is missing as well. (See also Abraham *et al.* 2007; Glicksohn 2011.)

Cultural Creativity

In many cultures, art is practiced as a group activity. For example, Maduro (1976) provides a study of *Mewari* painting community in a village in the Rajasthan province of India. It describes a strictly hierarchical group with each artist belonging to one of the *laborers, master craftsmen* or *creative artists* class. They mostly copy existing forms and patterns, with rarely an innovation, at least in the way it is considered in the Western art.

Sawyer (2006) discusses this and many other examples to argue that for such scenarios, one needs to take into account cultural context to evaluate creativity, and novelty may be neither necessary nor sufficient. However, the audience response is still a key factor. Notice that the culture itself can be an audience.

Computer and Creativity

We now consider the scenario at the other extreme, when technically there is no creator and the intent is missing. Even though the last two or three decades have seen a steady progress in the development of computer systems that produce artifacts in the domain of visual art (Cohen 1981; McCorduck 1991), music (Chordia & Rae 2010; López *et al.* 2010; Monteith *et al.* 2010), literature (Kurzweil 2001, Pérez y Pérez *et al.* 2010); and so on, generally they have received a negative press as regard to their creativity: computers cannot have emotions, programs do not have intents, creativity cannot be algorithmic, etc. etc. (Boden 2009; Sawyer 2006). In fact, such views blatantly expose the implicit assumptions underlying creativity: namely that it crucially needs a creator with emotions, intentions, and such. However, we have just seen a number of scenarios above involving humans where the intentions and emotions are missing and, even when they are there, what determines whether something is creative or not is the audience response. So why should we not apply the same yardstick for computer-generated artifacts?

Creativity in Viewer-Artifact Interaction

In all the scenarios above, we have seen that it is the audience response that determines whether an action (of the creator or the group), a process, or a product is creative. So suppose we drop the pretense, stop being apologetic about it, and embrace this view formally: We define creativity *as the process by which a cognitive agent acquires a novel perspective that is useful (or meaningful) to it in some way by interacting with an object or a situation.*

There are two aspects of this definition that need to be emphasized here. One is that we are taking completely the audience's perspective here, so the creator is not even mentioned. Needless to say, this is not the first time that such a position is articulated. Barthes' (1977) concluded: "We know that to restore to writing its future, we must reverse its myth: the birth of the reader must be ransomed by the death of the Author," and he traced this view to even earlier scholars. Moreover, even when one does not take this extreme position, most accounts of creativity do acknowledge the role of audience (Cropley *et al.* 2011; Csikszentmihalyi 1996; Horn & Salvendy 2006; Maher 2010). Our aim here is to explore the implications of this authorless view of creativity, especially for computational systems.

Secondly, both the novelty and the usefulness are defined from the agent's personal point of view. So in this way, this definition refers to little-c creativity, or P-creativity (Boden 1990; Kaufman & Beghetto 2009). However, as we will see soon, it can be extended to Big-C or H-creativity.

Implications of the Proposed View

We will now discuss how many intuitive notions associated with creativity can be rooted in the formulation proposed above. We start with the four Ps of creativity (Runco & Kim 2011).

Product: Potential of Artifacts for Creativity

If an agent can interact with an object to get a novel and useful perspective, we can impute creative potential to that object. If many people can get a novel and useful perspective, it should be emphasized that the perspective that each agent gets need not be the same, or need not be useful in the same way. So, for instance, different agents may see a work of modern art in various ways, and find it meaningful in different ways, and some may not see anything at all.

Though most accounts of creativity incorporate audience response to some extent, the extreme view we are examining here would allow Oracle readings of tea leaves and such as creative interactions. (See also Indurkhya 2007.) Just to contrast, in the traditional view, such objects lack a creator so, by association, lack creativity as well. This is a common argument to deny creativity to computer systems.

Process: Creativity in Generating Artifacts

Now we can take another step back and consider the process of generating a creative artifact. In other words, we need to consider the process of generating an artifact with which a viewer can interact and get a novel and useful perspective. So the viewer is always hovering in the background, and has a significant impact on whether the process is really generating an ordinary artifact as opposed to a creative artifact.

The main implication for modeling the generative aspect of creativity is that we cannot pursue it without considering

the audience, and making some assumptions about how they are likely to interact with the generated artifacts.

Person: Creativity of an Individual/Group

We can step back some more and consider the creativity of an individual or a group of individuals. Here we are looking at the ability of the individual to generate artifacts with which viewers can interact and get novel and useful perspectives. So the viewer is again in the background and is playing a critical role. Moreover, even though we speak of this person or that person as being creative, we are really focusing on certain artifacts that they have generated in their career, which have given their audience some novel and useful perspective.

The implication of this is that though we can certainly study personality traits of certain individuals who generated some artifacts during their career that were deemed creative by the audience, it does not follow that those personality traits in a different culture, in a different context and with a different audience will necessarily result in the generation of artifacts that would also be considered creative. This point is highlighted in the essay 'Late Bloomers' (Gladwell 2009), where early geniuses are contrasted with late bloomers. The relevant point here is that whether a work is accepted by the audience or not does not depend much on whether it was produced early or late in the career, but on the kind of work and the context and the culture in which it was produced.

Press: Context, Culture and H-Creativity

Press refers to the environmental factors that have an influence on the generation of the artifact; but that is taking the traditional perspective, where the focus is on the creator. If we are putting the viewer in the drivers' seat, then an analogous set of environmental factors can be identified that determine how a work is received by the viewer, and whether it is successful or not.

Let us first consider the artifact interaction with an individual viewer. Clearly, the context in which a viewer interacts with the artifact can have a major influence in what perspective is gleaned from it, and whether it is novel or meaningful. The most classic example of this might be Marcel Duchamp's *Fountain,* which was a urinal turned around. (See also 'When is Art?' in Goodman 1978.) There is also the effect of the viewer's background knowledge: when one views the *Parthenon* in Athens, one's knowledge of the history and culture of ancient Greek certainly effects one's perceptions and aesthetic experience.

Moving to larger groups and societies, there are many instances when a novel and potentially useful idea was not successful when introduced in one context, but the same idea was a big hit in another context. We mentioned above the example of Alexander Friedmann's expanding universe hypothesis, which was rejected when introduced because of Einstein's influence but was widely heralded later. Wegener's (1966) theory of continental drift suffered a similar fate when it was first introduced in 1915, even though that was no fault of Einstein.

There are also cases where the theory, although novel and carefully worked out, never received acceptance: for instance, Velikovsky's theory, which hypothesized Earth's encounters with a large comet expelled from Jupiter and provided explanations for many biblical events (Casti 1989, pp. 7–10).

History of marketing and product development also provides many such examples that are studied in business schools all over. Gladwell (2009), for instance, recounts Jim Wigon's not-so-successful odyssey to develop creative ketchup flavors, and contrasts this with mustard and spaghetti sauce, for which similar efforts were more readily accepted by the consumers.

The implication of all this simply is that one needs to study all these contextual factors that make an idea or an artifact novel and meaningful, and thereby eligible for the 'creativity' label. But this is essentially what is called H-creativity: novelty and usefulness for a culture or society. We should emphasize here that this novelty and usefulness with respect to a culture is not the same as popularity. Certain ideas or artifact can be popular in a society without being considered novel (by the members of the society themselves), and vice versa.

Who is Creative?

This is one question that is often asked with regard to artificial creativity systems. For example, who is being creative when the computer program Aaron generates a painting in which many people see some aesthetic value? Is it the program? Is it the programmer? In many complex computational systems, the programmer cannot see all the consequences of what their system can generate, and can be quite surprised by the artifacts it produces.

We would like to argue that this question is meaningless in the framework of creativity we are proposing here. Just to get away from the computational system scenario, consider a painting by a schizophrenic person that many people find interesting and insightful. Now obviously, the schizophrenic person is the creator of the painting. But is she or he creative? How can we determine this? Perhaps it is a natural way for them to express themselves; they may not see what all the fuss is about the artifact they created. In other words, if we zoom out and look at the larger picture, they are creating artifacts that many people find insightful, and so in that sense we can ascribe them creativity as explained above. But if we zoom in on the processes by which they generate these artifacts, where is the creativity? They may not be trying to generate some aesthetically pleasing object, and may not even be aware of any audience. The point is that there is nothing distinctive about the generation process itself that we can label it as creative.

Creativity and Computational Systems

Once we acknowledge that it is meaningless to ask who is being creative, the stigma surrounding the potential creativity of computational systems recedes away. The goal becomes simply to create artifacts that give them some novel and meaningful perspectives.

This seemingly small shift of focus has far reaching consequences, and our society is already moving towards it. To start with, modeling the audience, their cultural tastes and preferences, their cognitive processes that influence their response to novel stimuli, and so on, becomes very crucial. In the last 10-15 years, research in neuroscience has revealed that at least some of our aesthetic values are hardwired in the structure of the brain (Ramachandran & Hirstein 1999; Zeki 2000). Then to add to that, machine learning techniques can *learn* about the cultural preferences of an audience based on the past data. For instance, Ni *et al.* (2011) trained their program with the official UK top-40 singles chart over the past 50 years to learn as to what makes a song popular. A program like this might successfully predict, for instance, the winner of the future Eurovision competitions.

To reiterate a subtle point, creativity is not the same as popularity. So to be able to predict whether a song, or a book, or a video will become popular (Szabo & Huberman 2010) is not the same thing as evaluating their creativity. Nonetheless, we expect that similar techniques, perhaps with some adaptations, are more likely to yield the key to creativity.

Going one step further, once audience-based models of creativity are articulated, we can design, implement and experiment with computational systems that generate artifacts that are more likely to appeal to the audience, both with respect to their novelty and meaningfulness. One could even argue that computational systems are more ideally suited than humans to explore this space of creative possibilities (Harry 1992; Indurkhya, to appear).

Designing Creativity-Support Systems

A related issue is how to stimulate and enhance creativity in people, if it is possible at all. Indeed, a number of approaches have been proposed and tried out over the years (de Bono 1975; Gordon 1961; Holstein 1970; Rodari 1996; Shapira & Liberman 2009). One key observation that recurs in many of these studies is that trying to associate unrelated objects or situations stimulates creativity (Indurkhya 2010). In our past research, we have explored some approaches to design creativity-support systems based on this observation (Indurkhya 1997; Indurkhya *et al.* 2008; Ishii *et al.* 1998), but much more remains to be done.

Conclusions

Though most of the research on computational creativity has implicitly assumed that the creative value is in the artifact, they have been sort of apologetic about it. For example, Colton (2008) argues that it is not enough to generate an interesting or creative artifact, but one must also take into account the process by which the artifact was generated. Krzeczkowska *et al.* (2010) took pains to project some notion of purpose in their painting tool so that it might be perceived as creative.

In this paper we have sought to not only drop this veil of apology, but move to the other extreme by proposing a formulation of creativity that puts the onus on the viewer by characterizing it as the process of getting a new and meaningful insight about an object or situation.

We have argued that this formulation reflects more accurately what actually goes on in the whole creativity cycle. Moreover, other situations, like when we deem an artifact or an individual as being creative, we are really implicitly relying on the viewer-artifact interaction to make this judgment. Therefore, creativity of an agent and that of an artifact are best seen as derived concepts based on our proposed formulation of creativity.

We hope that this will stimulate further discussion about the nature of creativity and, more importantly, will generate new approaches to the design and development of computational creativity systems.

References

Abraham, A., Windmann S., McKenna P., & Güntürkün, O. 2007. Creative Thinking in Schizophrenia: The Role of Executive Dysfunction and Symptom Severity. *Cognitive Neuropsychiatry 12*(3), 235–258.

Barthes, R. 1977. The Death of the Author. In R. Barthes, *Image Music Text* (trans. S. Heath), London: Fontana. (Original work published 1967.)

Boden, M. A. 1990. *The Creative Mind*, London: Weidenfeld and Nicholson.

Boden, M. A. 2009. Computer Models of Creativity, *AAAI AI Magazine 30*(3), 23-33.

Casti, J. L. 1989. *Paradigms lost: Images of man in the mirror of science.* New York: William Morrow & Co.

Chordia, P., & Rae, A. 2010. Tabla Gyan: An Artificial Tabla Improviser, *Proceedings of the International Conference on Computational Creativity: ICCC-X,* Lisbon, Portugal.

Cohen, H., 1981. On Modeling of Creative Behavior. Rand Corporation Report Series, P-6681.

Colton, S., 2008. Creativity versus the Perception of Creativity in Computational Systems. *Proceedings of the AAAI Spring Symposium on Creative Systems*.

Cropley, D.H., Kaufman, J.C., Cropley, A. J. 2011. Measuring Creativity for Innovation Management, *Journal of Technology Management & Innovation 6*(3), 13-30.

Csikszentmihalyi, M. 1996. *Creativity: Flow and the psychology of discovery and invention*. New York: HarperCollins.

de Bono, E. 1975. *New Think: The Use of Lateral Thinking in the Generation of New Ideas.* New York: Basic Books.

Gladwell, M. 2009. *What the Dog Saw and Other Adventures,* London: Allen Lane.

Glicksohn, J. 2011. Schizophrenia and Psychosis. In M.A. Runco & S.R. Pritzker (Eds.) *Encyclopedia of Creativity, (2nd ed.),* New York: Academic Press.

Goodman, N. 1978. *Ways of Worldmaking.* Indianapolis, IN (USA): Hackett Publishing.

Gordon, W.J.J. 1961. *Synectics: The Development of Creative Capacity.* New York: Harper & Row.

Harry, H. 1992. On the role of machines and human persons in the art of the future. *Pose 8,* 30–35.

Holstein B.I. 1970. *Use of Metaphor to Induce Innovative Thinking in Fourth Grade Children.* Ph.D. thesis, School of Education, Boston University, Boston, Mass.

Horn, D. and Salvendy, G. 2006. Consumer-based assessment of product creativity: A review and reappraisal. *Human Factors and Ergonomics in Manufacturing & Service Industries 16*(2), 155-175.

Indurkhya, B. 1997. On Modeling Creativity in Legal Reasoning. *Proceedings of the Sixth International Conference on AI and Law,* Melbourne, Australia.

Indurkhya, B. 2007. Creativity in Interpreting Poetic Metaphors. In T. Kusumi (Ed.) *New Directions in Metaphor Research,* Tokyo: Hitsuji Shobo.

Indurkhya, B. 2010. On the Role of Metaphor in Creative Cognition. *Proceedings of the International Conference on Computational Creativity: ICCC-X,* Lisbon, Portugal.

Indurkhya, B. to appear. Computers and Creativity. In C. Forceville & T. Veale (eds.) *Agile Mind,* Mouton.

Indurkhya, B., Kattalay, K., Ojha, A., & Tandon, P. 2008. Experiments with a Creativity-Support System based on Perceptual Similarity. In H. Fujita and I. Zualkernan (eds.) *New Trends in Software Methodologies, Tools and Techniques,* Amsterdam: IOS Press.

Ishii, Y., Indurkhya, B., Inui, N., Nose, T., Kotani, Y., & Nishimura, H. 1998. A System Based on Rodari's 'Estrangement' Principle to Promote Creative Writing in Children. *Proceedings of Ed-Media & Ed-Telecom 98,* Freiburg, Germany.

Kaufman, J.C., & Beghetto, R.A. 2009. Beyond Big and Little: The Four C Model of Creativity. *Review of General Psychology 13*(1), 1–12.

Kozbelt, A. 2005. Factors Affecting Aesthetic Success and Improvement in Creativity: A Case Study of the Musical Genres of Mozart. *Psychology of Music 33*(3), 235–255.

Krzeczkowska, A., Colton, S., & Clark S. 2010. Automated Collage Generation – With Intent. *Proceedings of the International Conference on Computational Creativity: ICCC-X,* Lisbon, Portugal.

Kurzweil, R. 2001. *Cybernetic Poet*.
http://www kurzweilcyberart.com/poetry/ Last accessed on 29 January, 2012.

López, A.R., Oliveira, A.P., & Cardoso, A. 2010. Real-Time Emotion-Driven Music Engine. *Proceedings of the International Conference on Computational Creativity: ICCC-X,* Lisbon, Portugal.

Maduro, R. 1976. *Artistic creativity in a Brahmin painter community.* Berkeley, CA: Center for South and Southeast Asian Studies.

Maher, M.L. 2010. Evaluating Creativity in Humans, Computers, and Collectively Intelligent Systems, *DESIRE'10: Creativity and Innovation in Design,* Aarhus, Denmark.

McCorduck, P. 1991. *Aaron's code: Meta-art, artificial intelligence, and the work of Harold Cohen*. New York: W. H. Freeman.

Monteith, K., Martinez, T. & Ventura, D. 2010. Automatic Generation of Music for Inducing Emotive Response. *Proceedings of the International Conference on Computational Creativity: ICCC-X,* Lisbon, Portugal.

Ni, Y., Santos-Rodriguez, R., Mcvicar, M., & De Bie, T. 2011. Hit Song Science Once Again a Science? Presented at the *4th International Workshop on Machine Learning and Music: Learning from Musical Structure*, Sierra Nevada, Spain.

Painter, K. 2002. Mozart at Work: Biography and a Musical Aesthetic for the Emerging German Bourgeoisie. *The Musical Quarterly 86*(1), 186–235.

Pérez y Pérez, R., Negrete, S., Peñalosa, E., Ávila, R., Castellanos, V., & Lemaitre, C. 2010. MEXICA-Impro: a Computational Model for Narrative Improvisation. *Proceedings of the International Conference on Computational Creativity: ICCC-X,* Lisbon, Portugal.

Ramachandran, V.S., & Hirstein, W. 1999. The Science of Art: A neurological theory of aesthetic experience, *Journal of Consciousness Studies 6*(6-7), 15–51.

Rodari, G. 1996. *The Grammar of Fantasy* (J. Zipes, Trans.) New York: Teachers & Writers Collaborative.

Runco, M.A., & Kim, D. 2011. The Four Ps of Creativity: Person, Process, Product and Press. In M.A. Runco & S.R. Pritzker (Eds.) *Encyclopedia of Creativity, (2nd ed.),* New York: Academic Press.

Sacks, O. 1995. *An Anthropologist on Mars: Seven Paradoxical Tales*. New York: Alfred A. Knopf.

Sawyer, K. 2006. *Explaining Creativity.* Oxford University Press.

Shapira, O., & Liberman N. 2009. An Easy Way to Increase Creativity. *Scientific American (Mind Matters),* July 21, 2009.

Singh, S. 2004. *Big Bang.* New York: Harper Collins.

Szabo, G., & Huberman, B.A. 2010. Predicting the Popularity of Online Content. *Communications of the ACM 53*(8), 80–88.

Wegener, A. 1966. *The origin of continents and oceans (4th ed.)* (trans. J. Biram), London: Dover (Original work published 1915.)

Zeki, S. 2000. *Inner Vision: An Exploration of Art and the Brain.* Oxford (UK): Oxford University Press.

Computational and Collective Creativity: Who's Being Creative?

Mary Lou Maher
University of Maryland
mlmaher@umd.edu

Abstract

Creativity research has traditionally focused on human creativity, and even more specifically, on the psychology of individual creative people. In contrast, computational creativity research involves the development and evaluation of creativity in a computational system. As we study the effect of scaling up from the creativity of a computational system and individual people to large numbers of diverse computational agents and people, we have a new perspective: creativity can ascribed to a computational agent, an individual person, collectives of people and agents and/or their interaction. By asking "Who is being creative?" this paper examines the source of creativity in computational and collective creativity. A framework based on ideation and interaction provides a way of characterizing existing research in computational and collective creativity and identifying directions for future research.

Human and Computational Creativity

Creativity is a topic of philosophical and scientific study considering the scenarios and human characteristics that facilitate creativity as well as the properties of computational systems that exhibit creative behavior. "The four Ps of creativity", as introduced in Rhodes (1987) and more recently summarized by Runco (2011), decompose the complexity of creativity into separate but related influences:

- Person: characteristics of the individual,
- Product: an outcome focus on ideas,
- Press: the environmental and contextual factors,
- Process: cognitive process and thinking techniques.

While the four Ps are presented in the context of the psychology of human creativity, they can be modified for computational creativity if process includes a computational process. The study of human creativity has a focus on the characteristics and cognitive behavior of creative people and the environments in which creativity is facilitated. The study of computational creativity, while inspired by concepts of human creativity, is often expressed in the formal language of search spaces and algorithms.

Why do we ask who is being creative? Firstly, there is an increasing interest in understanding computational systems that can formalize or model creative processes and therefore exhibit creative behaviors or acts. Yet there are still skeptics that claim computers aren't creative, the computer is just following instructions. Second and in contrast, there is increasing interest in computational systems that encourage and enhance human creativity that make no claims about whether the computer is being or could be creative. Finally, as we develop more capable socially intelligent computational systems and systems that enable collective intelligence among humans and computers, the boundary between human creativity and computer creativity blurs. As the boundary blurs, we need to develop ways of recognizing creativity that makes no assumptions about whether the creative entity is a person, a computer, a potentially large group of people, or the collective intelligence of human and computational entities. This paper presents a framework that characterizes the source of creativity from two perspectives, ideation and interaction, as a guide to current and future research in computational and collective creativity.

Creativity: Process and Product

Understanding the nature of creativity as process and product is critical in computational creativity if we want to avoid any bias that only humans are creative and computers are not. While process and product in creativity are tightly coupled in practice, a distinction between the two provides two ways of recognizing computational creativity by describing the characteristics of a creative *process* and separately, the characteristics of a creative *product*. Studying and describing the processes that generate creative products focus on the cognitive behavior of a creative person or the properties of a computational system, and describing ways of recognizing a creative product focus on the characteristics of the result of a creative process.

When describing creative processes there is an assumption that there is a space of possibilities. Boden (2003) refers to this as conceptual spaces and describes these spaces as structured styles of thought. In computational systems such a space is called a state space. How such spaces are changed, or the relationship between the set of known products, the space of possibilities, and the potentially creative product, is the basis for describing processes that can generate potentially creative artifacts.

There are many accounts of the processes for generating creative products. Two sources are described here: Boden (2003) from the philosophical and artificial intelligence perspective and Gero (2000) from the design science perspective. Boden (2003) describes three ways in which creative products can be generated: combination, exploration,

and transformation: each one describes the way in which the conceptual space of known products provides a basis for generating a creative product and how the conceptual space changes as a result of the creative artifact. Combination brings together two or more concepts in ways that hasn't occurred in existing products. Exploration finds concepts in parts of the space that have not been considered in existing products. Transformation modifies concepts in the space to generate products that change the boundaries of the space. Gero (2000) describes computational processes for creative design as combination, transformation, analogy, emergence, and first principles. Combination and transformation are similar to Boden's processes. Analogy transfers concepts from a source product that may be in a different conceptual space to a target product to generate a novel product in the target's space. Emergence is a process that finds new underlying structures in a concept that give rise to a new product, effectively a re-representation process. First principles as a process generates new products without relying on concepts as defined in existing products.

While these processes provide insight into the nature of creativity and provide a basis for computational creativity, they have little to say about how we recognize a creative product. As we move towards computational systems that enhance or contribute to human creativity, the articulation of process models for generating creative artifacts does not provide an evaluation of the product. Computational systems that generate creative products need evaluation criteria that are independent of the process by which the product was generated.

There are also numerous approaches to defining characteristics of creative products as the basis for evaluating or assessing creativity. Boden (2003) claims that novelty and value are the essential criteria and that other aspects, such as surprise, are kinds of novelty or value. Wiggins (2006) often uses value to indicate all valuable aspects of a creative products, yet provides definitions for novelty and value as different features that are relevant to creativity. Oman and Tumer (2009) combine novelty and quality to evaluate individual ideas in engineering design as a relative measure of creativity. Shah, Smith, and Vargas-Hernandez (2003) associate creative design with ideation and develop metrics for novelty, variety, quality, and quantity of ideas. Wiggins (2006) argues that surprise is a property of the receiver of a creative artifact, that is, it is an emotional response. Cropley and Cropley (2005) propose four broad properties of products that can be used to describe the level and kind of creativity they possess: effectiveness, novelty, elegance, genesis. Besemer and O'Quin (1987) describe a Creative Product Semantic Scale which defines the creativity of products in three dimensions: novelty (the product is original, surprising and germinal), resolution (the product is valuable, logical, useful, and understandable), and elaboration and synthesis (the product is organic, elegant, complex, and well-crafted). Horn and Salvendy (2006) after doing an analysis of many properties of creative products, report on consumer perception of creativity in three critical perceptions: affect (our emotional response to the product), importance, and novelty. Goldenberg and Mazursky (2002) report on research that has found the observable characteristics of creativity in products to include "original, of value, novel, interesting, elegant, unique, surprising."

Amabile (1982) says it most clearly when she summarizes the social psychology literature on the assessment of creativity: While most definitions of creativity refer to novelty, appropriateness, and surprise, current creativity tests or assessment techniques are not closely linked to these criteria. She further argues that "There is no clear, explicit statement of the criteria that conceptually underlie the assessment procedures." In response to an inability to establish and define criteria for evaluating creativity that is acceptable to all domains, Amabile (1982, 1996) introduced a Consensual Assessment Technique (CAT) in which creativity is assessed by a group of judges that are knowledgeable of the field. Since then, several scales for assisting human evaluators have been developed to guide human evaluators, for example, Besemer and O'Quin's (1999) Creative Product Semantic Scale, Reis and Renzulli's (1991) Student Product Assessment Form, and Cropley et al's (2011) Creative Solution Diagnosis Scale.

Maher (2010) presents an AI approach to evaluating creativity of a product by measuring novelty, value and surprise that provides a formal model for evaluating creative products. Novelty is a measure of how different the product is from existing products and is measured as a distance from clusters of other products in a conceptual space, characterizing the artifact as similar but different. Value is a measure of how the creative product compares to other products in its class in utility, performance, or attractiveness. The measure of value uses clustering algorithms and distance measures operating on the value attributes of existing products. Surprise has to do with how we develop expectations for the next new idea. This is distinguished from novelty because it is based on tracking the progression of one or more attributes, and changing the expected next difference.

Computational creativity can be described by identifying the generative processes that are associated with being creative and how the process changes the conceptual space. Alternatively, computational creativity can be asserted when the product is recognized as creative, independently of the process. However, computational creativity is more complicated than a single process that generates a self-contained product, partly due to the different roles that people and computers play in computational creativity but also due to recent phenomena of scaling up participation to achieve collective human-computer creativity.

Collective Creativity

Collective creativity is associated with two or more people contributing to a creative process. Using the internet to develop and encourage creative communities has led to large scale collective creativity. Some examples of such creative communities are: *Designcrowd.com*, *Quirky.com, 99Designs.com* and *OpeningDesign.com*. Designcrowd and 99Designs are examples of websites that source creative work from a very large community of people that identify themselves as designers. *Quirk*y crowdsources innovative product development, where the community works together with an in-house design team to design products from idea to market. *OpeningDesign* is a platform for architecture and urban planning, encouraging people from different backgrounds to participate in projects and providing a space for opinion polls and crowdsourcing jobs.

These platforms rely on community participation, both amateur and professional, and their websites support community discussion and various amounts of involvement. They attract a range of contributions, from the casual observer who might be motivated to comment once or twice, to the active contributor who closely tracks progress, contributes new ideas, and responds often and with minimal delay. Maher, Paulini, and Murty (2010) show how the nature of the contributions and collaboration can be considered along a spectrum of approaches, ranging from *collected* intelligence to *collective* intelligence. *DesignCrowd* collects individual designs and is an example of *collected* intelligence and *Quirky* is an example of *collective* intelligence in design by encouraging collaboration and voting.

Large scale participation from individuals that may or may not have expertise in the class of products being designed or created can synthesis ideas that go beyond the capability of a single person or a more carefully constructed team. Page (2007) describes how diverse individuals bring different perspectives and heuristics to problem solving, and shows how that diversity can result in better solutions than those produced by a group of like-minded individuals. Hong and Page (2004) prove a theorem that "Diversity Trumps Ability" every time. Page (2007) argues that diversity improves problem solving, even though our individual experiences in working with a diverse group may be associated with the difficulty of understanding other viewpoints and reaching consensus. Many of the successful examples of collective creativity encourage diversity but do not require that everyone understand others' perspectives or even necessarily to reach concensus.

A recent study of communication in Quirky.com shows how the crowd contributes to ideation and evaluation as part of a larger design process (Paulini, Maher, and Murty, 2011). Their analysis shows that a design process that includes crowdsourcing shares processes of ideation and evaluation with individual and team design, and also includes a significant amount of social networking. Collective creativity is an emergent property of an online community where team design is structured and managed intentionally to produce an innovative product.

Who is being creative?

Creativity can be the result of introducing a novel and surprising idea and developing that idea into a product that is valuable in the context of an existing conceptual class of products. A creative idea can originate by bringing a different perspective or set of heuristics (as described by Page (2007)) to a conceptual class or existing patterns of design. This diversity can be achieved through social, computational and collective creativity. The various processes as described by Boden and Gero show how different algorithms or heuristics result in creative ideas. The various approaches to evaluating creativity show how creative ideas can be evaluated. The field of computational creativity now has the basis for developing and evaluating creative systems, and can benefit from characterizing individual contributions to the field. By asking "Who is being creative?" we can identify where the focus of computational creativity is now, and where there are research opportunities. Did the computer generate the creative idea or did a person, or was it an emergent structure from the interaction of people and computational systems? In this section we structure a framework around the concept of human/computer ideation and interaction, and map individual contributions onto a space of possibilities. The contributions include a sample of computational creativity drawing on the proceedings of ICCC 2011 (Ventura et al, 2011) and ICCCX (Ventura et al, 2010), including Quirky (quirky.com) and Scratch (Maloney et al, 2010) to fill gaps in ICCC coverage.

Ideation

Ideation is a process of generating, synthesizing, evaluating, implementing ideas that lead to a potentially creative product or solution. Ideation is a creative process and an idea is a product of that process. Using the term ideation to characterize computational creativity provides a basis for analyzing human, computational, and collective creativity with respect to the origin of a creative idea. While it may be hard to track precisely where an idea comes from in a complex creative process, we can identify where in a human-computer collective there is potential for creative ideas to be expressed and evaluated. Figure 1 places systems that contribute to computational creativity within a space according to the origin of the creative idea as human or computational agent. The "human" and "computational agent" dimensions of this space characterize the role of each in the computational creativity system.

Along the human dimension the framework includes two categories that describe the role of the human in computational creativity: model or generate.

- *Model*: the role of the human is developing a computational model or process. The computational system is effectively being creative because it is the source of the creative ideas or artifact. For example, The Painting Fool is a computational system that generates artistic paintings (Colton, 2011).
- *Generate*: the human generates the creative idea and the computational system facilitates or enhances human creativity by providing information, by providing a digital environment for generating the creative artifact, and/or by providing a perceptual interface to digital content that influences creative cognition. For example, Scratch, is a computational system for people to create interactive stories, animations, games, music, and art (Maloney et al, 2010).

Along the computational dimension the framework includes three categories that describe the role of the computational system: support, enhance, generate.

- *Support*: the computational system supports human creativity by providing tools and techniques. Scratch is an example of a creativity support tool.
- *Enhance*: the computational system extends the ability of the person to be creative by providing knowledge or changing human perception in ways that encourage creative cognition. For example, Scuddle uses a genetic algorithm to generate movement catalysts for dancers (Carlson et al, 2011).
- *Generate*: the computational system generates creative ideas that the human then interprets, evaluates or integrates as a creative product. The Painting Fool is a computational system that generates creative paintings.

Figure 1 shows a distribution of computational systems in which the human generates the creative ideas, aka creativity support tools, and the computational system generates creative ideas. The contributions in the space that is empty in Figure 1 includes theoretical contributions rather than the development of computational systems, for example the contributions to models of process that generate creative products and models for evaluating creative products.

Interaction

Interaction plays an important role in computational creativity, particularly interaction between computers and humans (as the generators or users of the computational system). Traditionally, human-computer interaction has been a one-to-one interaction in which one person interacts with one computational device or environment. Recently, interaction has changed in scale. Figure 2 places the same systems from Figure 1 within a space that characterizes the interaction between people and computers where the dimensions of this space express scale: from a single human or computational system to many.

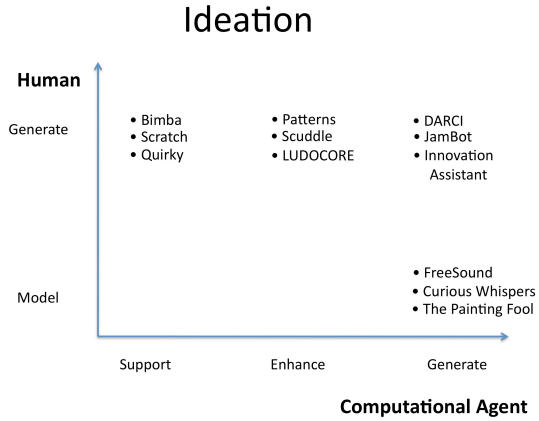


Figure 1. Ideation and Computational Creativity

Along the human dimension there are three categories that describe the scale of the human interaction: individual, group, or public.

- *Individual*: the computational system is developed to support one person working alone, for example Scuddle.
- *Group*: the computational system supports a group or a predefined team of people. This is exemplified by the collaborative technologies that support design and drawing such as Groupboard1. This area is not well represented in the ICCC series.
- *Society*: the computational system encourages crowdsourcing and collective intelligence, for example Quirky.

Along the computational agent dimension there are three categories that describe the scale of the computational agent interaction: individual, team, or multi-agent society.

- *Individual*: there is one computational system with a centralized control that is interacting with a person or people, for example The Painting Fool.
- *Team*: there are multiple, centrally organized agents that interact with one or more people. For example, Curious Whispers is a collection of autonomous mobile robots that communicate through simple songs (Saunders et al, 2010).
- *Multi-agent society*: the computational system is a multi-agent society with distributed control. For example, the designer agents and consumer agents in Gomez de Silva Garza and Gero (2010). This area is not well represented in the ICCC series.

From Figure 2 we see that the contributions in the ICCC series focus on the interaction is between one person and one computational system.

[1] http://www.groupboard.com/products/

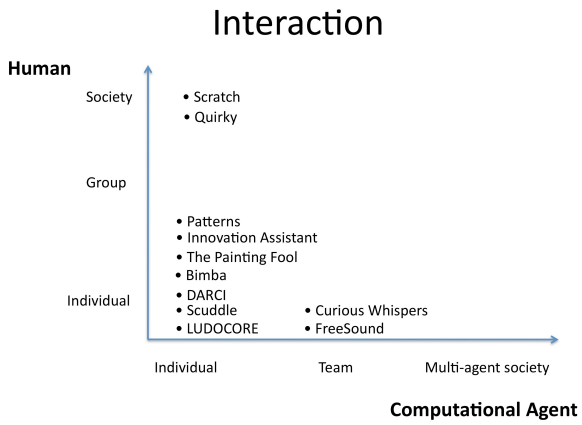


Figure 2. Interaction and Computational Creativity

Conclusions

As we develop a better understanding of processes and products in creative people or systems, we are able to develop more capable computational creativity. Ideation and interaction distinguish research in computational creativity by asking: Who is being creative? The word "who" is used to refer to one or more people or computational systems. When creativity is ascribed to the plural "who", that is when the ideas come from multiple sources, there is an assumption of interaction. An area of research in computational creativity that has received little attention is the role and scale of interaction. Interaction at the scale of one person and one computational system has been the norm in computational creativity, with a recent trend in developing collaborative environments to support or enhance creativity, multi-agent models of creativity and online communities that achieve collective creativity. This paper shows that there is an opportunity for researchers in computational creativity to build on our theoretical and practical advances in understanding creative processes and the evaluation of creative products to address the concepts of interaction and scale.

References

Boden, M. 2003. *The Creative Mind: Myths and Mechanisms*, Routledge; 2 edition

Carlson, K., Schiphorst, T. and Pasquier, P. 2011. Scuddle: Generating Movement Catalysts for Computer-Aided Choreography, *Proceedings of the Second International Conference on Computational Creativity*.

Colton, S. 2011. The Painting Fool: Stories from Building an Automated Painter, in J. McCormack and M. d'Inverno *Computers and Creativity*.

Gomez de Silva Garza, A. and Gero, J.S. 2010. Elementary Social Interactions and Their Effects on Creativity: A Computational Simulation, *International Conference on Computational Creativity*, pp 110-119. http://eden.dei.uc.pt/~amilcar/ftp/e-Proceedings_ICCC-X.pdf

Gero, J.S. 2000. Computational Models of Innovative and Creative Design Processes, *Technological Forecasting and Social Change* 64, 183-196.

Hong, L. and Page, S. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers, *The National Academy of the Sciences*.

Maher, M.L.: 2010. Evaluating Creativity in Humans, Computers, and Collectively Intelligent Systems, *DESIRE'10: Creativity and Innovation in Design*, Aurhus, Denmark.

Maher, M.L., Paulini, M. and Murty, P. 2010. Scaling up: From individual design to collaborative design to collective design, In J S Gero (Ed) *Design Computing and Cognition DCC10*, Springer, 581-600.

Maloney, J., Resnick, M., Rusk, N., Silverman, B., Eastmond, E. 2010. The Scratch Programming Language and Environment. *ACM Transactions on Computing Education* (Nov).

Page, S. 2007. *The Difference: How the Power of Diversity Creates Better Groups*, Princeton University Press.

Paulini, M., Maher, M.L., and Murty, P. 2011. The Role of Collective Intelligence in Design: A protocol study of online design communication, in *Proceedings of the 16th International Conference on Computer-Aided Architectural Design Research in Asia*, 687-696.

Oman, S and Tumer, I. 2009. The Potential of Creativity Metrics for Mechanical Engineering Concept Design in *Proceedings of the 17th International Conference on Engineering Design (ICED'09)*, Vol. 2, eds: Bergendahl, N.; Grimheden, M.; Leifer, L.; Skogstad, P.; Lindemann, U., pp 145-156.

Rhodes M. 1987. An analysis of creativity. In *Frontiers of Creativity Research: Beyond the Basics*, ed: SG Isaksen, Buffalo NY: Bearly, pp 216-222.

Ritchie, G. 2007. Some Empirical Criteria for Attributing Creativity to a Computer Program. *Minds and Machines*, 17(1):67-99.

Runco, M.A. 2007. *Creativity: Theories and Themes: Research, Development and Practice,* Elsevier.

Saunders, R., Gemeinboeck, P., Lombard, A., Bourke, D. and Kocabali, B. 2010. Curious Whispers: An Embodied Artificial Creative System, *International Conference on Computational Creativity*, pp 100-109. http://eden.dei.uc.pt/~amilcar/ftp/e-Proceedings_ICCC-X.pdf

Shah J., Smith S., Vargas-Hernandez N. 2003. Metrics for measuring ideation effectiveness, *Design Studies*, 24(2), 111-134.

Ventura, D., Pease, A., Perez y Perez, R., Ritchie, G., and Veale, T. (eds). 2010. *International Conference on Computational Creativity*. http://eden.dei.uc.pt/~amilcar/ftp/e-Proceedings_ICCC-X.pdf

Ventura, D., Gervas, P., Harrell, F., Maher, M.L., Pease, A., and Wiggins, G.: (eds) 2011. *Proceedings of the Second International Conference on Computational Creativity*. http://iccc11.cua.uam.mx/proceedings/index.html

Wiggins, G. 2006. A Preliminary Framework for Description, Analysis and Comparison of Creative Systems, *Knowledge-Based Systems* 19, 449-458.

A Quantitative Study of Creative Leaps

Lior Noy*, Yuval Hart*, Natalie Andrew*, Omer Ramote, Avi Mayo and Uri Alon
Molecular Cell Biology
Weizmann Institute of Science
Rehovot, Isreal
lior.noy@weizmann.ac.il

Abstract

We present a novel quantitative approach for studying creative leaps. Participants explored the space of shapes composed of ten adjacent squares, searching for 'interesting and beautiful' shapes. By recording players' actions we were able to quantitatively study aspects of their exploration process. In particular our goal is to identify populated sub-regions in the shape space and study the dynamics of 'creative leaps': a jump from one such area to another. We present here the experimental system, our methods of analysis and some preliminary results. We show that the network of shapes created by human participants is different from the class of networks created by applying a simple random-walk algorithm. Chosen shapes show an interesting negative correlation between their abundance and the probability to be chosen as beautiful. We further analyzed the human network unique signature using its *network motifs* profile. Intriguingly, this signature shows similarity to words-adjacency networks extracted from texts. Lastly, we find preliminary evidence that human players exhibit two types of exploration: 'scavenging', where shapes similar in their visual-iconic meaning are quickly accumulated, and 'creative leaps', where players shift to a new region in the shape space after a prolonged search. We plan to build upon this result to quantitatively study creative processes in general and creative leaps in particular.

Introduction

In his book "the Act of Creation" the author Arthur Koestler describes the similarities between three types of creative acts: the pun of the joker, the discovery of the scientist and the lyric expression of the poet (Koestler 1964). The crux of the creative act is the creative leap, the momentary intersection of two different matrices of association (Fig. 1, left). Consider a search resulting in a creative solution for a given problem. Before the creative leap the search is confined to some familiar sub-space (the horizontal plane in Fig. 1, left). Using chance or intuition the solver has managed somehow to reach a point on the plane which also belongs to another plane, a totally different class of solutions (the vertical plane in Fig. 1, left). The creative leap is the ability to recognize this transition point and to jump from one class of solutions to another.

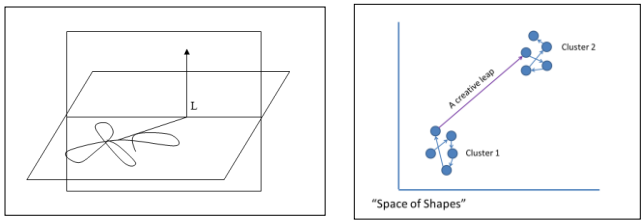


Figure 1. A Symbolic representation of creative leaps. Left: according to Koestler the heart of any creative act is a creative leap between two intersecting domains. Right: a hypothetical creative space. Solutions are grouped into two clusters. Searching within a cluster requires short moves and creates similar solutions. In order to move to a different cluster of solutions the agent needs to perform a creative leap.

Little is known about the dynamics of creative leaps. Previous work has described creative leaps of exceptional creators (Miller 1996) while empirical work has focused mainly on moments of insight in problem solving, such as the Remote Association Test, using both behavioral (Dominowski and Dallob 1995) and brain studies (Sandkühler 2008). It is difficult to capture creative leaps in a laboratory setting. Moreover, many solution spaces might be high-dimensional and complex, with no clear metric defining the similarity between points. For example, consider the space of all answers to the following question used in a group creativity test: "how can the number of tourists visiting your city be increased" (Nijstad and Stroebe 2006). While this problem has solutions that belong to different classes (for example 'increase advertisement' vs. 'improve infrastructure') it is not clear how to define and construct the space of all such ideas.

Our goal is to study a creative task with an underlying solution space that is (a) simple and well defined to enable a quantitative investigation of the search dynamic (b) that contains clusters of solutions, with the possibility of performing creative leaps between them (see Fig. 1, right). Our approach resembles recent work by Jennings that similarly studied people's search trajectories in a visual domain (Jennings 2010; Jennings et al. 2011).

We searched for a parameterized space that will be complex enough to allow for possible creative leaps, but not too complex to allow a computational description of human search in this space. We suggest using the set of all N-

size *polyominoes* – the set of two dimensional shapes composed of N adjacent squares (Golomb 1994).

Besides its well defined structure which allows for establishing a metric on the search space, the polyominoes space provides a crucial advantage: the shape space exploration complexity is tunable by changing the parameter N. We can thus aim to have an exploration process which is on one hand not too trivial and on the other hand not too complex to quantify. In that we hope to capture the gist of what Boden describes as 'an exploratory frame of mind' (Boden, 2004). Since this exploration process resembles a creative process undertaken by, say, a graphic designer designing a new icon in a limited space, we hope to gain insights in the growing field of computational models for design processes (Gero, 2000).

We analyzed the network of shapes and moves created by human participants and compared the human exploration with a simple random-walk algorithm that transverses the network of shapes discovered by the human participants. This comparison shows that the human search behavior is not simply the results of a random travel between the shapes. Our results suggest that humans perform two types of searches: 'scavenging', a simple search in an area of shapes, which can be explained by an algorithmic search, and 'insight' moves, or leaps, that cannot be explained by simple algorithm. The first type of moves corresponds to the within cluster exploration in Fig. 1, while the second type contains, we hope, the creative leaps.

We next describe our experimental setup, the methods of analysis we employed and some initial findings supporting the notion that creative leaps can be quantitatively studied using the suggested approach.

Experimental Setup

System

We developed a system to experimentally test human trajectories in the shape space of polyominoes. We are currently experimenting with *decominoes*, 10-size polyominoes (consisting of 4655 unique shapes and 36,446 shapes if rotations and mirror images are counted).

We tested several variants of the creative task and report here results from the 'journey in shape space': exploring the space by moving one square at a time, transforming one legitimate shape to another. The starting point shape is always the horizontal line. We ask people to "explore the space of 'shifting shapes' and to discover shapes that you find interesting and beautiful".

We developed an experimental setup using Processing, an open source, cross-platform, programming language used for visualization (see Fig. 2).

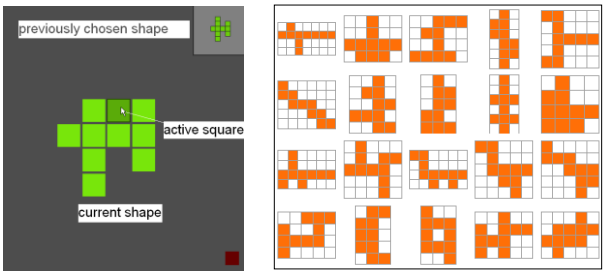


Figure 2. Exploring the space of shapes. Left: a screen shot of the 'Shape Shifter' game. At each step players move one square to create a new polyomino. Shapes can be stored in the 'shape gallery' by pressing the gray rectangle at the top-right corner. Right: examples of different shapes created by human players.

Procedure

123 participants (58 females and 65 males, ages 12-75 years, mean = 34.3), recruited through emails and social networks, were invited to participate in a short experiment in creativity. At any point players could store the current shape to a 'shape gallery'. The players moved freely between shapes, within a time limit of 25 minutes (no participant reached this limit). When choosing to finish the exploration they continued to the 'rating stage'. In this last stage players observed the 'shapes gallery' and were asked to choose 'the five most creative shapes you discovered'. We recorded square moves between shapes and their timing, as well as each player chosen gallery shapes and the final five shapes.

Analysis

A random-walk algorithm over the entire shape network

We used a network representation (a graph) of the shape space in the following way. Each shape is a node in the graph. Shape A and B are connected by an edge if shape A can be reached from shape B by moving a single square in a valid way. This structure is a directed graph representing all possible valid moves

The algorithm explores the network by first randomly removing one square from the current shape. The next decomino in the path is then generated by placing the 10th square in a new random location (self-loops are not excluded). This extends the path by one step. The path is further extended by repeating these steps up to a predetermined steps number.

This algorithm was used to establish both the entire shape space network and a random walk generated network to compare with the human generated network of travelled shapes. For the entire shape space the algorithm was run until all possible 36,446 decominoes were generated (with mean path length of 150,000 steps). For comparison with the human network, the algorithm was run 123 times (the number of human participants) with a number of steps which is sampled from the number of steps distribution of the human players.

A random-walk algorithm over the human generated network

In order to create computer generated networks which are more closely related to the human networks we restricted the algorithm to travel only on edges which were travelled by at least one human player.

First the human generated decominoes network is generated and the allowed steps are listed. Although the network is naturally directed, the computerized walker is allowed to move on the undirected network (that is, the computer can also move backward on any human edge).

The algorithm is seeded and a new shape is chosen randomly from the set of shapes which are connected by allowed edges. The length of the path is sampled from the distribution of lengths of paths traversed by the human players. This process is repeated 123 times.

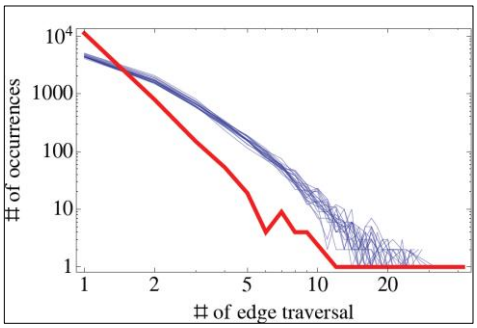


Figure 3. Comparing human and computational exploration networks. The number of occurrences of edges where edges are grouped by the number of times they were traversed. Shown are the values for human players' network (red) and the random walk network restricted to the human network shapes (mean of 10 simulation in dark blue, each specific simulation in light blue).

Our current goal is to compare the features of the human generated network to a network generated by a random-walk algorithm and to study if there is a noticeable difference between the two, in order to show that the human behavior cannot be explained as a result of a random-walk in the shape-space.

Triad Significance Profile Calculation

The 13 network motif frequencies of the human and random generated networks were calculated. The normalized Z score of each of the 13 possible triads was then calculated. Z score is computed by the difference of the triad frequency to the mean frequency of the same triad in a computerized agents' network, measured in STD units. Frequency mean and STD were calculated from 10 simulations of the computational networks.

Results

Human and Computational Networks

We first asked whether the exploration network created by human players is different from the network created by a random-walk algorithm traveling the entire shape networks. We find that the exploration network created by human players is much more compact. Furthermore, the players' network obeys a power-law distribution of node degree frequencies (how many edges go in or out from a specific node), while the computational algorithm produces a Gaussian-like distribution of node degree frequencies. In addition, human exploration on the network of all allowed edges is very constrained and compact relative to a random exploration process of the whole shapes space.

We next asked whether the type of exploration players perform is dictated only by some constraint on shapes available to people's perception. We thus compared the human exploration network with an ensemble of networks created by allowing a random-walk algorithm to choose shapes randomly, but restricting it to shapes that were selected by the human players. We find that the algorithm travels much less than the human players and so create a much smaller network than the players' network. Furthermore, the properties of the computational exploration networks, such as the distribution of nodes degrees is markedly different from the human exploration network (Fig. 3).

Consensus in Participants' Choices

A possible concern regarding our creative task is whether there is some consensus among different participants regarding their aesthetic choices. While we do not expect to have total agreement – for example some players preferred iconic shapes, while other preferred more abstract ones, a total lack of consensus could raise doubts on the validity of this task to measure human creativity.

To assess the consensus in participants' choices we plotted the *selection ratio*, the percentage of times a shape was chosen (number of times chosen divided by number of times traversed) against the number of times a shape was traversed (Fig. 4). We differentiated between shapes ranked as interesting shapes in the last stage of the game (in blue) and those that were only chosen to the gallery (in red).

We note that there is a large number of shapes with high (>50%) selection ratio, with few shapes exhibiting selection ratio of more than 90%. At least for these shapes there seems to be a consensus among the different human participants. In addition, shapes that were ranked in the last stage had a statistically significant higher selection ratio (ranked: centered around (23.34, 50) with STD (19.41, 20); not-ranked: centered around (15.6, 20) with STD (6.7, 13); non-paired *t*-test, $p<10^{-7}$).

We also note the negative correlation (Pearson correlation = -0.25, $p<0.05$) between the prevalence of a shape (how many times it was traversed) and its selection ratio. Intriguingly, this might suggest that shapes 'less traveled by' are appreciated more by the people who have reached them.

A Network Motifs Signature

In order to further characterize the human exploration network we measured its network motifs signature, termed

triad significance profile (TSP). This network signature is calculated by taking the frequencies of all three node subgroups of a network and normalizing each frequency by the triad frequency in a network created by a similar random process (Milo 2002). In our case, we compared triad frequencies of human network with triad frequencies created by the random walk algorithm on the human network (see Analysis). Previous studies in our lab showed that networks with similar structure and function have a similar TSP signature. Thus, this method offers another quantitative classification to networks.

This preliminary calculation (Fig. 5) indicates that the network motifs significance profile shares a similar frequency signature of text networks (Milo 2004), suggesting that the human visual exploration process in shape space consists of visual rules similar to those of language networks, having categories of words with a certain formulated way of combining between different categories. Future work should check the dependency of the calculated triad significance profile on the randomization process used to create the base-line random network.

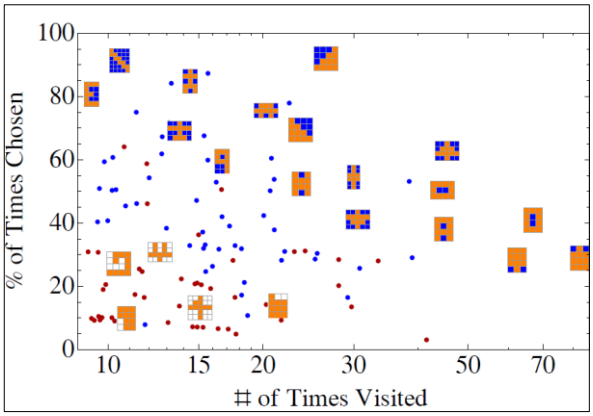


Figure 4. Consensus in participants' choice of shapes. Y-axis: the number of times a shape was admitted into the gallery out of the number of times it was visited. X-axis: the number of times a shape was visited. Only shapes that were visited at least 10 times are presented. Dots in blue represent shapes that were also ranked in the final stage while red dots represent shapes that were chosen to the gallery but were not ranked. Correspondingly, shapes shaded in blue are representative of the set of finally chosen shapes.

Initial Evidence for Creative Leaps

In order to more closely examine the exploration process of individual players, we focused on the 'chosen to the gallery' shapes (Fig. 6), enumerating both the number of steps between two sequential shapes (the number above each shape) and the time interval between selections of the two shapes (the y axis). For several players we observe an interesting pattern: the time and number of steps between two sequential chosen shapes is declining at the beginning, usually creating similar content shapes. Then, a long traversal exploration process is commenced, usually leading to shapes belonging to a new cluster of similar shapes. As exemplified in Fig. 6, the player moves from "Animals" shapes to "Space invaders" shape to "Symbolic male/female" shapes. One can interpret this saw-tooth pattern as consisting of scavenger explorations connected by a creative leap, which serves to reach a new iconographic domain.

We hope to utilize these processes to cluster the shapes automatically into different domains and thus create a semi-metric on the shape space. Another utility to aid building the metric comes from the use of the rating process at the end of the game. Subjects are requested to choose the five most creative shapes. Our assumption is that subjects will choose shapes that they see as most distinct from one another, thus providing another metric measure on the shape space.

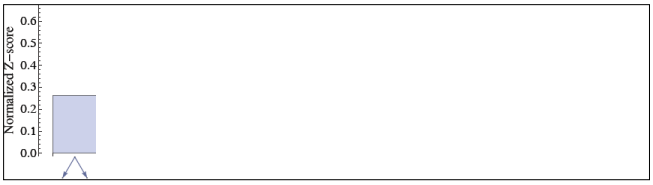


Figure 5. The triad significance profile (TSP) of the human players' network suggests a similarity to word-adjacency networks of texts. The main feature of the TSP is the under-representation of triangle-shaped triads 7 to 13.

Conclusions and Future Work

We presented a novel quantitative approach for studying creative leaps. Our goal is to study a creative task using computational tools. Specifically we aim to define the space of products of the creative task, to detect clusters of similar products and to study creative leaps between them.

Working toward this goal we developed a web-based game in which players explored a visual space composed of 10-size polyominoes, while searching for interesting and beautiful shapes. As a first step we tested whether human behavior in this task can be explained as a result of a random-walk algorithm. We therefore compared the exploration network created by human players to two computational exploration networks. The first network was created by random walks on all possible shapes, and the second one was created by random walks restricted to shapes chosen by human players. We compared general properties of these networks, such as in/out degree, and found a significant difference between the human and the computational networks. Compared to a network made by a random walk on shapes chosen by players, the computer's random walk is much smaller, suggesting that the trajectories of human exploration contain also segments of directed motion toward interesting regions of the space. Following the *fogginess* metaphor of Jennings (2011) these segments might correspond to the areas of the landscape with have 'good visibility'.

We also used the concept of network motifs to characterize the human search network. We identified which of a known super-families of networks (e.g. social, transcrip-

tion networks, and language originated), matches the human exploration network. We find that the human network is similar to language-originated network, and are planning to further study the connection between these two networks.

We further find preliminary evidence of players' paradigm shift while playing the game. Players show periods of 'scavenging', where they exploit shapes similar in iconic meaning (e.g. animals, letter, symmetric shapes) accompanied by long walks on the grid of possible shapes, which leads to a different region in the shapes space. The 'saw-tooth' pattern we have found in the time between chosen shapes (Fig. 6) might be the first clue for the existence of clusters in our shape space. We plan to corroborate these finding by different methods that can be used to detect clusters of shapes in this visual domain. In particular, we plan to use the human choices embedded in our task at multiple levels (which shape to move to; which shapes to insert to the gallery; which shapes to choose in the final stage) as a different probe into the structure of the shape space.

This paper presents work-in-progress aiming to develop a computational platform for studying human search in creative tasks, and in particular to study creative leaps. We are currently performing a large-scale human experiment with this platform and plan to apply a host of quantitative methods to further test the preliminary results presented here. Using these methods we hope to be able to measure and study the dynamics of creative leaps.

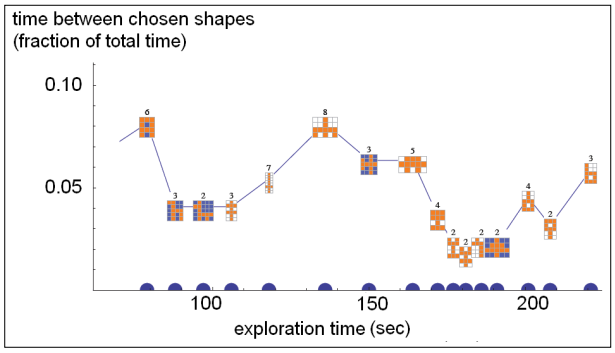


Fig 6. Preliminary evidence for clusters in the shape space. Looking at the time differences between chose shapes we often see 'saw-tooth' patterns. Humans seem to reach a fruitful region, 'scavenge' it, that is, to quickly pick a few similar shapes, and then to move to another region, a move that takes much more time. Notice for example the two clusters of similar shapes around 100 and 180 seconds. Only chosen shapes are shown, and shapes in the 'top five' (chosen between all gallery shapes) appear with a blue background. The number above each shape is the number of moves from the previous shape.

References

Boden, M.A. 2007. Creativity in a nutshell. *Think* 5 (15): 83.

Dominowski, R.L. and Dallob, P. 1995. Insight and problem solving. In: Sternberg, R. J.; Davidson J. E. (Eyds). The nature of insight, pp. 33–62. MIT Press, Cambridge, MA.

Gero, J.S. 2000. Computational models of innovative and creative design processes. *Technological forecasting and social change* 64 (2): 183.

Golomb, S. W. 1994. *Polyominoes* (2nd ed.). Princeton University Press, Princeton, NJ.

Jennings, K.E. 2010.Search strategies and the creative process, Proceedings First International Conference on Computational Creativity, 130

Jennings, K.E.; Simonton, D.K. and Palmer, S.E. 2011. Understanding exploratory creativity in a visual domain, *Proceedings of the 8th ACM conference on Creativity and cognition*, 223.

Koestler, A. 1964. *The Act of Creation*. Penguin Books, NY.

Miller, A. I. 2000. Insights of genius: Imagery and creativity in science and art. MIT Press, Cambridge, MA.

Milo, R.; Shen-Orr, S.; Itzkovitz, S.; Kashtan, N.; Chklovskii, D.; and Alon, U. 2002. Network motifs: simple building blocks of complex networks. *Science* 298 (5594): 824.

Milo, R.; Itzkovitz, S.; Kashtan, N.; Levitt, R.; Shen-Orr, S.; Ayzenshtat, I.; Sheffer, M.; and Alon, U. 2004. Superfamilies of evolved and designed networks. *Science* 303 (5663): 1538.

Sandkühler, S. and Bhattacharya, J. 2008. Deconstructing insight: EEG correlates of insightful problem solving. *PLoS One* 3 (1): e1459.

On the Notion of Framing in Computational Creativity

John Charnley, Alison Pease and Simon Colton
Computational Creativity Group
Department of Computing, Imperial College
180 Queens Gate, London SW7 2RH, United Kingdom.
ccg.doc.ic.ac.uk

Abstract

In most domains, artefacts and the creativity that went into their production is judged within a context; where a context may include background information on how the creator feels about their work, what they think it expresses, how it fits in with other work done within their community, their mood before, during and after creation, and so on. We identify areas of framing information, such as motivation, intention, or the processes involved in creating a work, and consider how these areas might be applicable to the context of Computational Creativity. We suggest examples of how such framing information may be derived in existing creative systems and propose a novel dually-creative approach to framing, whereby an automated story generation system is employed, in tandem with the artefact generator, to produce suitable framing information. We outline how this method might be developed and some longer term goals.

Introduction

Michael Craig-Martin's 1973 work, *An Oak Tree*, comprises a glass of water on a shelf and an accompanying text, in which Craig-Martin claims that the object which appears to be a glass of water is really an oak tree. The text takes the form of a question and answer session written by Craig-Martin about how he has changed the water into a tree:

A. [.....] I've changed the physical substance of the glass of water into that of an oak tree.

Q. It looks like a glass of water.

A. Of course it does. I didn't change its appearance. But it's not a glass of water, it's an oak tree.

...

Q. Haven't you simply called this glass of water an oak tree?

A. Absolutely not.

Craig-Martin is rather mysterious as to how he has accomplished the change:

Q. Was it difficult to effect the change?

A. No effort at all. But it took me years of work before I realised I could do it.

Q. When precisely did the glass of water become an oak tree?

A. When I put the water in the glass.

Q. Does this happen every time you fill a glass with water?

A. No, of course not. Only when I intend to change it into an oak tree.

The status of the piece as a work of art is then raised:

Q. Do you consider that changing the glass of water into an oak tree constitutes an art work?

A. Yes.

Q. What precisely is the art work? The glass of water?

A. There is no glass of water anymore.

Q. The process of change?

A. There is no process involved in the change.

Q. The oak tree?

A. Yes. The oak tree.

This is an example of human creativity which is taken seriously in its field. First shown in 1974, it was bought by the National Gallery of Australia in Canberra in 1977, and has been exhibited all over the world with the text translated into at least twenty languages. Like many important works, opinion is divided: artist Michael Daley referred to it as "self-deluding" and "pretentious" (Daley August 31 2002), while art critic Richard Cork wrote:

> "I realise that one of the most challenging moments occurred in 1974 when the Rowan Gallery mounted an exhibition of Michael Craig-Martin's work." (Cork October 9 2006).

Researchers in Computational Creativity (CC) can learn from this work. The main point that we consider in this paper is that *the artefact* (the glass of water) *has no creative value without the title and accompanying text*. The value, the creativity associated with the piece, lies in the narrative surrounding the glass of water. This point has clear implications for CC, which has traditionally focused on artefact generation, to the extent that the degree of creativity judged to be in the system is often considered to be entirely dependent on characteristics of the set of artefacts it produces (for instance, see (Ritchie 2007)). Very few systems in CC currently generate their own narrative, or framing information. Artefacts are judged either in isolation or in conjunction with a human-produced narrative, such as the name of the system and any scientific papers which describe how it works. Researchers in CC will be familiar with what Bedworth and Norwood call "carbon fascism" (Bedworth and Norwood 1999) (the bias that only biological creativity can produce valuable artefacts), and, for the most part, computer-generated creative artefacts are not taken seriously

by experts in the domain in which the artefacts belong. We believe that enabling creative software to produce its own framing information will help to gain acceptance from these experts. While *An Oak Tree* may be a rather extreme example of the importance of framing information, we hold that such information almost always plays some role in creative acts, and is a fundamental aspect of human creativity. We consider here which types of framing information we could feasibly expect a piece of software to produce, and begin to propose ways in which we could formalise this. Specifically, we consider three areas in computational terms: *motivation* (why did you do X?), *intention* (what did you mean when you did X?), and *processes* (how did you do X?). We make the following contributions:

1. We highlight the importance of framing information in human creativity.
2. We propose an approach to automatically generating framing information, in which a separate creative act of automated story generation is performed alongside traditional artefact generation.

Framing in human creativity

Sir John Tusa is a British arts administrator, radio and television journalist, known for his BBC Radio 3 series *The John Tusa Interview*, in which he interviews contemporary artists. These interviews have been reproduced as two books in which he explores the processes of creativity (Tusa 2003; 2006). We have analysed all thirteen interviews in his most recent collection, in order to provide the starting point for a taxonomy of framing information. In the following discussion, unless otherwise specified, all page numbers refer to this collection of interviews (Tusa 2006). We identified two categories which artists spoke about: INTERNAL, or inward looking, in which the artist talks about their own *Work*, *Career* and *Life*, and EXTERNAL, or outward looking, in which the artist talks about their view of their *Audience* and *Field*.

The artist's work

Discussion about artists' work is very common in the Tusa interviews. This might concern a specific piece; such as how an artist feels about a piece, what they think it expresses, or how it relates to everyday concepts; or this might concern details of the generative process, such as how the work is created, how processes involved in its creation fit together, or whether a new technique or material changed the way that something was done. As an example below, Cunningham (MC) relates his work to scientific and religious concepts:

> *MC:* ...it was the statement of Einstein's which I read at that time, where he said, 'There are no fixed points in space.' And it was like a flash of lightning; I felt, Well, that's marvellous for the stage. Instead of thinking it's front and centre, to allow any point, very Buddhist, any point in the space to be as important as any other. (p. 66)

The artist's career

A picture of the structure of an artist's career, in terms of his or her past, present and possible future directions, can aid understanding of current work. Questions about previous work include asking how two pieces differ; what category work from certain periods falls into; classification of a career into different stages, such as early and late, or pre-work X and post-work X.

Examples from (Tusa 2006) include the questions: "So you think you are recognisably the same person, creatively the same person as you would have been if you'd stayed in New York?" (to Forsythe, p. 93); "What's the next stage of your evolution as a maker of ballets?" (also to Forsythe, p. 105), and "When you look back over the last twenty years, would you ever have guessed that the work that you do would have travelled so far ..... I mean this is an extraordinary journey. How aware have you been of the evolution as you've been through it?" (to McBurney, pp. 181-2).

The artist's life

Audiences are interested in the personalities and influences behind society's "creative heroes". Topics of interest include political, intellectual, personal, cultural and religious influences; value systems; reasons for working in a particular area; important events in the life of the artist, and so on.

John Tusa asks many questions in this vein. For instance, he asks: "Are you an optimist or a pessimist as a person?" (to Rovner); "When did you discover that you had this condition called Dysgraphia, where I think the brain wants to write words as pictures?" (to Viola); "What do you feel, as you're coming in to work?" (to McBurney); and "What music do you like?" (to Piano), and has some rather poignant exchanges, such as one with Viola in which he asks about a near-death experience (pp. 221-3); and this exchange with Rovner (MR):

> *JT:* Are you lonely as an artist?
>
> *MR:* You mean as an artist or as a person?
>
> *JT:* Well as a person who is an artist.
>
> *MR:* I'm alone. I don't know if I'm lonely. I am single, you know I'm a single person, I'm a single person. (p. 213)

The artist's view of their audience

The perception that an artist has of his or her audience may influence their work. Queries in this topic included questions about effect of a particular field on audiences, and what the effect of certain pieces of work are on the collective subconscious. Egoyan, for instance, discusses responsibility to one's audience with Tusa (p75).

The artist's view of the field

Embedding a particular artist's work into the context of a body of work is one of the purposes of framing information. Queries include definitional questions about particular fields, and their relationship to other fields; how a piece fits into a field; in which field an artist sees themselves; the influence of external characteristics such as politics, or how modern advancements such as new techniques have affected a field; the history of a field and directions in which it could go, and so on. For instance, Egoyan discusses how he thinks video compares to film (p76), and Forsythe talks about how his work fits into great classical ballets (p. 106).

Framing for Computational Creativity

Analysis of the interview responses suggests a new direction for CC: *enabling creative software to generate some of its own framing information*. As with human artworks, the appeal of computer creativity will be enhanced by the presence of framing. However, there are obvious restrictions on the scope to which the various forms of framing apply in the computer generated context. Here we consider three areas in computational terms: motivation (why did you do X?), intention (what did you mean when you did X?), and processes (how did you do X?).

Motivation

Many creative systems currently rely upon human intervention to begin, or guide, a creative session and the extent to which the systems themselves act autonomously varies widely. In some sense, the level to which these systems could be considered self-motivating is inversely proportional to the amount of guidance they receive. However, it is possible to foresee situations where this reliance has been removed to such an extent – and the human input rendered so remote – that it is considered inconsequential to the creative process. For instance, the field of Genetic Programming (Koza 1992) has resulted in software which can, itself, develop software. In the CC domain, software may eventually produce its own creative software which, in turn, produces further creative software, and so forth. In such a scenario, there could be several generations in an overall geneology of creative software. As the distance between the original human creator and the software that directly creates the artefact increases, the notion of self-motivation becomes blurred.

Beyond this, the scope for a system's motivation towards a particular generative act is broad. For example, a suitably configured system may be able to perform creative acts in numerous fields and be able to muster its effort in directions of its own choosing. With this in mind, we can make a distinction between *motivation to perform creative acts in general*, *motivation to create in a particular field* and *motivation to create specific instances*.

In the human context, the motivation towards a specific field may be variously influenced by the life of the artist, their career and their attitudes, in particular towards their field and audience. Several of these are distinctly human in nature and it currently makes limited sense to speak of the *life* or *attitudes* of software in any real sense. By contrast, we *can* speak of the *career* of a software artist, as in the corpus of its previous output. This may be used as part of a process by which a computer system decides which area to operate within. For example, we can imagine software that chooses its field of operation based upon how successful it has previously been in that area. For instance, it could refer to external assessments of its historic output to rate how well-received it has been, focusing its future effort accordingly.

The fact that a computer has no *life* from which to draw motivation does not preclude its use as part of framing information. All those aspects missing from a computer could, alternatively, be simulated. For example, we have seen music software that aims to exhibit characteristics of well-known composers in attempts to capture their compositional style (Cope 2006). The extent to which the simulation of human motivation enhances the appeal of computer generated artefacts is, however, still unquantified. The motivation of a software creator may come from a bespoke process which has no basis in how humans are motivated. The details of such a process, and how it is executed for a given instance, would form valid framing information, specific to that software approach.

Intention

The aims for a particular piece are closely related to motivation, described above. A human creator will often undertake an endeavour because of a desire to achieve a particular outcome. Factors such as attitudes to the field contribute to this desire. Certainly, by the fact that some output is produced, every computer generative act displays intent. The aims of the process exist and they can, therefore, be described as part of the framing. In the context of a computer generative act, we might distinguish between *a priori* intent and intentions that arise as part of the generative process. That is, the software may be pre-configured to achieve a particular goal although with some discretion regarding details of the final outcome, which will be decided during the generative process. The details of the underlying intent will depend upon the creative process applied. For example, as above, software creators might simulate aspects of human intent.

Intent has been investigated in collage-generation systems (Krzeczkowska et al. 2010). Here, the software based its collage upon events from the news of that day with the aim of inviting the audience to consider the artwork in the context of the wider world around them. This method was later generalised to consider wider combinations of creative systems and more-closely analyse the point in the creative process at which intentionality arose (Cook and Colton 2011).

Processes

In an act of human creativity, information about the creative process may be lost due to human fallibility, memory, awareness, and so on. However, in a computational context there is an inherent ability to perfectly store and retrieve information. The majority of creative systems would have the ability to produce an audit trail, indicating the results of key decisions in the generative process. For example, an evolutionary art system might be able to provide details of the ancestry of a finished piece, showing each of the generations in between. The extent to which the generative process can be fully recounted in CC is, nevertheless, limited by the ability to fully recreate the sources of information that played into the generative process. Software may, for instance, use information from a dynamic data source in producing an artefact, and it may not be possible to recreate the whole of this source in retrospect.

One system that produces its own framing is an automated poetry generator currently being developed (Colton, Goodwin, and Veale 2012). In addition to creating a poem, this system produces text which describes particular aspects of

its poetry that it found appealing and aspects of how it generated its output. In order to fully engage with a human audience, creative systems will need to adopt some or all of the creative responsibility in generating framing information.

Details of the creative process are valid aspects of framing information, which are relevant to both computational and human creative contexts. There is a notion of an appropriate level of detail: extensive detail may be dull and the appreciation of artefacts is sometimes enhanced by the absence of information about the generative process.

Examples of framing for Computational Creativity

There are many ways in which creative systems might generate their own framing information. For example, an automated art system, such as AARON (McCorduck 1991), could store details of all its previous artworks and provide an assessment of how a new piece differs, in various respects, from its past output. A poetry system, such as (Colton, Goodwin, and Veale 2012), might reveal the general mood of the inspiring source it used as a basis for an affective poem. Mathematical software, such as HR (Colton 2002), could be given the ability to compare the conjectures it finds against on-line mathematical databases and report on how its output relates to known theorems. By corollary, an art system could appeal to image databases to suggest similarities to other artists. A simple enhancement to the collage generation program of (Krzeczkowska et al. 2010) could see it provide the text of the news story that formed the inspiration for the collage. In this mode, the framing information would become as important an aspect of the overall presentation as the collage itself. The artwork would be a combination of both the collage and underlying story, rather than the collage alone. This list is by no means exhaustive. The varied nature of framing information that we have been describing shows that the opportunities for enhancing works with framing are extensive.

A dually-creative approach to framing

Framing information has the potential to greatly impact an audience's assessment of an artefact. In some instances, framing is arguably as much a part of the overall creative presentation as the artefact itself: this was seen in Craig-Martin's *An Oak Tree*, described above, as well as, for example, elements of Marcel Duchamp's *readymades* series such as *Fountain*. The information can be as simple as a title for the artefact, or might encompass much of the type of framing indicated in our analysis. Framing can add to the mystique and mystery surrounding an artefact, as we have described.

Framing information need not be factually accurate. Information surrounding human creativity can be lost, deliberately falsified or made vague for artistic impact. Thus, the generation of framing information can itself be seen as a creative act. The overall impact of the *package* – namely the artefact and the associated framing information – will depend on both the assessed quality of the artefact, together with the impression given by the framing information. We propose one approach to artefact-with-framing generation, where the two are produced simultaneously, by a dually-creative process. Under this approach, the most appropriate creative paradigm for the framing information would be a form of automated storytelling. One part of a combined system would create the artefact itself and a storytelling aspect would generate a framing story. The framing story could be as simple or as complex as those which accompany human creations. Tools which were able to perform tasks such as metaphor and analogy (see (Gentner, Holyoak, and Kokinov 2001; Gibbs Jr. 2008)) might be integrated into the storytelling aspect.

In the previous section, we discussed aspects of framing which might be relevant to the CC setting. This information could form much of the input to the story generation system, becoming part of the basis of the story. For example, purely factual information about how the software arrived at the final product could be retained. Given that there is no requirement for the framing story to be factually correct, some or all of the story might be fictional and there is no prescription for the extent to which the framing story should directly correspond with the artefact. Consider, for example, a framing story which describes all aspects of the creative process in full detail compared with a framing story consisting of a random seemingly-unrelated word. Both have artistic value, but in entirely different ways.

An initial approach to the fact that no configuration of a particular automated story generation system would be able to generate the variety of framing stories that we have witnessed in human creativity, might be to develop a small number of story-telling paradigms, each based upon a particular story template. One challenge might then be to achieve an appropriate balance between fact and fiction in the generated stories. In future, we might hand such decisions over the software. For example, a sufficiently-able software suite might decide which story-telling paradigm is most appropriate for a particular effect, the balance between fact and fiction and how extensive the framing should be. In a more complex manifestation, the story might form an interactive dialogue, providing answers to audience queries in a manner akin to an interview. As with human creativity, the answers to those questions may be entirely at the whim of the generating system. Going further, software might employ story generation approaches to simulate aspects of the framing information which might otherwise be absent, such as a religious belief or other motivation. This could, in turn, feed back into the generation of the creative artefact itself. Storytelling for framing information represents an interesting challenge for our existing and future automated story generation systems.

Related work

In (Colton, Pease, and Charnley 2011; Pease and Colton 2011), two generalisations were introduced with the aim of enabling more precise discussion of the kinds of behaviour exhibited by software when undertaking creative tasks. The first generalisation places the notion of a generative act, wherein an artefact such as a theorem, melody, artwork or poem is produced, into the broader notion of a *creative act*. During a creative act, multiple types of generative acts are

undertaken which might produce framing information, F, aesthetic considerations, A, concepts, C, and exemplars, E; in addition to generative acts which lead to the invention of novel generative processes for the invention of information of types F, A, C and/or E.

The second generalisation places the notion of assessment of the aesthetic and/or utilitarian value of a generated artefact into the broader notion of the impact of a creative act, X. In particular, an assumption was introduced that in assessing the artefacts resulting from a creative act, we actually celebrate the entire creative act, which naturally includes information about the methods underlying the generation of the new material, and the framing information, which may put X into various contexts or explain motivations, etc., generally adding value to the generated artefacts over and above their intrinsic value.

The introduction of these two generalisations enabled the FACE and IDEA descriptive models to be introduced as the first in the fledgling formalisation known as *Computational Creativity Theory*. In this paper we have extended this model by exploring the notion of *framing*.

Future work and conclusions

Creativity is not performed in a vacuum and the human context gives an artefact meaning and value. Implicit in the Computational Creativity Theory models so far developed is the notion that the FACE information/artefacts resulting from creative acts can be seen as invitations to a dialogue. For instance, when a person appreciates a painting, they are encouraged to ask questions of it, and look for answers, either explicitly from the artist or some perceived notion of how artists work, via visual interrogation of the piece itself, or through certain cultural contexts; for example, by understanding the culture in the time and place when the painting was produced.

Despite the importance of framing information as part of the overall artistic endeavour, we are only aware of a very small number of systems that generate framing information to accompany their creative output. We have proposed one approach to this, whereby automated story generation is used to generate framing information. There are no real bounds to what information such framing can contain, its basis in fact versus fiction, or the format in which it is presented. Consequently, we suggest that initial attempts be restricted to a small number of simplified paradigms, taking their basis from a more complete investigation into how humanly-produced framing information relates to CC. Expanding upon this starting point, we imagine software taking over some of the creative responsibility for the framing information, such as determining the story-telling paradigm and the story's emphasis or level of detail.

Craig-Martin, via his narrative of *An Oak Tree*, opens up a dialogue with the viewer on the nature of essence, proof, faith, matter, reality, art, and so on. The viewer engages with this narrative, which includes the manner of presentation of the piece, Craig-Martin's background as an artist and a person, critics' and artists' responses to the piece, stories surrounding the work and effects that it has on everyday life (for instance, there is a myth that Australian customs officials barred it from entering the country since it was classified as "vegetation", and in February 2012 the first three hits from google images on the search term "an oak tree" are images of Craig-Martin's work). We anticipate that the direction outlined in this paper will form an important axis of development for CC systems. Our long-term goal is to help to develop CC to such an extent that one day a piece of creative software will appear in the table of contents of a collection of Tusa-style interviews, to discuss its work and itself, alongside other contemporary artists.

Acknowledgements

This work has been funded by EPSRC grant EP/J004049. We are grateful to the three reviewers, who raised interesting points.

References

Bedworth, J., and Norwood, J. 1999. The Turing test is dead. In *Proceedings of the 3rd conference on creativity and cognition*.

Colton, S.; Goodwin, J.; and Veale, T. 2012. Full face poetry generation. In *Proceedings of the Third International Conference on Computational Creativity*.

Colton, S.; Pease, A.; and Charnley, J. 2011. Computational creativity theory: The FACE and IDEA descriptive models. In *Proceedings of the 2nd International Conference on Computational Creativity*.

Colton, S. 2002. *Automated Theory Formation in Pure Mathematics*. Springer-Verlag.

Cook, M., and Colton, S. 2011. Automated collage generation – with more intent. In *Proceedings of the Second International Conference on Computational Creativity*.

Cope, D. 2006. *Computer Models of Musical Creativity*. Cambridge, MA: MIT Press.

Cork, R. October 9, 2006. Losing our vision. *New Statesman*.

Daley, M. August 31, 2002. Tracey left on the shelf. *The Guardian*.

Gentner, D.; Holyoak, K.; and Kokinov, B. 2001. *The Analogical Mind: Perspectives from Cognitive Science*. Cambridge, MA: MIT Press.

Gibbs Jr., R. W., ed. 2008. Cambridge, UK: Cambridge University Press.

Koza, J. R. 1992. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA, USA: MIT Press.

Krzeczkowska, A.; El-Hage, J.; Colton, S.; and Clark, S. 2010. Automated collage generation – with intent. In *Proceedings of the 1st International Conference on Computational Creativity*.

McCorduck, P. 1991. *AARON's Code: Meta-Art, Artificial Intelligence, and the Work of Harold Cohen*. W.H. Freeman and Company.

Pease, A., and Colton, S. 2011. Computational creativity theory: Inspirations behind the FACE and the IDEA models. In *Proceedings of the 2nd International Conference on Computational Creativity*.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17:67–99.

Tusa, J. 2003. *On Creativity*. London: Methuen.

Tusa, J. 2006. *The Janus aspect: artists in the twenty-first century*. London: Methuen.

Small-Scale Systems and Computational Creativity

Nick Montfort and Natalia Fedorova
Program in Writing & Humanistic Studies
Massachusetts Institute of Technology
77 Massachusetts Ave, 14N-233
Cambridge, MA 02139
nickm@nickm.com phd.natali@gmail.com

Abstract

Creative computational systems have often been large-scale endeavors, based on elaborate models of creativity and sometimes featuring an accumulation of heuristics and numerous subsystems. An argument is presented for facilitating the exploration of creativity through small-scale systems, which can be more transparent, reusable, focused, and easily generalized across domains and languages. These systems retain the ability, however, to model important aspects of aesthetic and creative processes. Examples of extremely simple story generators are presented along with their implications for larger-scale systems. A case study focuses on a system that implements the simplest possible model of ellipsis.

Introduction

For a variety of institutional, intellectual, and other reasons, the typical computational system developed to model or produce creativity is a sizable one. Some of these systems, such as Harold Cohen's AARON, even become lifelong projects of their creators, continuing to accumulate rules and heuristics for decades.

There are certainly virtues to large-scale systems, which have revealed a great deal about formal models of creativity and creative computing. We present the argument that small-scale systems can also make contributions, serving to complement more extensive projects and to lead into them. Specifically, the argument is advanced that it makes sense to welcome such systems in new ways in conferences, in thesis work, and in the developing large-scale systems.

Rather than directly claiming that these small-scale systems are creative based on some formal definition, we argue that they *engage creativity* and are relevant to larger-scale systems that have been argued to be creative.

Small-Scale Systems that Engage Creativity

Many of the systems that will be discussed here are small – often limited to around 1 KB – and most were developed in a matter of hours or days. These are not systems built around a model of creativity; many of them, in fact, were not created with any particular research purpose in mind. However, each of these systems does explore one or more aspects of creativity relevant to its domain. These systems, without modeling creativity directly, nevertheless inquire about creativity. They also can focus larger-scale investigations of creativity that implement complete models.

The systems discussed here all use randomness within some framework of regularity. It can be creative to introduce randomness in a context where, individually or as a culture, regularity is the norm – and vice versa. But the connection between regular elements (a recurring vocabulary, a poetic form, etc.) and randomness (deployed in many different ways) is much more complex, as is the question of when randomness is a quick and easy substitute for a more sophisticated process and when it is the best method. While we believe that small-scale systems can be used to address issues of randomness and creativity, full discussion of this topic must be left for later.

Creative Text Generators of the 1950s and 1960s

By 1952, Christopher Strachey's innovative and certainly small-scale love letter generator was running on the Manchester Mark I and producing texts such as "YOU ARE MY EROTIC APPETITE: MY SWEET ENTHUSIASM. MY LOVE FONDLY WOOS YOUR CURIOUS TENDERNESS. YOU ARE MY WISTFUL SYMPATHY." The system runs today in emulation (Link 2007) and has been discussed recently as "the first experiment in digital literature" (Wardrip-Fruin 2011). Its purpose, it seems, was not to shine with brilliance but to parody the formulaic process of love-letter writing. By being a parody of a banal writing process, this small-scale system did serve as a model – a model of a *lack* of creativity – and demonstrated that computational processes could relate to human writing processes.

In 1959 Theo Lutz published on his small-scale system to generate stochastic texts based on Kafka's The Castle, pairs of "elementary sentences" with a logical connective. These include (in English translation from the German) "A CASTLE IS FREE AND EVERY FARMER IS FAR." and

"NO COUNT IS QUIET THEREFORE NOT EVERY CHURCH IS ANGRY." By drawing on a well-known author and transforming the text in a way that intensified his disquieting juxtapositions, Lutz created a system with a literary purpose. His system's operation, and its results, were consonant with Kafka's description of a formally valid social system in which the particular combinations were often meaningless.

In the United States, and in connection with the Fluxus movement, Alison Knowles and James Tenney published the 1968 chapbook *A House of Dust*. It consisted of 20 connected sheets of computer paper on which a poem generated by a Fortran program was printed, with each stanza of the same form. An example:

A HOUSE OF GLASS
IN A DESERTED FACTORY
USING ALL AVAILABLE LIGHTING
INHABITED BY COLLECTORS OF ALL TYPES

This project showed that variations of a regular stanza could be interesting when lined up and read one after the other, and that a creative language generator could produce a reasonably lengthy work that is compelling and worth reading.

"about so many things" and "Arrested" from Electronic Flipbooks
Nannette Wylde, 1998

These very simple systems for text generation were written in Macromedia Director. They merely place some strings that are selected uniformly at random into a simple template, the nature of which is self-evident. "Arrested" presents a sort of stanza that describes different situations in which people are arrested, and is, as Wylde describes it, "a play on preconceptions regarding social, ethnic, religious, and political affiliations." In the case of "about so many things," the template is simply "He" followed by a sentence completion and the "She" followed by a sentence completion, to produce text such as: "He likes chocolate / She thinks things should be different." The sentence completions range from being rather gender-neutral to being quite different when applied to people of different genders. For instance: "feels stressful," "is a good parent," "has a crush on the teacher," "is a firefighter."

The READ_ME file for "about so many things" explains: "the activities are drawn from the same pool of possibilities. Any line of text could be applied to either subject. In essence, the work explores the release of societal constraints regarding gender roles." Many sentences have different connotations when associated with people of different genders. By simply assigning sentences at random to be about either "He" or "She," "about so many things" produces interesting texts that provoke the reader to think about cultural preconceptions related to gender. The lesson for large-scale creative text generator is that determining the gender of a character, or transforming the gender of a character in an existing story, can be an important decision that is part of the creative process.

"The Two" and "Through the Park"
Nick Montfort, 2008

"The Two" builds on the core conceit of "About so Many Things." It uses even less text and only a slightly more complex template, such that the original Python program fits into 1 KB and the JavaScript version is not much larger. (The 1 KB limit is inspired in part by the demoscene, but also by poetic compression. While limitations of this sort are useful in many ways, and do enforce certain types of simplicity, they do not guarantee algorithmic simplicity or clear and readable code.) Two people described by their roles are introduced in the first line of each generated stanza; pronouns introduced in the second line require that the reader assign specific genders to the people in those two roles; and a conclusion involving both of them is provided:

The indigent turns to the librarian.
She smacks him.
They pray together.

In this case, two characters are introduced, the first of which is stereotypically male in U.S. culture. The second is usually culturally assumed to be female. Then, "she" and "he" appear on the second line, suggesting an obvious but disturbing resolution of reference: The librarian smacking the indigent. Since the typical reader's assumptions about the behavior of librarians and indigents will not line up with this interpretation, the reader may be compelled to consider the other interpretation, that the indigent is female and the librarian male. In either case, this generated text (and many of the texts that are generated) will challenge the reader's assumptions and stereotypes. This and other small-scale systems by this developer have been described and compared to systems of other sizes (Montfort, 2012).

Another 1 KB Python program that has also been made available in a JavaScript version is "Through the Park." This system is an attempt to build a very simple model of ellipsis or elision, the omission of part of a story. A carefully constructed list of sentences is reduced by a fixed number (removing sentences at random but keeping their order) and the resulting shorter story is output. The method of ellipsis has no intelligence or creativity to it, but with carefully constructed sentences it can nevertheless be effective. "Through the Park" is the subject of a short case study in the next section.

"The Semi-Automatic Doodle Machine" from Microcodes
Páll Thayer, 2010

This tiny program (at 756 characters, "a bit longer than most" in the Microcodes series, as Thayer notes) produces some simple instructions for that non-artistic but potentially creative drawing practice known as doodling. The program first prints "Use a pencil and a 210mm x 210mm sheet of paper. Start with your hand at the upper-left corner." and then prints some instruction such as "With pencil up, move 8mm to the right," printing a new one endlessly each time ENTER is pressed.

As a creative text generator, the program is curious be-

cause it generates instructions rather than a story or poem. Of course, the program is not framed as generating creative writing, but rather that non-artistic form of drawing known as a doodle. Seen as a generator of visual art, the program is rather hilarious. It uses a person as a sort of plotter, inverting the typical relationship between "user" and computer. With its tedious, precise instructions about how to do a task that has no external value, it might be a parody of creativity assistance software. It also highlights how computation can be applied at different stages of the creative process, questioning whether the entire pipeline of creative generation needs to be built for a system to be effective.

"Through the Park": A Case Study

The small-scale system "Through the Park" is about as simple as it can be while incorporating any computational elements at all. It provides a highly simplified model of an important narrative technique, however, a technique useful in full-scale story generation systems.

The Importance of Ellipsis

One way of understanding ellipsis is as one possible tempo at which a narrative may be related. In this view, it is the leaping over of one or more events in no time at all, which corresponds to telling the story at the fastest possible speed – an infinite speed (Prince 1982). The importance of this narrative technique has been articulated by narrative theorists, including Seymour Chatman: "Ellipsis is as old as *The Illiad*. But ... ellipsis of a particularly broad and abrupt sort is characteristic of modern narratives" (1978, p. 71). These omissions can allow the reader's imagination to fill the story in, as Fielding explains at the beginning of book III of *Tom Jones:*

> The reader will be pleased to remember, that … we gave him a hint of our intention to pass over several large periods of time ...
>
> In so doing, we do not only consult our own dignity and ease, but the good and advantage of the reader: for besides that by these means we prevent him from throwing away his time, in reading without either pleasure or emolument, we give him, at all such seasons, an opportunity of employing that wonderful sagacity, of which he is master, by filling up these vacant spaces of time with his own conjectures; for which purpose we have taken care to qualify him in the preceding pages.

Understanding ellipses has been the subject of some research, but generating ellipses has not been as well-studied. As recently as 2006, it appeared that computational narrative systems did not incorporate an ability to use ellipsis (Gervás et al.). Those in the field have noted the relevance of this technique to cinematic and textual story generation, however. The interactive fiction system Curveship (Montfort 2007 p. 107) can generate ellipses but does not determine how to do so. Ellipsis was also supported in the Mimesis system, because "narrative effects in [3D] environments are often achieved by selecting elements of the story world to elide from the narrative discourse (e.g., temporal and causal ellipsis) ..." (Young 2007 p. 14)

A Minimal Ellipsis System

"Through the Park" was prompted by a conversation with Michael Mateas about how to develop the simplest story generator grounded in a meaningful narrative technique. The first version of it, a 1 KB Python program, was posted on *Grand Text Auto* on November 20, 2008. It has 25 sentences. Nine are removed during execution and the remaining 16 are printed in their original order. The sentences are:

> The girl grins and grabs a granola bar.
> The girl puts on a slutty dress.
> The girl sets off through the park.
> A wolf whistle sounds.
> The girl turns to smile and wink.
> The muscular man paces the girl.
> Chatter and compliments cajole.
> The man makes a fist behind his back.
> A wildflower nods, tightly gripped.
> A snatch of song reminds the girl of her grandmother.
> The man and girl exchange a knowing glance.
> The two circle.
> Laughter booms.
> A giggle weaves through the air.
> The man's breathing quickens.
> A lamp above fails to come on.
> The man dashes, leaving pretense behind.
> Pigeons scatter.
> The girl runs.
> The man's there first.
> Things are forgotten in carelessness.
> The girl's bag lies open.
> Pairs of people relax after journeys and work.
> The park's green is gray.
> A patrol car's siren chirps.

The system is meant to tell a version of, or at least alludes to, the folktale Little Red Riding Hood. On *Grand Text Auto,* readers were asked if they considered a system this simple to be a story generator. While not all commenters agreed that it was one, game developer Gregory Weir was the first to reply, echoing Fielding in some ways:

> It's definitely a story generator. I like how my interpretation of the story can vary drastically on which cues are included. This is partly due to a few sharply-charged cues: the girl's smile, the knowing glance, the blank stare, and the police siren. Depending on which of these are included, cues like the girl's bag or the movement can be erotic or horrific.
>
> It does depend heavily on the mind's ability to fill in gaps … (Montfort 2008a)

The sentences were consciously written to suggest (although not directly assert) that the two characters might be in a friendlier or more antagonistic relationship, and that the situation is more playful or sinister.

Developing this generator led to an improved understanding of ellipsis and of the characteristics (both ontological and linguistic) of story elements and their representa-

tions. In this simple system, there is no representation of the underlying fabula or story levels that is separate from a potential text, which may or may not be included in the final, realized discourse. Linguistically, it is problematic to include pronouns or other words that refer to other sentences; if such words are used, "she" or "he" might appear before "the girl" or "the man" are introduced. The more cohesive a text is, the harder it is to elide a sentence from it without adjusting the other sentences.

The underlying events in a story also should be able to stand apart, but for narrative interest, it is appropriate that they are, in Weir's terms, "charged" with varying emotional implications. While it seems valuable for the events to be of different valences, it is also helpful that they contribute to a consistent scenario and agree on, for instance, who the two main characters are and what the setting is.

A more general model would allow different events/sentences to have different probabilities of being omitted; an even more general one would allow for conditional probabilities. Since experience with "Through the Park" suggests some qualities of the relationship between intersentential cohesion, the relationship between underlying events, and the opportunity for ellipsis, there are insights that could be applied in the development of more elaborate systems.

Generality across Languages

Gregory Rabassa has stated that "translation is essentially the closest reading one can give a text" (1989), suggesting that the translation of a computational system to produce linguistic or narrative creativity would at least have to involve a very deep analysis and understanding of the system. Large-scale systems are seldom translated because of the great effort that would be needed; small-scale systems are more manageable and can be translated in fairly short amounts of time, sometimes even by volunteers.

Fedorova translated "Through the Park" to Russian, demonstrating that the system does not only work in English. The small size of the system and simplicity of its operation facilitated this. The need to maintain an ambiguity of tone or emotion did complicate the translation process to some extent, further highlighting the particular way in which the original sentences were constructed. However, each of the sentences could be translated, resulting in a Russian system that produced ellipses with the same sorts of effects as the original English system.

Because "Through the Park" works at the sentence level, modifying the discourse without making adjustments to syntax, it is less language-specific than some other creative text generators are. "The Two" uses the ambiguity of gender of noun phrases in its first lines to achieve its effect; this ambiguity is not easy to achieve in all languages.

Generality across Story Domains

A system that is so specific that it can only tell one story, or one class of stories, is probably not worth much time or attention. While large systems are often difficult to convert to other story domains, adaptation is a good sign that the system is general. In the case of large-scale systems, such adaptations would often be difficult and time consuming; they are easier in smaller-scale systems.

The simple underlying system in "Through the Park" was re-used by writer and artist J. R. Carpenter to create two story generators, "Excerpts from the Chronicles of Pookie & JR" and "I've Died and Gone to Devon." The former program was used to produce much of the text of Carpenter's book *Generation[s]*. "Excerpts" was ported to JavaScript in 2009 by Ravi Rajakumar (independently of the port of "Through the Park"), translated into Spanish and Catalan in 2011 by Laura Borràs Castanyer, and translated into Russian by Natalia Fedorova in 2012.

Another system that uses "Through the Park" as a basis is Fedorova's "Halfway Through." This system has one Russian and one English array of sentences; it mingles an inner soliloquy with overheard phrases.

"Through the Park" is not the most-reused small-scale creative text generator (for instance, Montfort's *Taroko Gorge,* which was also originally a 1 KB Python program, has been appropriated and reworked online more than ten times) nor the most-translated (for instance, Montfort's "The Two," another originally 1 KB Python program ported to JavaScript, has been translated to French, Spanish, and Russian). Still, that it has been ported, translated, and re-used attests to its accessibility and flexibility.

Benchmarks, Baselines, and Subsystems for Larger-Scale Systems

A small-scale system can be used as a benchmark or baseline for evaluating larger-scale systems use similar techniques, driven by more elaborate methods. For instance, it could be worthwhile to compare a sophisticated system that elides parts of a story for a particular purpose (to generate suspense, to increase reader interest) against a system that elides at random, as "Through the Park" does. Even without a purpose-built story, such a system would reveal something about how effective the technique of elision or omission is when applied without any special logic, intelligence, or creativity. As a first step, developers of a creative system for ellipsis should show that it can exceed, by whatever metric, the effectiveness of a random one.

If an elaborate creative system to address one particular aspect of story-generation does not exceed the small-scale baseline, all is not lost. A larger-scale system that incorporates several subsystems can simply use the simple, random system to deal with that particular technique (ellipsis, assignment of gender to characters, or something else) while using more elaborate methods elsewhere.

Allowing for Small-Scale Work

Small-scale systems can be of direct as well as indirect significance. They can be easily understood and modified, even without the involvement of their original creators. The new systems that are developed in this way can contribute to new types of cultural production, having value

inside and outside the computational creativity community. They can be provocative, challenging the ideas that have been developed using large-scale systems and helping to develop some that have been overlooked. They can be used in teaching as the starting point for literary work or more elaborate exercises in computational expression. Finally, they can be used to sketch, as an artist would, in preparation for undertaking a large-scale work.

Despite the worth of small-scale projects and the slight effort that is needed to execute them, the context of computer science, and many interdisciplinary contexts, discourages work on such sketches and encourages researchers to proceed more directly to the development of large-scale systems. There are a few cases where small-scale systems are seen to have a place – for use as examples, for instance, or as subsystems in a larger system – but not many.

A dissertation project usually corresponds to a large-scale system, and the master's thesis and undergraduate capstone projects are typically reduced versions. Most conference papers are based on work with large-scale systems; even short papers, such as this one, are often invited not for the discussion of small-scale systems but for the dissemination of intermediate results about work in progress.

Ph.D. students in every field are already expected to understand their research area thoroughly by reviewing and understanding the relevant literature. It seems appropriate for them to spend as much time as they would reading a handful of articles in the development of one, or a few, small-scale systems. Such systems allow for different perspectives and approaches to be attempted; they also encourage a focus on the essential and on extreme abstraction of method and of the domain of creativity.

There are institutions that support, or could support, the development of small-scale systems. In particular, the hackathon, codefest, demo party, or other sort of competition, as often arranged outside of an academic context as inside it, could be employed to encourage the development of small-scale creativity systems. Although adding such an event to an existing conference would not change the paradigm for system development radically – those who were able to attend and compete would be there because their paper about a large-scale system was accepted – an event for quick development of systems could call attention to the value of such systems.

Small-scale systems have definite benefits, despite the institutional preference for using and discussing large-scale ones. These systems are easily portable across platforms, easily translated, easily generalized to different domains, and capable of capturing the essential aspects of important narrative techniques. Since they are also quick to put together, it would be sensible to do more to allow and encourage their development.

Acknowledgements

Our thanks to the anonymous reviewers, particularly for the suggestions that led to the discussion of randomness and the section on benchmarks, baselines, and subsystems.

References

Fielding, H. 2011. *The History of Tom Jones, a Foundling.* In Wikisource. Modified April 7. http://en.wikisource.org/wiki/The_History_of_Tom_Jones,_a_Foundling

Gervás, P., B. Lönneker-Rodman, J. C. Meister, F. Peinado. 2006 Narrative Models: Narratology Meets Artificial Intelligence. In *Proc. of Toward Computational Models of Literary Analysis, International Conference on Language Resources and Evaluation.*

Knowles, A. and J. Tenney. 1968. *A House of Dust.* Cologne: Verlag Gebrüder König. Excerpted in Pearson, L, ed., It Is Almost That: A Collection of Image+Text Work by Women Artists & Writers, Los Angeles: Siglio. 2011. 194-199.

Link, D. n.d. *Manchester Mark I Emulator.* http://www.alpha60.de/research/muc/

Lutz, T. 1959. Stochastische texte. In *Augenblick* 4:1, 3-9. Trans. H. MacCormack, 2005. http://www.stuttgarter-schule.de/lutz_schule_en.htm

Montfort, N. 2007. Generating Narrative Variation in Interactive Fiction. Ph.D. diss., University of Pennsylvania, Dpt of Computer and Information Science. http://www.cis.upenn.edu/grad/documents/montfort_000.pdf

Montfort, N. 2008a. Story Generation in 1k. On *Grand Text Auto.* November 20. http://grandtextauto.org/2008/11/20/story-generation-in-1k/

Montfort, N. 2008b. The Two. On *nickm.com.* http://nickm.com/poems/the_two.html

Montfort, N. 2008c. Through the Park. On *nickm.com.* http://nickm.com/poems/through_the_park.html

Montfort, N. 2012. XS, S, M, L: Creative Text Generators of Different Scales. January. Technical Report, The Trope Tank. TROPE-12-02. http://trope-tank.mit.edu/TROPE-12-02.pdf

Prince, G. 1982. *Narratology: The Form and Functiong of Narrative.* New York: Mouton Publishers.

Rabassa, G. 1989. No Two Snowflakes Are Alike. In *The Craft of Translation.* Chicago: Univ. of Chicago Press.

Thayer, P. 2010. "The Semi-Automatic Doodle Machine" From Microcodes. January. http://pallit.lhi.is/microcodes/

Wardrip-Fruin, N. 2011. Digital media archaeology : interpreting computational processes. In E. Huhtamo and J. Parikka, eds., *Media archaeology: approaches, applications, and implications.* Berkeley, Calif.: Univ. of California Press.

Wylde, N. 1998. About So Many Things and Arrested. *Electronic Flipbooks.* CD-ROM.

Young, M. 2007. Story and discourse: A bipartite model of narrative generation in virtual worlds. In *Interaction Studies* 8:2, 177–208.

Automatic Generation of Melodic Accompaniments for Lyrics

Kristine Monteith, Tony Martinez, and Dan Ventura
Computer Science Department
Brigham Young University
Provo, UT 84602 USA
kristinemonteith@gmail.com, martinez@cs.byu.edu, ventura@cs.byu.edu

Abstract

Music and speech are two realms predominately species-specific to humans, and many human creative endeavors involve these two modalities. The pairing of music and spoken text can heighten the emotional and cognitive impact of both - the complete song being much more compelling than either the lyrics or the accompaniment alone. This work describes a system that is able to automatically generate and evaluate musical accompaniments for a given set of lyrics. It derives the rhythm for the melodic accompaniment from the cadence of the text. Pitches are generated through the use of *n*-gram models constructed from melodies of songs with a similar style. This system is able to generate pleasing melodies that fit well with the text of the lyrics, often doing so at a level similar to that of human ability.

Introduction

Programmers and researchers have often attempted to endow machines with some form of intelligence. In some cases, the end goal of this is purely practical; a machine with the capacity to learn could provide a multitude of useful and resource-saving tasks. But in other cases, the goal is simply to make machines behave in a more creative or more "human" manner. As one author explains, "Looked at in one way, ours is a history of self-imitation...We are ten times more fascinated by clockwork imitations than by real human beings performing the same task." (McCorduck 2004).

One major area of human creativity involves the production of music. Wiggins (2006) states that, "...musical behavior is a uniquely human trait...further, it is also ubiquitously human: there is no known human society which does not exhibit musical behaviour in some form." Naturally, many computer science researchers have turned their attention to musical computation tasks. Researchers have attempted to classify music, measure musical similarity, and predict the musical preferences of users (Chai and Vercoe 2001; McKay and Fujinaga 2004). Others have investigated the ability to search through, annotate, and identify audio files (Dannenberg et al. 2003; Dickerson and Ventura 2009). More directly in the realm of computational creativity, researchers have developed systems that can automatically arrange and compose music (Oliveira and Cardoso 2007; Delgado, Fajardo, and Molina-Solana 2009).

Like music, speech is an ability that is almost exclusively human. While species such as whales or birds may communicate through audio expressions, and apes may even be taught simple human-like vocabularies and grammars using sign language, the complexities of human language set us apart in the animal kingdom. Major research efforts have been directed toward machine recognition and synthesis of human speech (Rabiner 1989; Koskenniemi 1983) Computers programs have been designed to carry on conversations, some of them doing so in a surprisingly human-like manner (Weizenbaum 1966; Saygin, Cicekli, and Akman 2000). More creative programming endeavors have involved the generation of poetry (Gervás 2001; Rahman and Manurung 2011) or text for stories (Riedl 2004; Pérez y Pérez and Sharples 2004; Gervás et al. 2005; Ang, Yu, and Ong 2011).

Gfeller (1990) points out the similarities between speech and music: "Both speech and music are species specific and can be found in all known cultures. Both forms of communication evolve over time and have structural similarities such as pitch, duration, timbre, and intensity organized through particular rules (i.e. syntax or grammar) that result in listener expectations." Studies show that music and the spoken word can be particularly powerful when paired together. For example, in one study, researchers found that a sung version of a story was often more effective at reducing an undesirable target behavior than a read version of the story (Brownell 2002). Music can help individuals with autism and auditory processing disorders more easily engage in dialog (Wigram 2002). The pairing of music with language can even help individuals regain lost speech abilities through a process known as Melodic Intonation Therapy (Gfeller 1990; Schlaug, Marchina, and Norton 2008). On the other hand, lyrics have the advantage of being able to impart discursive information where the more abstract nature of music makes it less fit to do so (Kreitler and Kreitler 1972). Lyrics can also contribute to the emotional impact of a song. One study found that lyrics enhanced the emotional impact of a selection with sad or angry music (Ali and Peynircioglu 2006). Another found that lyrics tended to be a better estimator of the overall mood of a song than the melody when the lyrics and the melody disagree (Wu et al. 2009).

This work describes a system that can automatically com-

pose melodic accompaniments for any given text. For each given lyric, it generates hundreds of different possibilities for rhythms and pitches and evaluates these possibilities with a number of different metrics in order to select a final output. The system also incorporates an awareness of musical style. It learns stylistic elements from a training corpus of melodies in a given genre and uses these to output a new piece with similar elements. In addition to self-evaluation, the generated selections are further evaluated by a human audience. Survey feedback indicates that the system is able to generate melodies that fit well with the cadence of the text and that are often as pleasing as the original accompanying tunes. Colton, Charnley, and Pease (2011) suggest a number of different measures that can be used to evaluate systems during the creative process. We direct particular attention to two of these–precision and reliability–and demonstrate that, for simpler styles, our system is able to perform well with regard to these metrics.

Related Work

Conklin (2003) summarizes a number of statistical models which can be used for music generation, including random walk, Hidden Markov Models, stochastic sampling, and pattern-based sampling. These approaches can be seen in a number of different studies. For example, Chuan and Chew (2007) use Markov chains to harmonize given melody lines, focusing on harmonization in a given style. Cope (2006) also uses statistical models to generate music in a particular style, producing pieces indistinguishable from human-generated compositions. Pearce and Wiggins (2007) provide an analysis of a number of strategies for melodic generation, including one similar to the generative model used in this paper.

Delgado, Fajardo, and Molina-Solana (2009) use a rule-based system to generate compositions according to a specified mood. Oliveira and Cardoso (2007) describe a wide array of features that contribute to emotional content in music and present a system that uses this information to select and transform chunks of music in accordance with a target emotion.

Researchers have also directed efforts towards developing systems intended for accompaniment purposes. Dannenberg (1985) presents a system of automatic accompaniment designed to adapt to a live soloist. Lewis (2000) also details a "virtual improvising orchestra" that responds to a performer's musical choices.

While not directly related to generating melodic accompaniment for lyrics, a number of studies have looked at aligning musical signals to textual lyrics (the end result being similar to manually-aligned karaoke tracks). For example, Wang and associates (2004) use both low-level audio features and high-level musical knowledge to find the rhythm of the audio track and use this information to align the music with the corresponding lyrics.

Methodology

In order to generate original melodies, a set of melodies is compiled for each different style of composition. These melodies were isolated from MIDIs obtained from the Free MIDI File Database[1] and the "I Love MIDIs" website[2]. These selections help determine both the rhythmic values and pitches that will be assigned to each syllable of the text. The system catalogs the rhythmic patterns that occur for each of the various numbers of notes in a given measure. The system also creates an n-gram model representing what notes are most likely to follow a given series of notes in a given set of melodies. Models were developed for three stylistic categories: nursery rhymes, folk songs (bluegrass), and rock songs (Beatles).

For each lyric, the system first analyzes the text and assigns rhythms. It determines where the downbeats will fall for each given line of the text. One hundred different downbeat assignments are generated randomly, and evaluated according to a number of aesthetic measures. The system selects the random assignment with the highest score for use in the generated melody. The system then determines the rhythmic values that will be assigned to each syllable in the text by counting the number of syllables in a given measure and finding a rhythm that matches that number of syllables in one of the songs of the training corpus. Once rhythmic values are assigned, the system assigns pitches to each value using the n-gram model constructed from the training corpus. Once again, one hundred different assignments are generated and evaluated according to a number of metrics. Further details on the rhythm and pitch generation are provided in the following subsections.

Rhythm Generation

Rhythms are generated based on patterns of syllabic stress in the lyrics. Each word of the text is located in the CMU Pronunciation Dictionary[3] to determine the stress patterns of the constituent phonemes. (Each phoneme in the dictionary is labeled 0, 1, or 2 for "No Stress," "Primary Stress," or "Secondary Stress.") The system also looks up each word to determine if it occurs on a list of common articles, prepositions, and conjunctions.

The system then attempts to find the best positions for downbeats. For each given line of text, the system generates 100 possible downbeat assignments. The text of each line is distributed over four measures, so four syllables are randomly selected to carry a downbeat. Each assignment is then scored, and the system selects the assignment receiving the highest score for use in the melodic accompaniment. Downbeat assignments that fall on stressed syllables are rated highly, as are downbeats that fall on the beginning of a word and ones that do not fall on articles, prepositions, or conjunctions. Downbeat assignments that space syllables more evenly across the allotted four measures are also rated more highly (i.e. assignments that have a lower standard deviation for number of syllables per measure receive higher scores). See Figure 4 for further details on the precise downbeat scoring metrics. Figure 1 illustrates a possible downbeat assignment for a sample lyric.

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

[2] http://www.ilovemidis.com/ForKids/NurseryRhymes/

[3] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Lyrics:	Pat	a	cake	pat	a	cake
Phonemes:	PAET	AH	KEYK	PAET	AH	KEYK
Stress:	1	0	1	1	0	1
Downbeats:	true	false	false	true	false	false

Lyrics:	ba-	ker's	man
Phonemes:	BEY	KERZ	MAEN
Stress:	1	0	1
Downbeats:	true	false	true

Figure 1: Sample downbeat assignments for *Pat-A-Cake* lyrics

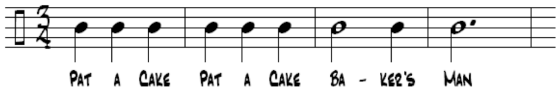


Figure 2: Default rhythm assignments for *Pat-A-Cake* lyrics

Once the downbeats are assigned, a rhythmic value is assigned to each syllable. The system randomly selects a piece in the training corpus to provide rhythmic inspiration. This selection determines the time signature of the generated piece (e.g. three beats or four beats to a measure). For each measure of music generated, the system looks to the selected piece and randomly chooses a measure that has the necessary number of notes. For example, if the system needs to generate a rhythm for a measure with three syllables, it randomly chooses a measure in the training corpus piece that has three notes in it and uses its rhythm in the generated piece. If no measures are available that match the number of syllables in the lyric, the system arbitrarily assigns rhythmic values, with longer values being assigned to earlier syllables. For example, in a measure with three syllables using a three-beat pattern, each syllable would be assigned to a quarter note. In a measure with four syllables, the first two syllables would be assigned to quarter notes and the last two syllables to eighth notes. Figure 2 illustrates the default rhythms assignment for a sample lyric.

Pitch Generation

Once the rhythm is determined, pitches are selected for the various rhythmic durations. Selections from a given style corpus are first transposed into the same key. Then an *n*-gram model with an *n* value of four is constructed from these original melodic lines. The model was created simply from the original training melodies, with no smoothing. For the new, computer-generated selections, melodies are initialized with a series of random notes, selected from a distribution that models which notes are most likely to begin musical selections in the given corpus. In order to foster song cohesion, each line of the song is initialized with the same randomly generated three notes. Additional notes for each line are randomly selected based on a probability distribution of what note is most likely to follow the given three notes as indicated by the *n*-gram model of the style corpus.

The system generates several hundred possible series of pitches for each line. Each possible pitch assignment is then scored. To encourage melodic interest, higher scores are given to melodic lines with a higher number of distinct

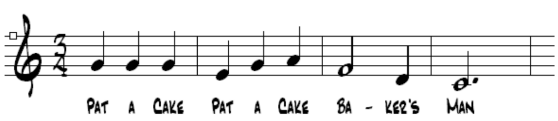


Figure 3: Sample pitch assignments for *Pat-A-Cake* lyrics

pitches and melodies featuring excessive repeated notes are penalized. Melodies with a range greater than an octave and a half or with interval jumps greater than an octave are penalized since these are less "sing-able." Melodic lines that do not end on a note in a typical major or minor scale and final melodic lines that do not end on a tonic note are given a score of zero. More precise details about the scoring of pitch assignments are given in Figure 4. Possible pitch assignments for a sample lyric are shown in Figure 3.

Results

Accompaniments were generated for lyrics in three stylistic categories: nursery rhymes, folk songs (bluegrass), and rock songs (Beatles). In each case, an attempt was made to find less commonly known melodies, so that the generated music could be more fairly compared to the original melodic lines. Melodic lines were generated for the following:

Nursery rhymes:

- Goosey Goosey Gander
- Little Bo Peep
- Pat-a-Cake
- Rub-a-Dub-Dub
- The Three Little Kittens

Folk songs:

- Arkansas Traveler
- Battle of New Orleans
- Old Joe Clark
- Sally Goodin
- Wabash Cannonball

Rock songs:

- Act Naturally
- Ask Me Why
- A Taste of Honey
- Don't Pass Me By
- I'll Cry Instead

Three melodies were generated for each of the fifteen lyrics considered. One was generated using a corpus of songs that matched the style of the lyrics (e.g. to generate a melody for *Goosey Goosey Gander* the four other nursery

MelodicAccompaniment($Lyric$,$StyleCorpus$)
for all $LINE_i$ in $Lyric$ **do**
  $STR_i \leftarrow$ patterns of syllabic stress in $LINE_i$
  $POS_i \leftarrow$ parts of speech for each syllable in $LINE_i$
  $BEG_i \leftarrow$ boolean values indicating that a syllable in $LINE_i$ begins a word
  for $i = 1 \rightarrow 100$ **do**
    $DB_j \leftarrow$ randomly assign downbeats to four syllables
    $score_j \leftarrow$ ScoreDownbeats(DB_j,STR_i,POS_i,BEG_i)
  end for
  $DB_i \leftarrow DB_j$ that coincides with the largest $score_j$
  $RHYTHM_i \leftarrow$ SelectRhythms(DB_i)
  for $i = 1 \rightarrow 100$ **do**
    $PITCHES_j \leftarrow$ assign pitches using *n*-gram model from $StyleCorpus$
    $score_j \leftarrow$ ScorePitches($PITCHES_j$)
  end for
  $PITCHES_i \leftarrow PITCHES_j$ that coincides with the largest $score_j$
  $MELODY_i \leftarrow$ combine $RHYTHM_i$ and $PITCHES_i$
  $MELODY += MELODY_i$
end for
return $MELODY$

ScoreDownbeats(DB_j,STR_i,POS_i,BEG_i)
for $k = 1 \rightarrow j$ **do**
  If DB_{jk} and $STR_{ik} = 1$ then $score += 1$
  If DB_{jk} and $POS_{ik}! = Art|Prep|Conj$ then $score += 0.5$
  If DB_{jk} and BEG_{ik} then $score += 0.5$
  $x \leftarrow maxSyllablesPerMeasure$
  $score += (x - stdDevSyllablesPerMeasure) * 0.5$
  $score += (x - numPickupSyllables) * 0.25$
  $score += (x - numSyllablesLastMeasure) * 0.25$
end for
return score

SelectRhythms(D_i, S_i)
$M \leftarrow$ divide S_i into measures based on D_i
$C \leftarrow$ randomly select a song in $StyleCorpus$
$R \leftarrow 0$
for all M_j in M **do**
  $R_j \leftarrow$ randomly selected measure from C with the same # of notes as syllables in M_j
  R += R_j
end for
return R

ScorePitches($PITCHES_j$)
$score \leftarrow uniquePitches(PITCHES_j)/size(PITCHES_j)$
If $MaxRepeatPitches(PITCHES_j) < maxRepeatPitches$ then $score += 1$
If $Range(PITCHES_j) < maxRange$ then $score += 1$
If $MaxInterval(PITCHES_j) < maxInterval$ then $score += 1$
If $!EndsOnScaleNote(PITCHES_j)$ then $score = 0$
If $LastLine(j)$ and $!EndsOnTonic(PITCHES_j)$ then $score = 0$
return score

Figure 4: Algorithm for automatically generating melodic accompaniment for text

	bluegrass	nursery	rock	average
bluegrass	1.34	3.09	1.19	1.87
nursery	1.14	3.32	1.19	1.88
rock	1.25	3.28	1.11	1.88
original	1.50	3.50	1.47	2.16

Table 1: Average responses to the question "How familiar are you with these lyrics?" Each row represents a compositional style and each column a category of lyrics.

rhyme songs were used to build the *n*-gram model) and two more were generated in the remaining two creative styles[4].

Study participants were divided into four groups. Each group was asked to listen to versions of songs for each of the fifteen lyrics, with selections for each group being a mixture of lyrics with the original human-composed melodies and lyrics with the three types of computer-generated melodies. Subjects were not informed that any of the melodies were computer-generated until after data collection. Fifty-two subjects participated in the study, and each melodic version was played for thirteen people.

After each selection, subjects were asked to respond to the following questions (1=not at all, 5=very much):

- How familiar are you with these lyrics? 1 2 3 4 5
- How familiar are you with this melody? 1 2 3 4 5
- How pleasing is the melodic line? 1 2 3 4 5
- How well does the music fit with the lyrics? 1 2 3 4 5
- Is this the style of melody you would have expected to accompany these lyrics? 1 2 3 4 5
- Are you familiar with any other melodies for these lyrics? YES NO

Table 1 shows the average responses to the question about familiarity of lyrics for each of the three categories. In each case, lyrics were rated as more familiar when they were paired with their original melodies as opposed to the computer-generated melodies. However, none of these differences were significant at the $p < 0.05$ level. The majority of subjects were relatively unfamiliar with the bluegrass and rock lyrics. The nursery rhyme lyrics were slightly more familiar, but in many cases, subjects were familiar with the lyrics but not any specific tune.

Table 2 shows the average responses to the question about familiarity of melody for each of the three categories. On average, subjects were slightly more familiar with the original melodies in the bluegrass and rock categories than they were with the lyrics. The original nursery rhymes melodies were rated as slightly less familiar on average than the lyrics. System-generated melodies received an average score of less than two for familiarity in each of the three categories (significantly lower than original melodies with a statistical significance of $p < 0.01$).

Subjects were likely to be less receptive to new melodies if they were very familiar with the old ones. (One respondent

[4]Selections generated for these experiments are available at http://axon.cs.byu.edu/emotiveMusicGeneration

	bluegrass	nursery	rock	average
bluegrass	1.62	1.49	1.40	1.50
nursery	1.53	2.17	1.34	1.68
rock	1.41	1.39	1.24	1.35
original	2.31	2.94	1.81	2.35

Table 2: Average responses to the question "How familiar are you with this melody?" Each row represents a compositional style and each column a category of lyrics.

	bluegrass	nursery	rock	average
bluegrass	3.50	3.50	3.56	3.52
nursery	3.37	3.24	3.09	3.23
rock	2.70	2.17	2.16	2.34
original	3.79	3.79	2.95	3.51

Table 3: Average responses to the question "How pleasing is the melodic line?" Each row represents a compositional style and each column a category of lyrics.

mentioned that hearing a new melody to a familiar childhood song was a little "unnerving".) Tables 3 through 7 report only the responses where subjects indicated that they were not familiar with an alternate melody for a given set of lyrics.

As shown in Table 3, the system was able to generate melodies that received the same average ratings for pleasing melodic lines as the original melodies. The average rating for songs in the bluegrass style was almost identical to that of the original melodies. The average ratings for pleasantness of generated nursery rhythm melodies was not significantly different than the original tunes.

For over a third of the lyrics, a computer-generated melody in at least one style was rated as more pleasing than the original melody. These tunes are listed in Table 4 along with their average ratings. For example, the original melody for *Battle of New Orleans* received a rating of 3.33 for average melodic pleasantness. The computer-generated melody for this lyric in a nursery rhyme style received a rating of 3.92. The original melody for *Little Bo Peep* received an average melodic pleasantness rating of 3.22. The bluegrass-styled computer-generated melody received a rating of 3.80, and the nursery-rhyme-styled generated melody received a rating of 3.43.

Table 5 shows that the original melodies were rated on average as fitting a little better with the lyrics (although the difference between the original melodies and the songs composed in the bluegrass style is not statistically significant). However, as shown in Table 6 a number of the individual computer-generated melodies were still rated as fitting better with the lyrics than the original melodies. For example, the rock version of *Old Joe Clark* received a rating of 3.00 from this metric while the original version received a rating of 2.75. Both the bluegrass and nursery-rhyme versions of *Ask Me Why* received higher ratings than the original version.

Table 7 reports responses to the question "Is this the style of melody you would have expected to accompany these lyrics?" Not surprisingly, the original melodies were

	Battle of New Orleans	Little Bo Peep	Rub A Dub Dub	Act Naturally	Ask Me Why	I'll Cry Instead
bluegrass	3.23	**3.60**	**3.80**	**3.50**	**4.23**	**3.79**
nursery	**3.92**	**3.43**	3.17	2.91	**3.14**	**2.92**
rock	2.83	2.60	2.13	2.54	2.00	**2.36**
original	3.33	3.22	3.50	2.70	2.83	2.12

Table 4: Average responses to the question "How pleasing is the melodic line?" for six songs where system-generated melody in one or more styles scored higher than the original melody.

	bluegrass	nursery	rock	average
bluegrass	3.59	3.20	3.18	3.32
nursery	3.35	3.36	2.71	3.14
rock	3.23	2.18	2.26	2.56
original	3.88	4.27	2.90	3.68

Table 5: Average responses to the question "How well does the music fit with the lyrics?" Each row represents a compositional style and each column a category of lyrics.

more "expected" on average than melodies composed in new styles. The computer-generated melodies composed in the style of the original melodies were also generally more expected with one exception: bluegrass melodies for rock lyrics tended to receive higher expectation ratings.

In a number of cases, the system was able to compose an unexpected melody that still received high ratings for pleasing melodies and a lyric/note match. Two such examples are shown in Table 8. In both cases, the songs received above average ratings for melodic pleasantness and average ratings for music/lyric match, but below average ratings for style expectedness.

Discussion

The original nursery rhymes were composed predominantly with notes of the major scale, and the rhythms in these songs were similarly simple. (Songs generated with corpus-inspired rhythms were quite similar to songs generated with the system's default rhythms.) With the exception of a flat seventh introduced by the mixolydian scale of *Old Joe Clark*, the bluegrass melodies also feature pitches exclusively from the major scale. Bluegrass rhythms also tended to be similarly straightforward. With simpler rhythms and fewer accidentals, more of the melodies generated in these two styles are likely to "work." The original bluegrass melodies tended to have more interesting melodic motion, and this appears to have translated into more interesting system-generated melodies. In contrast, the rock songs featured a much wider variety of scales and accidentals. These

	Arkansas Traveler	Old Joe Clark	Three Little Kittens	Ask Me Why	A Taste of Honey	I'll Cry Instead
bluegrass	**4.08**	2.71	**4.25**	**3.54**	2.57	**3.43**
nursery	3.08	2.75	3.80	**3.07**	**2.85**	**2.38**
rock	3.08	**3.00**	2.18	1.77	2.08	**2.27**
original	3.91	2.75	4.17	2.75	2.79	2.15

Table 6: Average responses to the question "How well does the music fit with the lyrics?" for six songs where system-generated melody in one or more styles scored higher than the original melody.

	bluegrass	nursery	rock	average
bluegrass	3.47	2.85	2.91	3.08
nursery	3.22	3.46	2.44	3.04
rock	3.12	1.82	2.14	2.36
original	3.69	4.27	2.79	3.58

Table 7: Average responses to the question "Is this the style of melody you would have expected to accompany these lyrics?"

extra tones do add color to the generated selections, but further refinements may be necessary to select which more complicated melodies are "fresh" or "original" instead of just "weird."

Wiggins (2006) proposes a definition for computational creativity as "The performance of tasks which, if performed by a human, would be deemed creative." The task of simply composing any decent new melody for an established tune could be considered creative. Composing one that improved on the original constitutes an even greater degree of creative talent. By this metric, our system fits the definition of "creative."

	Pat-A-Cake (bluegrass)	Act Naturally (bluegrass)
How pleasing is the melodic line?	3.80	3.50
How well does the music fit with the lyrics?	3.20	3.17
Is this the style of melody you would have expected?	2.60	2.50

Table 8: Average responses to questions for two songs where the melodic accompaniment was surprising but still worked.

Colton (2008) suggests that, for a computational system to be considered creative, it must be perceived as possessing skill, appreciation, and imagination. A basic knowledge of traditional music behavior allows a system to meet the "skillful" criteria. Our system takes advantage of statistical information about rhythms and melodic movement found in the training songs to compose new melodies that behave according to traditional musical conventions. A computational system may be considered "appreciative" if it can produce something of value and adjust its work according the preferences of itself or others. Our system addresses this criterion by producing hundreds of different possible rhythm and pitch assignments and evaluating them against some basic rules for pleasantness and singability. The "imaginative" criterion can be met if the system can create new material independent of both its creators and other composers. Since all of the generated melodies can be distinguished from songs in the training corpora, this criterion is met at least on a basic level. Our system further demonstrates its imaginative abilities by composing melodies in alternate styles that still manage to demonstrate an acceptable level of melodic pleasantness and synchronization with the cadence of the text.

Boden (1995) argues that unpredictability is also a critical element of creativity, and a number of researchers have investigated the role of unpredictability in creative systems (Macedo 2001; Macedo and Cardoso 2002) Our system meets the requirement of unpredictability with its ability to compose in various and sometimes unexpected styles. It is able to generate melodies that surprise listeners but still achieve high ratings for pleasantness.

Colton, Charnley, and Pease (2011) propose a number of different metrics in conjunction with their FACE and IDEA models that can be used to assess software during a session of creative acts. Equations for calculating these metrics are listed in Figure 5, were S is the creative system, (c_i^g, e_i^g) is a concept/expression pair generated by the system, $\overline{a^g}$ is an aesthetic measure of evaluation, and t is a minimum acceptable aesthetic threshold. Two of the measures suggested are precision (obtained by dividing the number of generated works by the number that met a minimum acceptable aesthetic level) and reliability (obtained from taking the system's best creation as calculated by some aesthetic measure and subtracting the system's worst). Table 9 reports the results of these calculations for the system's compositions in each of the three styles and compares them to the same metrics calculated for the original songs using responses to the question "How pleasing is the melodic line?" as the scoring metric. In order to calculate precision, we consider the worst score obtained by an original, human-composed melody to be the minimum acceptable threshold value. While the prize for most pleasing melody still goes to a human-composed song, all of the songs composed in a bluegrass and nursery style and two-thirds of the rock songs meet the basic criteria of being better than the worst original melody. The system is generating original melodies that are better than some established, human-generated songs a remarkable percentage of the time. The reliability of the system in generating bluegrass and nursery-style melodies is also worth mentioning. The reliability measures for these two categories are

$$average(S) = \frac{1}{n}\sigma_{i=1}^{n}\overline{a^g}(c_i^g, e_i^g)$$
$$best\_ever(S) = max_{i=1}^{n}(\overline{a^g}(c_i^g, e_i^g))$$
$$worst\_ever(S) = min_{i=1}^{n}(\overline{a^g}(c_i^g, e_i^g))$$
$$precision(S) = \frac{1}{n}|\{(c_i^g, e_i^g) : 1 < i < n \wedge \overline{a^g}(c_i^g, e_i^g) > t\}|$$
$$reliability(S) = best\_ever(S) - worst\_ever(S)$$

Figure 5: Assessment metrics proposed by Colton, Charnley, and Pease (2011)

	bluegrass	nursery	rock	original
average	3.52	3.23	2.34	3.51
best ever	4.23	3.92	3.83	4.50
worst ever	2.93	2.58	1.73	2.12
precision	1.00	1.00	0.67	1.00
reliability	1.30	1.33	2.11	2.38

Table 9: Assessment metrics calculate on average responses to the question "How well does the music fit with the lyrics?"

1.30 and 1.33 as compared to the 2.38 reliability measure for original songs. (Note that, for reliability, smaller scores are more desirable.) While the system probably shouldn't quit its day job to become a classic rock songwriter quite yet, it is considerably reliable at producing reasonable and pleasing melodies in the other two genres.

Similar results can be seen in Table 10 where responses to the question "How well does the music fit with the lyrics?" are used as the aesthetic measure. As with the previous calculations, the "worst ever" score for an original melody was used as a minimum aesthetic threshold for the generated melodies. Again, all of the nursery rhyme and bluegrass-styled compositions meet this threshold, as do two-thirds of the rock-styled songs. A song generated in the nursery rhyme or bluegrass style also more reliably matches the lyrics than an arbitrarily selected human-generated song.

Previous versions of our system analyzed each melody in a given training corpus according to a number of different metrics and used the results in the construction of neural networks designed to evaluate generated melodies (Monteith, Martinez, and Ventura 2010). For the sake of simplicity and computational speed, the most pertinent of these findings were distilled into rules for use by the system in these experiments. In other words, the information gathered by the system to date about melody generation has been simplified and codified so that more focus could be directed towards matching rhythms to text. However, the system could likely benefit from the use of additional metrics and further "observation" of human-generated and approved tunes in its attempts to create pleasing melodies. A similar process of evaluation could be applied to the process of rhythm generation, particularly in the assignment of downbeats. Currently, the system relies on a small set of arbitrary, pre-coded rules to determine downbeat placement. It would likely require a much larger training corpus than we currently have available, but perhaps more natural-sounding placements could be obtained if the system could learn from a corpus of "good" lyric/melody pairings the types of words

	bluegrass	nursery	rock	original
average	3.32	3.14	2.56	3.68
best ever	4.25	3.86	4.23	4.75
worst ever	2.57	2.36	1.63	2.15
precision	1.00	1.00	0.67	1.00
reliability	1.68	1.49	2.61	2.60

Table 10: Assessment metrics calculate on average responses to the question "How well does the music fit with the lyrics?"

and syllables best suited for supporting downbeats. Audience feedback could help determine an optimal weighting of the various evaluation criteria.

References

Ali, S. O., and Peynircioglu, Z. F. 2006. Songs and emotions: are lyrics and melodies equal partners? *Psychology of Music* 4(4):511–534.

Ang, K.; Yu, S.; and Ong, E. 2011. Theme-based cause-effect planning for multiple-scene story generation. In *Proceedings of the International Conference on Computational Creativity*, 48–53.

Boden, M. 1995. Creativity and unpredictability. *Stanford Humanities Review* 4.

Brownell, M. D. 2002. Musically adapted social stories to modify behaviors in students with autism: four case studies. *Journal of Music Therapy* 39:117–144.

Chai, W., and Vercoe, B. 2001. Folk music classification using hidden markov models. In *Proceedings of the International Conference on Artificial Intelligence*.

Chuan, C., and Chew, E. 2007. A hybrid system for automatic generation of style-specific accompaniment. In *Proceedings International Joint Workshop on Computational Creativity*, 57–64.

Colton, S.; Pease, A.; and Charnley, J. 2011. Computational creativity theory:the FACE and IDEA descriptive models. In *Proceedings of the 2nd International Conference on Computational Creativity*, 90–95.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *Creative Intelligent Systems: Papers from the AAAI Spring Symposium*, 14–20. Stanford, CA: AAAI Press.

Conklin, D. 2003. Music generation from statistical models. In *Proceedings of the AISB Symposium on Artificial Intelligence and Creativity in the Arts and Sciences*, 30–35.

Cope, D. 2006. *Computer Models of Musical Creativity*. Cambridge, Massachusetts: The MIT Press.

Dannenberg, R. B.; Birmingham, W. P.; Tzanetakis, G.; Meek, C.; Hu, N.; and Pardo, B. 2003. The MUSART testbed for query-by-humming evaluation. In *Proceedings of the International Conference on Music Information Retrieval*, 41–51.

Dannenberg, R. B. 1985. An on-line algorithm for real-

time accompaniment. In *Proceedings of the International Computer Music Conference*, 279–289.

Delgado, M.; Fajardo, W.; and Molina-Solana, M. 2009. Inmamusys: Intelligent multi-agent music system. *Expert Systems with Applications* 36(3-1):4574–4580.

Dickerson, K., and Ventura, D. 2009. Music recommendation and query-by-content using self-organizing maps. In *Proceedings of the International Joint Conference on Neural Networks*, 705–710.

Gervás, P.; Díaz-Agudo, B.; Peinado, F.; and Hervás, R. 2005. Story plot generation based on CBR. *Journal of Knowledge-Based Systems* 18(4–5):235–242.

Gervás, P. 2001. An expert system for the composition of formal spanish poetry. *Journal of Knowledge-Based Systems* 114(3-4):181–188.

Gfeller, K. 1990. Music, the language of emotions. In Unkefer, R., ed., *Music Therapy in the Treatment of Adults with Mental Disorders; Theoretical Basis and Clinical Interventions*. New York: Schirmer Books.

Koskenniemi, K. 1983. *Two-level morphology: A general computational model of word-form recognition and production*. University of Helsinki: Department of General Linguistics.

Kreitler, H., and Kreitler, S. 1972. *Psychology of the arts*. Durham, NC: Duke University Press.

Lewis, G. 2000. Too many notes: Computers, complexity and culture in voyager. *Leonardo Music Journal* 10:33–39.

Macedo, L., and Cardoso, A. 2002. Assessing creativity: the importance of unexpected novelty. In *Proceedings of the ECAI'02 Workshop on Creative Systems: Approaches to Creativity in Artificial Intelligence and Cognitive Science*, 31–38.

Macedo, L. 2001. Creativity and surprise. In *Proceedings of the AISB Symposium on Artificial Intelligence and Creativity in Arts and Science*, 84–92.

McCorduck, P. 2004. *Machines Who Think*. Natick, MA: A. K. Peters, Ltd., 2nd edition.

McKay, C., and Fujinaga, I. 2004. Automatic genre classification using large high-level musical feature sets. In *Proceedings of the Fifth International Symposium on Music Information Retrieval*, 525–530.

Monteith, K.; Martinez, T.; and Ventura, D. 2010. Automatic generation of music for inducing emotive response. In *Proceedings of the International Conference on Computational Creativity*, 140–149.

Oliveira, A., and Cardoso, A. 2007. Towards affective-psychophysiological foundations for music production. In *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction*, 511–522.

Pearce, M. T., and Wiggins, G. A. 2007. Evaluating cognitive models of musical composition. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, 73–80.

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

Rabiner, L. R. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2):257–286.

Rahman, F., and Manurung, R. 2011. Multiobjective optimization for meaningful metrical poetry. In *Proceedings of the International Conference on Computational Creativity*, 4–9.

Riedl, M. 2004. Narrative generation: Balancing plot and character. *Ph.D. Dissertation, North Carolina State University*.

Saygin, A. P.; Cicekli, I.; and Akman, V. 2000. Turing test: 50 years later. *Minds and Machines* 10(4):463–518.

Schlaug, G.; Marchina, S.; and Norton, A. 2008. From singing to speaking: Why patients with Broca's aphasia can sing and how that may lead to recovery of expressive language functions. *Music Perception* 25:315–323.

Wang, Y.; Kan, M.-Y.; Nwe, T. L.; Shenoy, A.; and Yin, J. 2004. LyricAlly: automatic synchronization of acoustic musical signals and textual lyrics. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, 212–219. New York, NY, USA: ACM Press.

Weizenbaum, J. 1966. ELIZA - a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9(1):36–45.

Wiggins, G. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Journal of Knowledge Based Systems* 19(7):449–458.

Wigram, T. 2002. Indications in music therapy: evidence from assessment that can identify the expectations of music therapy as a treatment for autistic spectrum disorder (ASD): meeting the challenge of evidence based practice. *British Journal of Music Therapy* 16:11–28.

Wu, Y.-S.; r. Chu, W.; Chi, C.-Y.; Wu, D. C.; T.-H.Tsai, R.; and j Hsu, J. Y. 2009. The power of words: Enhancing music mood estimation with textual input of lyrics. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, 1–6.

Full-FACE Poetry Generation

Simon Colton[1], **Jacob Goodwin**[1] **and Tony Veale**[2]
[1]Computational Creativity Group, Department of Computing, Imperial College London, UK. ccg.doc.ic.ac.uk
[2] School of Computer Science and Informatics, University College Dublin, Ireland. afflatus.ucd.ie

Abstract

We describe a corpus-based poetry generation system which uses templates to construct poems according to given constraints on rhyme, meter, stress, sentiment, word frequency and word similarity. Moreover, the software constructs a mood for the day by analysing newspaper articles; uses this to determine both an article to base a poem on and a template for the poem; creates an aesthetic based on relevance to the article, lyricism, sentiment and flamboyancy; searches for an instantiation of the template which maximises the aesthetic; and provides a commentary for the whole process to add value to the creative act. We describe the processes behind this approach, present some experimental results which helped in fine tuning, and provide some illustrative poems and commentaries. We argue that this is the first poetry system which generates examples, forms concepts, invents aesthetics and frames its work, and so can be assessed favourably with respect to the FACE model for comparing creative systems.

Introduction

Mainstream poetry is a particularly human endeavour: written by people, to be read by people, and often about people. Therefore – while there are some exceptions – audiences expect the opportunity to connect on an intellectual and/or emotional level with a person, which is often the author. Even when the connection is made with characters portrayed in the poem, the expectation is that the characters have been written from a human author's perspective. In the absence of information about an author, there is a default, often romantic, impression of a poet which can be relied upon to provide sufficient context to appreciate the humanity behind a poem. Using such an explicit, default or romantic context to enhance one's understanding of a poem is very much part of the poetry reading experience, and should not be discounted.

Automated poetry generation has been a mainstay of computational creativity research, with dozens of computational systems written to produce poetry of varying sophistication over the past fifty years. In the literature review given below, it is clear that the emphasis has been almost entirely on artefact generation, i.e., producing text to be read as poetry, rather than addressing the issues of context mentioned above. Therefore, without exception, each of these systems has to be seen as an assistant (with various levels of autonomy) for the system's user and/or programmer, because that person provides the majority of the context. This is usually achieved by supplying the background material and templates; or curating the output; or writing technical papers to describe the sophistication of the system; or writing motivational text to enhance audience understanding, etc.

While such poetry assistants are very worthwhile, we aim instead to build a fully autonomous computer poet, and for its poems to be taken seriously in full disclosure of the computational setting. The first step towards this aim is to acknowledge that the poems generated will not provide the usual opportunities to connect with a human author, as mentioned above. A second step therefore involves providing a suitable substitute for the missing aspects of humanity. To partly address this, we have built a system to construct poems via a corpus-based approach within which existing snippets of human-written text are collated, modified and employed within the stanzas of poem templates. In particular, in the *Corpus-Based Poetry* section below, we describe how a database of similes mined from the internet, along with newspaper articles can be used to generate poems.

A third step, which also addresses the missing human element to some extent, involves providing a context within which a poem can be read. Software may not be able to provide an author-centric human context, but it can provide a context which adds value to a poem via an appeal to aspects of humanity, in particular emotions. In the section below entitled *Handing over High-Level Control*, we describe how the software uses a corpus of newspaper articles to (a) determine a mood for the day in which it is writing a poem, which it uses to (b) generate an aesthetic and templates within which to generate poems, then (c) selects and modifies corpus material to instantiate the templates with, ultimately producing poems that express the aesthetic as best as possible. To communicate aspects of the context, a final step has been to enable it to provide a commentary on its work, which can be referred to by readers if required.

In the *Illustrative Results* section, we present some poems along with the commentaries generated alongside them. Given our aim for the poems to be considered in full disclosure of their computational context, along with various other arguments given in (Pease and Colton 2011b), we believe it is not appropriate to use Turing-style tests in the evaluation of this poetry generation project. Instead, we turn initially to the FACE descriptive model described in (Colton, Charnley, and Pease 2011) and (Pease and Colton 2011a), which suggests mechanisms for evaluating software in terms of the types of generative acts it performs. In the *Conclusions and Future Work* section below, we argue that we can reasonably claim that our software is the first poetry generator to achieve ground artefact generation of each of the four types prescribed in the FACE model, namely: examples, concepts, aesthetics and framing information. We believe that such full-FACE generation is the bare minimum required before we can start to properly assess computer poets in the wider context of English literature, which is a longer term aim for this project. We describe how we plan to increase the autonomy and sophistication of the software to this end.

Background

Perhaps the first computational poetry generator, the *Stochastische Texte* system (Lutz 1959), sought recognisably Modernist literary affect using a very small lexicon made from sixteen subjects and sixteen predicates from Kafka's *Das Schloß*. The software randomly fitted Kafka's words into a pre-defined grammatical template. Poems by software in this genre – where a user-selected input of texts are processed according to some stochastic algorithm and assigned to a pre-defined grammatical and/or formal template – have been published, as in (Chamberlain and Etter 1984) and (Hartman 1996), and they remain popular on the internet, as discussed in (Addad 2010). Such constrained poetry generation follows on from the OULIPO movement, who inaugurated the poetics of the mathematical sublime with *Cent mille milliards de poèmes* (Queneau 1961), an aesthetic expressed today in digital poems like *Sea and Spar Between* (Montfort and Strickland 2010).

Most of the more sophisticated poetry generation software available on the internet is designed to facilitate *digital poetry*, that is, poetry which employs the new rhetorics offered by computation. For examples, see (Montfort and Strickland 2010), (Montfort 2009) and (Roque 2011). We distinguish this from a stronger definition of *computationally creative* poetry generation, where an autonomous intelligent system creates unpredictable yet meaningful poetic artefacts. Recent work has made significant progress towards this goal; in particular, the seminal evolutionary generator McGONAGALL (Manurung 2004) has made a programmatic comeback, as described in (Rahman and Manurung 2011) and (Manurung, Ritchie, and Thompson 2012). This work is based on the maxim that "deviations from the rules and norms [of syntax and semantics] must have some purpose, and not be random" and the authors specify that *falsifiable* poetry generation software must meet the triple constraints of grammaticality, meaningfulness and poeticness. McGONAGALL, the most recent incarnation of the WASP system described in (Gervás 2010), and the system described in (Greene, Bodrumlu, and Knight 2010), all produce technically proficient poems satisfying these criteria.

There are a number of systems which use corpora of human-generated text as source material for poems. In particular, (Greene, Bodrumlu, and Knight 2010) and (Gervás 2010) rely on small corpora of already-poetic texts. The Hiveku system (www.prism.gatech.edu/˜aledoux6/hiveKu/) uses real-time data from Twitter; (Wong and Chun 2008) use data from the blogosphere and search engines; and (Elhadad et al. 2009) have used Project Gutenberg and the Google n-grams corpus. These approaches all rely on user-provided keywords to start a search for source material and seed the poetry generation process. The haikus produced by the system described in (Wong and Chun 2008) using Vector Space manipulation demonstrate basic meaningfulness, grammaticality and poeticness, but are tightly constrained by a concept lexicon of just 50 keywords distilled from the most commonly used words in the haiku genre. The Electronic Text Composition (ETC) poetry engine (Carpenter 2004) is one of a few generators to use a very large corpus of everyday language in the service of meaningful poetry generation. Its knowledge base is constituted from the 85 million parsed words of the British National Corpus, which has been turned into a lexicon of 560,000 words and 49 million tables of word associations. ETC generates its own poem templates, and its corporal magnitude encourages surprising, grammatically well-formed output. A dozen of its poems were published under a pseudonym (Carpenter 2004).

The creative use of figurative language is essential to poetry, and is a notion alluded to, but declared beyond the scope of (Manurung, Ritchie, and Thompson 2012). One example of prior research in this direction is the system of (Elhadad et al. 2009), which generates haiku, based on a database of 5,000 empirically gathered word association norms. It was reported that this cognitive-associative source principle produced better poems than a WordNet based search. Other aspects of small-scale linguistic creativity relevant to poetry generation include the production and validation of neologisms (Veale 2006), and the elaborations of the Jigsaw Bard system (Veale and Hao 2011), which works with a database of simple and ironic similes to produce novel compound similes.

Concerning aspects of computational poetry at a higher level than example generation, the WASP system can be considered as performing concept formation, as it employs a cultural meta-level generation process, whereby a text is considered and evolved by a group of distinct expert subsystems "like a cooperative society of readers/critics/editors/writers" (Gervás 2010). However, the results of the iterative evaluation are not presented with the final output, and the system does not generate the aesthetics it evaluates, which are "strongly determined by the accumulated sources used to train the content generator", in a similar way to (Greene, Bodrumlu, and Knight 2010) and (Díaz-Agudo, Gervás, and González-Calero 2002).

To the best of our knowledge, there are no poetry generation systems which produce an aesthetic framework within which to assess the poems they produce. Moreover, none of the existing systems provide any level of context for their poetry. In general, the context within which the poems can be appreciated is either deliberately obfuscated to attempt to facilitate an objective evaluation, as per Turing-style tests, or is provided by the programmer via a technical paper, foreword to an anthology, or web page. There are a myriad of websites available which generate poems in unsophisticated ways and then invite the reader to interpret them. For instance, the RoboPoem website (www.cutnmix.com/robopoem), states that: "A great deal of poetry mystifies it's readers: It may sound pretty, but it leaves you wondering 'what the hell was that supposed to mean?"' then extols the virtue of randomly generating mysterious-sounding poetry. This misses the point that poets use their intellect to write poetry which might need unpicking, in order to better convey a message, mood or style. A random sequence of words is just that, regardless of how poem-shaped it may be. The RoboPoet (the smartphone version of which enables you to "generate nonsensical random poems while waiting at the bus-stop") and similar programs only serve to highlight that people have an amazing capacity to find meaning in texts generated with no communicative purpose.

Corpus-Based Poetry Generation

As we see above, using human-produced corpora is common in computational poetry. It has the advantages of (a) helping to avoid text which is utterly un-interpretable (as most human-written text is not like this), which would likely lead to a moment where readers remember that they are not reading the output of a fully intelligent agent, and (b) having an obvious connection to humanity which can increase the semantic value of the poem text, and can be used in framing information to add value to the creative act. However – especially if corpora of existing poems are used – there is the possibility of accusations of plagiarism, and/or the damning verdict of producing pastiches inherent with this approach. Hence, we have chosen initially to work with very short phrases (similes) mined from the internet, alongside the phrases of professional writers, namely journalists writing for the British Guardian newspaper. The former fits into the long-standing tradition of using the words of *the common man* in poetry, and the latter reflects the desire to increase quality while not appropriating text intended for poems.

The simile corpus comes from the Jigsaw Bard system[1] which exploits similes as *readymades* to drive a "modest form of linguistic creativity", as described in (Veale and Hao 2011). Each simile is provided with an evidence score that indicates how many times phrases expressing the simile were seen in the Google n-gram corpus[2] from which they were mined. There are 21,984 similes in total, with 16,579 having evidence 1 and the simile "As happy as a child's life" having the most evidence (1,424,184). Each simile can be described as a tuple of $\langle object, aspect, description \rangle$, for instance $\langle child, life, happy \rangle$. Our database of Guardian newspaper articles was produced by using (i) their extensive API[3] to find URLs of articles under headings such as *World* and *UK* on certain days (ii) the Jericho package[4] to extract text from the web pages pointed to by the URLs, and (iii) the Stanford CoreNLP package[5] to extract sentences from the raw text. As of writing, the database has all 12,820 articles made available online since 1st January 2012, with the *World* section containing the most articles at 1,232.

In addition to the corpora from which we directly use text, we also employ the following linguistic resources:
[1] The CMU Pronunciation Dictionary[6] of 133,000 words.
[2] The DISCO API[7] for calculating word similarities, using a database of distributionally similar words (Kolb 2008).
[3] The Kilgariff database of 208,000 word frequencies (Kilgarriff 1997), mined from the British National Corpus[8]. This database also supplies detailed part-of-speech (POS) tagging for each word, with major and minor tags given.
[4] An implementation[9] of the Porter Stemmer algorithm (Porter 1980) for extracting the linguistic stems of words.
[5] The well known WordNet[10] lexical database.
[6] An implementation[11] of the Text Rank keyphrase extraction algorithm (Mihalcea and Tarau 2004).
[7] The Afinn[12] sentiment dictionary, containing 2,477 words tagged with an integer from -5 (negative affect) to 5 (positive affect). We expanded this to a dictionary of around 10,000 words by repeatedly adding in synonyms for each word identified by WordNet.

Poetry generation is driven by a four stage process of: *retrieval, multiplication, combination* and *instantiation*. In the first stage, similes are retrieved, according to both sentiment and evidence. That is, a range of relative evidence values can be given between 1% (very little evidence) and 100% (the most evidence) along with a sentiment range of between -5 and 5 (as per [7]). Note that the sentiment value of the $\langle object, aspect, description \rangle$ triple is calculated as the average of the three words, with a value of zero being assigned to any word not found in [7]. Constraints on word frequencies, as per [3], can also be put on the retrieval, as can constraints on the pronunciation of words in the simile, as per [1]. In addition, an article from the Guardian can be retrieved from the database (with details of how the article is chosen given later), keyphrases can be extracted using [6], and these can be further filtered to only contain relatively unusual words (as per [3]), which often contain the most pertinent information in the article.

Simile Multiplication

In the second stage, variations for each simile are produced by substituting either an object, aspect or description word, or any combination thereof. The system is given a value n for the number of variations of given simile G required, plus a substitution **scheme** specifying which parts should be substituted, and a choice of three substitution **methods** to use. Denoting G_o, G_a and G_d for the object, aspect and description parts of G, the three methods are:

(d) Using DISCO [2] to retrieve the n most similar words to each word, as determined by that system.

(s) Using the corpus of similes to retrieve the n most similar words to each word. This is calculated with reference to G and the whole corpus. For instance, suppose G_d is to be substituted. Then all the matching similes, $\{M^1, \ldots, M^k\}$, for which $M^i_o = M^i_o$ or $M^i_a = M^i_a$ are retrieved from the database. The set M^i_d of words for $i = 1, \ldots, k$ are collated, and a repetition score $r(M^i_d)$ for each one is calculated as: $r(M^i_d) = |\{j \in 1, \ldots, k : M^j_d = M^i_d\}|$. Informally, for a potential substitute, this method calculates how many similes it appears in with another word from G. The n words with the highest score are used as substitutes.

(w) Using Wordnet [5] to retrieve the n most frequent synonyms of each word, with frequency assessed by [3].

Each variation, V, of G is checked and pruned if (i) the simile exists already in the database, (ii) the major POS of either V_o, V_a or V_d differs from the corresponding part of G, or (iii) the overall sentiment of V is positive when that of G is negative (or vice-versa). To determine the yield of variations each method can produce, we ran the system to

[1] afflatus.ucd.ie/jigsaw
[2] books.google.com/ngrams/datasets
[3] www.guardian.co.uk/open-platform
[4] jericho.htmlparser.net
[5] nlp.stanford.edu/software
[6] www.speech.cs.cmu.edu/cgi-bin/cmudict
[7] www.linguatools.de/disco/disco_en.html
[8] www.natcorp.ox.ac.uk
[9] www.tartarus.org/~martin/PorterStemmer
[10] wordnet.princeton.edu
[11] lit.csci.unt.edu
[12] fnielsen.posterous.com/tag/afinn

Scheme	d	s	w	Average
001	61.68	23.16	0.04	28.29
	2.02	1.68	3.22	2.31
010	59.04	25.58	4.50	29.71
	2.27	1.99	2.09	2.12
100	37.06	28.38	2.26	22.57
	2.08	1.75	1.93	1.92
011	44.50	47.78	0.26	30.85
	2.27	2.25	3.35	2.62
101	39.68	41.94	0.10	27.24
	2.25	1.89	2.83	2.32
110	37.06	40.54	5.84	27.81
	2.21	2.02	2.21	2.15
111	27.84	39.44	0.01	22.43
	2.40	2.10	2.67	2.39
Average	43.69	35.26	1.86	26.94
	2.21	1.95	2.61	2.26

Table 1: Top lines: the average yield (to 2 d.p) of variations returned by each method and substitution scheme when asked to produce 100 variations for 100 similes. Bottom lines: the average interpretation level required for similes generated by the method and scheme. Note that 101 means that the object and description were substituted but not the aspect in the $\langle o, a, d\rangle$ simile triple, etc.

generate 100 variations – before pruning – of 100 randomly chosen similes, for each method, with every possible substitution scheme. The results are given in table 1. We see that the d and s methods yield high numbers of variations, but the w method delivers very low yields, especially when asked to find substitutes for G_d. This is because the number of synonyms for a word is less than the number of similar words, and the number of synonyms for adjectives is particularly low. Unexpectedly, replacing more parts of a simile does not necessarily lead to more similes. On inspection, this is because the increase in degrees of freedom is balanced by an increase in likelihood of pruning due to (i), (ii) or (iii) above.

In addition to observing the quantity of variations produced, we also checked the variations qualitatively. We noticed subjectively that, even out of context, certain variations were very easy to interpret, others were more challenging, and for some no suitable interpretation could be derived. For each of the methods d, s and w, we extracted 1,000 variations from those produced for table 1, and the first author subjectively hand-annotated each variation with a value 1 to 4, with 1 representing obvious similes, 2 representing similes for which an interpretation was more difficult but possible, 3 representing similes which took some time to form an interpretation for, and 4 representing similes for which no interpretation was possible. Some example similes with annotations are given in table 2. On inspection of the level 4 variations, we noted that often the problem lay in the POS-tagging of an adjective as a noun. For instance, in table 2, *kind* is identified as a noun, hence similes with nouns like *form* instead of *kind* are allowed, producing syntactically ill-formed sentences. We plan to rule this out using context-aware POS tagging, available in a number of NLP packages.

The average interpretation level for each of the substitution methods and schemes is given in table 1. We turned this analysis into a method enabling the software to control (to some extent) the level of interpretation required.

Interp. Level	Method Scheme	Variation Original
1	d	as sad as the groan of a widow
	011	as lonely as the moan of a widow
2	s	as deadly as the face of a dagger
	110	as deadly as the sting of a scorpion
3	d	as shallow as the space-time of a fork
	110	as shallow as the curve of a spoon
4	w	as form as the pump of a dove
	011	as kind as the heart of a dove

Table 2: Example simile variations, given with the interpretation level required and the original versions.

Meth.	Bound.	Naïve %	Best %	Best Method
d	1/2	72.00	75.20	RandomForest
s	1/2	60.40	65.80	LogitBoost
w	1/2	68.10	72.00	Bagging
d	2/3	59.20	68.30	OneR
s	2/3	71.20	75.70	RotationForest
w	2/3	63.20	71.10	RandomCommittee
average		65.68	71.35	

Table 3: Ten-fold cross-validation results for the best classifier on the boundary problems for each method.

To do this, given a required interpretation level n for simile variations, pairings of substitution (method, scheme) which produce an average interpretation level between n and $n+1$ in table 1 are employed. So, for instance, if similes of interpretation level 1 are required, the software uses a $(s, 001), (s, 010), (s, 100), (s, 101)$ or $(w, 100)$ pairing to generate them. To increase the performance of the approach, we used the WEKA machine learning system (Hall et al. 2009) to train a predictor for the interpretation levels which could be used to prune any variation predicted to have an interpretation level different to n. To produce the data to do so, we recorded 22 attributes of each of the 3,000 annotated similes, namely: the word frequencies [3] of each part and the minimum, average and maximum of these; the pairwise similarity [2] of each pair of parts, and the min/av/max of these; the pairwise number of collocations of each pair in the corpus of similes and the min/av/max of these; the method used for finding substitutions (d, s or w); whether the object, aspect and/or description parts have been substituted from the original; and the interpretation level.

Unfortunately, using 30 different machine learning methods in WEKA (with default settings for each), the best predictive accuracy we could achieve was 47.3%, using the RotationForest learning method. We deemed this insufficient for our purposes. However, for each variation method, we were able to derive adequate predictors for two associated binary problems, in particular (i) to predict which side of the 1/2 boundary the level of interpretation an unseen simile will be on, and (ii) the same for the 2/3 boundary. The best methods, assessed under 10-fold cross validation, and their predictive accuracy for the boundary problems for the d, s and w variation methods are given in table 3. We found that in each case, a classifier which is significantly better (as tested by WEKA using a paired T-test) than the naïve classifier had been learned, and we can expect a predictive accuracy of around 71% on average. The best learning method was dif-

ferent for each boundary problem, but some methods performed well in general. While not the best for any, the RandomSubspace method was the only one which achieved a significant classifier for all the problems. The Bagging, RotationForest, and RandomForest methods all produced significant classifiers for five of the six problems.

WEKA enables the learned predictors to be used externally, so we implemented a process whereby the generative procedure above produces potential simile variants of a given level, then the result is tested against both boundary predictors appropriate to the method. If it is predicted to fall on the wrong side of either boundary, it is rejected. As a final validation of the process, we generated 300 new simile variations, with 100 of level 1, 2 and 3 each. We mixed them randomly and tagged them by hand as before. Our hand tagging corresponded with what the software expected 82% of the time, which we believe represents sufficient control.

Combination and Instantiation

The third and fourth phases of poetry generation are more straightforward. In the combination phase, similes, variations of them and keyphrases extracted from newspaper articles are combined as per user-given templates. The templates dictate what words in each of a pair of text fragments must match exactly, what the POS tags of these words and others in the fragments must be, and how they are to be combined. Templates often simply pair two phrases together according to certain constraints, to be used in the instantiation phase later. Alternatively, they can provide more interesting ways of producing a compound phrase. The process can be iterated, so that triples, quadruples, etc., can be generated.

As an example, suppose we have the keyphrase "excess baggage" from a newspaper article about travel. This can be matched with the simile "the emotional baggage of a divorce", and presented in various ways, from simple expressions such as "the emotional excess baggage of a divorce", to the more elaborate "Oh divorce! So much emotional excess baggage", as determined by the combination template. It is possible to drop certain words, for instance the keyphrase "gorgeous history" (about a 1980s pop group) and the simile "As gorgeous as the nature of a supermodel" could produce "a supermodel-gorgeous history", and variations thereof. As a final example, keyphrases such as "emotional jigsaw puzzle" (describing a surreal play in a review) can be elaborated by combination with the simile "As emotional as the journey of a pregnancy" to produce: "An emotional jigsaw puzzle, like the journey of a pregnancy".

The retrieval, multiplication and combination stages of the process perform the most important functions, which leaves the instantiation process able to simply choose from the sets of elaborated phrases at random, and populate the fields of a user-given template. Templates allow the extraction of parts of phrases to be interleaved with user-given text, and there are also some final constraints that can be applied to the choice of phrases for the template, in particular to reduce repetition by only choosing sets of phrases where the word stems (constructed by [4]) are different.

In terms of linguistic and semantic constraints, the four stage process is quite powerful, as highlighted with the example poem given in figure 1, produced using a highly constrained search for pairs of similes. We used no simile multiplication here, in order to highlight the linguistic rather than inventive abilities. The circadian aspects of the poem are part of the template, with only the similes provided by the software. We see that the poem contains only straightforward words, because during the retrieval stage, only similes with words having frequencies in the top 5% were retrieved (as determined by [3]). Moreover, the only direct repetition is there by design in the template, and no repetition even of word stems is allowed anywhere else. This was achieved during the instantiation process, which recorded the similes used, and avoided using any word where [4] suggested the same word stem with an existing word in the poem.

Circadian No. 39

Stealthy swiftness of a leopard,
Happy singing of a bird.

In the morning, I am loyal
Like the comfort of a friend.
But the morning grows more lifeless
Than the fabric of a rag.
And the mid-day makes me nervous
Like the spirit of a bride.

Active frenzy of a beehive,
Dreary blackness of a cave.

In the daytime, I am slimy
Like the motion of a snake.
But the sunlight grows more comfy
Than the confines of a couch.
And the day, it makes me tasty
Like the flavor of a coke.

Shiny luster of a diamond,
Homey feeling of a bed.

In the evening, I am solid
Like the haven of a house.
But the evening grows more fragile
Than the mindset of a child.
And the twilight makes me frozen
Like the bosom of a corpse.

Famous fervor of a poet,
Wily movement of a cat.

In the night-time, I am hollow
Like the body of a drum.
But the moonlight grows more supple
Than the coating of an eel.
And the darkness makes me subtle
Like the color of a gem.

Stealthy swiftness of a leopard,
Happy singing of a bird.

Figure 1: An example instantiation of a user-given template.

The poem also has strictly controlled meter and stress. For instance, each two-line stanza firstly uses a simile with $\langle sw, sw, sw \rangle$ pronunciation (where s and w are syllables, with s being the stressed one), and then uses a simile with $\langle sw, sw, s \rangle$ pronunciation. This is achieved during the retrieval stage, which uses the pronunciation dictionary [1] to select only similes of the right form, and the combination process, which puts together appropriate pairs of lines. There is similar regularity in the six-line stanzas. Possibly less obvious is the subtle rhyming at play, with the final phonemes of selected pairs of lines being the same (such as beehive and cave, snake and coke, drum and gem). Moreover, inadvertent rhyming – which can be jarring – is ruled out elsewhere, for instance, snake and couch were constrained to have no rhyming, as were house and child, drum and eel, etc. The rhyming constraints come into play during the combination phase, when sets of lines are collated for final use in the stanzas. Finally, we notice that the stanzas alternate in sentiment during the course of the poem, for instance the line "Happy singing of a bird" in the first two-line stanza, contrasts starkly with the line "Dreary blackness of a cave" in the second. This is also achieved during the combination phase, which can be constrained to only put together similes of certain sentiments, as approximated by [7].

Handing over High-Level Control

We see automated poetry generation as the simultaneous production of an artefact and a context within which that artefact can be appreciated. Normally, the context is provided by the programmer/user/curator, but, as described below, to give more autonomy to the software, we enabled it to provide its own context, situated in the events of the day in which it is writing poems. In order to deliver the context alongside each poem, we also implemented a rudimentary ability to provide a commentary on the poem, and how it was produced, as described in the second subsection below.

Context Generation

In overview, the software determines a mood for the day, then uses this to choose both a Guardian article from which to extract keyphrases which will be combined with simile variations and form lines of the poem, and an aesthetic within which to assess the generated poems. These are then used to produce a set of templates for the four-stage poem generation process described above. Finally, the software instantiates the templates to produce a set of poems, and chooses to output the one which maximises the aesthetic.

As in the automated collage generation of (Krzeczkowska et al. 2010), the software appeals to daily newspaper articles for raw material. We extend that approach by also using the articles to derive a mood, from which an aesthetic is generated. In particular, each of the 12,820 articles in the corpus has been assigned a sentiment value between -5 and 5, as the average of the sentiment of the words in the article, assessed by [7]. Thus, when a poetry generation session begins, the software is able to check the sentiment of the set N of newspaper articles posted during the previous 24 hours, and if it is less than the average, the software determines the mood as *bad*, or *good* otherwise. If the mood is good, then an article, A, from the happiest five articles from N is chosen, with melancholy articles similarly chosen during a bad mood. The keyphrases, $key(A)$, are then extracted from the article, and we denote as $words(A)$ the set of words appearing in $key(A)$. Note that very common words such as "a", "the", "of", etc., are removed from $words(A)$.

As an example, on 17/01/2012, the mood was assessed as bad, and a downbeat article about the Costa Concordia disaster was retrieved. In contrast, on 24/01/2012, the mood was assessed as good, and an article describing the buoyant nature of tourism in Cuba was retrieved, from which keyphrases such as "beach resorts", "jam-packed bar", "seaside boulevard" and "recent sunny day" were extracted using [6]. Note that [6] also returns a relevancy score for each keyphrase, e.g., "recent sunny day" was given a score of 0.48 for relevance, while "jam-packed bar" only scored 0.31.

The mood is sufficient to derive an aesthetic within which to create poems, but this will be projected partly through members of $words(A)$ appearing in the poem, and mood is only one aspect of the nature of a poem. Letting $words(P)$ denote the words in poem P, for more variety, the software can choose from the following four measures:

- **Appropriateness:** the distance between the average sentiment of the words in $words(P)$ from 5 if it is a good mood day, or from -5 if it is a bad mood day.
- **Flamboyance:** the average of $f(w)$ over $words(P)$, where $f(w) = 0$ if $w \in words(A)$ and $f(w) = 1/frequency(w)$ if $w \notin words(A)$, where frequency is calculated by [4].
- **Lyricism:** the proportion of linguistic constraints adhered to by P, with the constraints determined by the set of templates generated for the poem, as described below.
- **Relevancy** to the Guardian article: the average of $rel(w)$ over $words(P)$, where $rel(w) = 0$ if $w \notin words(A)$ and $rel(w)$ is the relevancy [6] of w, if $w \in words(A)$.

The choice of which set of measures, M, to use in the aesthetic for a poem is determined somewhat by A and $key(A)$. In particular, if A is assessed as being in the most emotive 10% of articles ever seen (either happy or sad), then M is chosen as either {Appropriateness} or {Appropriateness, Relevance} in order to give due consideration to the gravity or brevity of the article. If not, and the size of $key(A)$ is less than 20% of the average over the corpus, then it might be difficult to gain relevancy to A in the poem, hence M is chosen as $\{Relevance\}$. In all other cases, M is chosen randomly to consist of either 1 or 2 of the four measures – we found that mixing more than 2 diluted their effect, leading to poems with little discernible style.

The software also generates templates to dictate the structure of the poem. The number of stanzas, z, is taken to be between 2 and 10, with the number dictated by the size of $key(A)$, i.e., larger poems are produced when $key(A)$ is relatively large, compared to the rest of the corpus. The structure of the poem can be *equal*, i.e., of the form $A_1A_2 \dots A_z$ with each stanza A_i being of the same length (chosen randomly between 2 and 6 lines). The structure can also be chosen to be *alternating* of the form $A_1B_2A_3 \dots A_z$ or $A_1B_2A_3 \dots B_z$; or *bookended* of the form $A_1B_2 \dots B_{z-1}A_z$. The choice of structure is currently made randomly, and there is no relationship between pairs of stanzas, except that the templates constrain against the usage of a new phrase (combined from a keyphrase and simile as described above) in the template if one of the words has the same stem as a word in an already-used phrase. As part of the template generation, the software chooses the number of times (between 0 and 5) this constraint is allowed to be broken per phrase, as a level of repetition can add flavour to a poem. Note that the counts per phrase are reset to zero if the software runs out of phrases to add to the template.

If M contains the $Lyricism$ measure, then the templates are also constrained to express some linguistic qualities, which are added at the stanza level. In particular, the line structure of all stanzas of type A is chosen to be either equal, alternating or bookended in the same fashion as the stanza structure, with stanzas of type B also given a structure. This structure allows linguistic constraints to be added. For instance, if a stanza has alternating structure $abab$, the software chooses a single linguistic constraint from: *syllable-count*, *end-rhyme*, *start-rhyme*, and constrains all lines of type a accordingly. It does the same (with a possibly different linguistic constraint) for lines of type b. Note that *syllable-count* means that the two lines should have the same

number of syllables as each other (within a threshold of two syllables), *end-rhyme* means that the two lines should at least end in the same phoneme, with *start-rhyme* similar.

The random nature of the choices to fill in the final poem template ensures variety. In each session, the software generates 1,000 poems, and their scores for each of the measures in M are calculated. The average rank over the measures is taken as an overall rank for each poem, and the highest ranked is presented as the poem for the day. If the templates over-constrain the problem and no poems are produced, then a single constraint is chosen to be dropped, and the session re-started iteratively until poems are produced.

Commentary Generation

In addition to the four stage process of retrieval, multiplication, combination and instantiation, the software chooses a Guardian article, performs sentiment analysis, aesthetic invention, template construction and searches for appropriate poems. While some of these methods are at present rather rudimentary and perhaps a little arbitrary, it is our hypothesis that a well-formed commentary about how the software has produced a poem will provide a context for the poem and add value to the appreciation process, as argued above.

In order for the software to generate the commentary, we re-use the four stage process, but with the retrieval stage sampling not from corpora of human produced text, but rather from a set of recorded statements about how each of the processes worked, and what they produced. In particular, the software records details such as (a) the mood of the day (b) the Guardian article it retrieved and how emotive it was (c) the keyphrases extracted, which sentences they came from, and which were used in the final poem (d) the combinations of keyphrase and similes it produced (e) the nature of the poem structure dictated by the template, (f) the aesthetic weightings used, and (g) what successes and failures it had in instantiating the templates. We have produced by hand a number of (sets of) commentary templates that can present the statements in a supportive way. Currently, the software randomly chooses which set of templates to use to generate the commentary. The software chooses the title for each poem as the keyphrase occurring the most often in the poem, choosing randomly if there is a tie for the most used.

Illustrative Examples

We artificially situated the software in the days from 1/01/2012 to 10/02/2012, and asked it to produce a single poem for each day, along with a commentary. We added the constraint that the poem should be exactly four stanzas in length for presentation purposes in this paper. We curated three for presentation here, in figure 2 below. The commentaries are meant to provide enough context for proper appreciation of each poem, so we will not add detail here to the commentaries of the individual poems. Viewing the entire set of generated poem/commentaries subjectively, we were disappointed by the number of compound sentences available for the templates. Even with large sets of keyphrases extracted from an article, and extensive simile multiplication employed, we found that there were few opportunities for a simile to be used for embellishment, which meant that the software had limited choices for the final poem template, which led to an over-reliance on repeating lines, or using similar lines. More importantly, the differences in the aesthetic evaluations over the 1,000 poems generated for a day were not great, hence the aesthetic generation was driving the production of poems less than we would have liked.

Conclusions and Future Work

We agree with (Pease and Colton 2011b) that Turing-style tests encourage naïvety and pastiche in creative systems. However, eschewing their use leaves a hole regarding proper evaluation of our poetry generation system. Instead, we can turn to the FACE descriptive model put forward in (Colton, Charnley, and Pease 2011) and (Pease and Colton 2011a), which advocates describing a creative system in terms of the *creative acts* it performs, which are in turn tuples of generative acts. The generative acts produce outputs of four types: examples of concepts, concepts themselves, aesthetic measures which can evaluate concept/example pairs, and framing information. Looking at the literature review above, the WASP and Electronic Text Composition systems can be considered as generating concepts, as can any system which generates and employs a statistical model of written or verbal language (such as in Markovian approaches). It does not appear that any system invents aesthetic measures or produces framing information such as a commentary which can be used as a context for the poem. Hence, according to the FACE model, our approach can be considered favourably, as it has processes producing examples (instantiation), concepts (template generation), aesthetics (choosing measures) and framing information (producing commentaries), within the creative act of poem generation. This represents an advance in the state of the art of automatic poetry generation.

It is clear that many aspects of the process presented here are fairly rudimentary, often with random choice substituting a reasoned approach. Our main contribution has been to implement a rounded system which can function on the majority of levels required to be taken seriously as a poet, albeit in a simplistic manner. We plan further enhancements to all of the processes described above, including (i) implementing improved ways to generate phrases for templates, as the yield is currently too low to enable the software to use its more advanced linguistic constraining features (ii) working with other corpora (iii) enabling the software to automatically add higher level structures to poems via the kinds of narratives seen in the circadian poem given above, and (iv) turn the commentary generation processes into full-story telling, which may include the introduction of fictions. After the enhancements, we will work with a poet and explore gaining critical feedback via the publication of anthologies.

While the *imitation-game* aspect of Turing-style tests are not conducive for creativity, we do applaud the usage of *dialogue* they prescribe. Indeed, in the future, we imagine all creative systems being enhanced with a story generator able to produce both static framing information, and to reply with a story to any question asked of it in a dialogue situation. We believe that only with such abilities will software systems be taken seriously as creative entities in the cultural world.

It was generally a bad news day. I read an article in the Guardian entitled: "Police investigate alleged race hate crime in Rochdale". Apparently, "Stringer-Prince, 17, has undergone surgery following the attack on Saturday in which his skull, eye sockets and cheekbone were fractured" and "This was a completely unprovoked and relentless attack that has left both victims shocked by their ordeal". I decided to focus on mood and lyricism, with an emphasis on syllables and matching line lengths, with very occasional rhyming. I like how words like attack and snake sound together. I wrote this poem.

Relentless attack

a glacier-relentless attack
the wild unprovoked attack of a snake

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

the high-level role of eye sockets
a relentless attack, like a bloodhound

It was generally a good news day. I read a story in the Guardian culture section entitled: "South Africa's ANC celebrates centenary with moment in the sun". It talked of south africans, interfaith prayers and monochrome photos. Apparently, "The heroic struggle against a racist regime was remembered: those thousands who sacrificed their lives in a quest for human rights and democracy that took more than eight decades" and "At midnight he watched with amusement as Zuma lit the centenary flame, at the second attempt, with some help from a man in blue overalls marked 'Explosives'". I wanted to write something highly relevant to the original article. I wrote this poem.

Blue overalls

the repetitive attention of some traditional african chants
a heroic struggle, like the personality of a soldier

an unbearable symbolic timing, like a scream
blue overalls, each like a blueberry
some presidential many selfless leaders

oh! such influential presidents
such great presidents
blueberry-blue overalls

lark-blue overalls
a knight-heroic struggle

It was generally a bad news day. I read a story in the Guardian entitled: "Thai police hunt second bomb plot suspect in Bangkok". It talked of suspected bomb plotters, lebanese men and travel alerts. Apparently, "Sketches released late on Friday night by Thai police showed the suspect as a white Middle-Eastern man with short hair and stubble, around 1.8m (5ft 9in) tall". It's a serious story, but I have concentrated on flourishes today. I wrote this poem.

Foreign embassies

the wiry militant arm of a doorman
a white middle-eastern man, like a snowball
spaceship-foreign embassies
foreign embassies, each like a stranger

an impersonal suvarnabhumi international airport
a white middle-eastern man, like the surface of a porcelain
the sturdy design of a bangkok post

foreign embassies, each like a spaceship
an impersonal suvarnabhumi international airport
stranger-foreign embassies
the stout engineering of a bangkok post

a white middle-eastern man, like the skin of an earthenware
foreign embassies, each like a stranger
spaceship-foreign embassies

Figure 2: Illustrative poems and commentaries. For the Guardian articles on which these poems are based, please see: www.guardian.co.uk, followed by: /uk/2012/feb/09/police-race-hate-crime-rochdale /world/2012/jan/08/south-africa-anc-centenary /world/2012/jan/15/thai-second-bomb-suspect-bangkok

Acknowledgements

This work has been funded by EPSRC grant EP/J004049. Many thanks to the reviewers for their insightful comments.

References

Addad, E. 2010. Interactive poetry generation systems: an illustrated overview. netpoetic.com/2010/10/interactive-poetry-generation-systems-an-illustrated-overview/.

Carpenter, J. 2004. Electronic text composition project. http://slought.org/content/11199.

Chamberlain, W., and Etter, T. 1984. *The Policeman's Beard is Half-Constructed: Computer Prose and Poetry*. Warner Books.

Colton, S.; Charnley, J.; and Pease, A. 2011. Computational creativity theory: the FACE and IDEA models. In *Proceedings of the 2nd International Conference on Computational Creativity*.

Díaz-Agudo, B.; Gervás, P.; and González-Calero, P. 2002. Poetry generation in COLIBRI. *Advances in Case-Based Reasoning* 2416.

Elhadad, M.; Gabay, D.; Goldberg, Y.; and Netzer, Y. 2009. Gaiku: Generating haiku with word associations norms. In *Proc. of the Workshop on Computational Approaches to Linguistic Creativity*.

Gervás, P. 2010. Engineering linguistic creativity: Bird flight and jet planes. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*.

Greene, E.; Bodrumlu, T.; and Knight, K. 2010. Automatic analysis of rhythmic poetry with applications to generation and translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. 2009. The WEKA data mining software: An update. *SIGKDD Explorations* 11(1).

Hartman, C. 1996. *Virtual Muse: Experiments in Computer Poetry*. Wesleyan University Press.

Kilgarriff, A. 1997. Putting frequencies in the dictionary. *International Journal of Lexicography* 10 (2).

Kolb, P. 2008. DISCO: A multilingual database of distributionally similar words. In *Proceedings of KONVENS*.

Krzeczkowska, A.; El-Hage, J.; Colton, S.; and Clark, S. 2010. Automated collage generation – with intent. In *Proceedings of the 1st International Conference on Computational Creativity*.

Lutz, T. 1959. Stochastische texte. *Augenblick* 4(1).

Manurung, R.; Ritchie, G.; and Thompson, H. 2012. Using genetic algorithms to create meaningful poetic text. *JETAI* 24(1):43–64.

Manurung, H. 2004. An evolutionary algorithm approach to poetry generation. *PhD. Thesis*, University of Edinburgh.

Mihalcea, R., and Tarau, P. 2004. Textrank: Bringing order into texts. In *Proc. of the Conference on Empirical Methods in NLP*.

Montfort, N., and Strickland, S. 2010. Sea and spar between. http://blogs.saic.edu/dearnavigator/winter2010/nick-montfort-stephanie-strickland-sea-and-spar-between/.

Montfort, N. 2009. The ppg256 series of minimal poetry generators. In *Proceedings of Digital Arts and Culture 2009*.

Pease, A., and Colton, S. 2011a. Computational creativity theory: Inspirations behind the FACE and IDEA models. In *Proceedings of the 2nd International Conference on Computational Creativity*.

Pease, A., and Colton, S. 2011b. On impact and evaluation in computational creativity: A discussion of the Turing test and an alternative proposal. In *Proc. AISB symp. on AI and Philosophy*.

Porter, M. 1980. An algorithm for suffix stripping. *Program* 14 (3).

Queneau, R. 1961. *Cent mille milliards de poèmes*. Gallimard.

Rahman, F., and Manurung, R. 2011. Multiobjective optimization for meaningful metrical poetry. In *Proceedings of the 2nd International Conference on Computational Creativity*.

Roque, A. 2011. Language technology enables a poetics of interactive generation. *The Journal of Electronic Publishing* 14.

Veale, T., and Hao, Y. 2011. Exploiting readymades in linguistic creativity: A system demonstration of the jigsaw bard. In *Proc. of the 49th Annual Meeting of the ACL*.

Veale, T. 2006. Tracking the lexical Zeitgeist with Wordnet and Wikipedia. In *Proceedings of the 17th European Conference on Artificial Intelligence*.

Wong, M., and Chun, A. 2008. Automatic haiku generation using VSM. In *Proceedings of ACACOS*.

Illustrating a Computer Generated Narrative

Rafael Pérez y Pérez, Nora Morales, Luis Rodríguez
División de Ciencias de la Comunicación y Diseño
Universidad Autónoma Metropolitana, Cuajimalpa
Av. Constituyentes 1054 C. P. 11950, México D. F.
{rperez/nmorales/lrodriguez}@correo.cua.uam.mx

Abstract

This work describes a computer model that generates visual narratives. It is part of a research project on narrative generation. A visual narrative is defined as a sequence of pictorial-scenes; each scene contains characters, locations and symbols representing dramatic tensions. A computer generated plot is transformed into a visual narrative by converting each textual action into a pictorial scene. We present details of the composition process and explain how the graphic elements employed to produce a coherent narration are generated. We describe the questionnaire that we employed to evaluate the system, discuss the results and outline future developments.

Introduction

Narrative is a fundamental manifestation of human culture. "Most scholars now see narrative… and a host of rhetorical figures not as 'devices' for structuring or decorating extraordinary texts but instead as fundamental social and cognitive tools" (Eubanks 2004). Traditionally, the word "narrative" has been understood as a kind of synonymous of written text. However, many forms of storytelling (and knowledge) are visual; that is, "what we see is as important, if not more so than what we hear or read" (Rose, 2001:1). The use of images in the construction of narratives has given origin to the concept of Visual Narrative. Thus, following McCloud, in this work visual narrative is defined "as juxtaposed pictorial and other static images in deliberate sequence, intended to convey information and/or to produce an aesthetic response in the viewer" (McCloud, 1993).

The processes involved in the codification and understanding of visual stories have been the subject of study in the field of psychology for some time (see e.g. Arnheim, 1969) and have firmly established the links between thought and perception. The human ability to organize our experiences in the form of stories or narrative structures has been called Narrative Intelligence (Blair and Meyer, 1997). Narrative Intelligence has also been defined as "the human ability and perhaps even compulsion to make sense of the world through narrative and storytelling" (Mateas and Sengers, 1999). In this way, "Human narrative intelligence might have evolved because the structure and format of narrative is particularly suited to communicate about the social world." (Dautenhahn, 2001). Thus, we envision Visual Narrative as a form of Narrative Intelligence.

Because of its importance in shaping human experience and knowledge, it is not surprising that AI researchers have developed a substantial amount of work related to understanding stories and on how to generate them. One of the common aspects of these efforts is its inherent interdisciplinary approach. Research on AI and narrative has drawn on ideas and theories from different fields such as art, cultural studies, drama, psychology and more recently design. As result from this interdisciplinary work, we can distinguish three main results: Narrative is now recognized as a source for informing System Design; research paradigms and methodologies that address complex questions have been developed and validated; the relationship between AI, Computational Creativity and the Humanities has proven enriching and useful.

The work presented in this paper is part of a research project in narrative generation. We have developed a computer model of creative writing called E-R; a program called MEXICA (Pérez y Pérez and Sharples 2001) is an implementation of such a model. The purpose of this work is to expand our plot-generation model with mechanisms that allow illustrating its textual outputs to produce visual narratives. We refer to this new module as the Visual Narrator. This paper describes our first prototype. Although it is possible to find computer systems that generate or evaluate images (e.g. Norton et al. 2011; Colton 2011), or systems where a visual portrayal of characters plays an important role (e.g. Riedl et al. 2008; Cassell 2001; Rickel and Johnson 1999) as far as we know this is the first plot generator capable of illustrating its own output. It is worth noticing that, as antecedents of this work, we published a paper that employs animations to represent computer-generated daydreams (Pérez y Pérez at al. 2007) and a grammar that generates pre-Hispanic images (Álvarez et al. 2007).

Our computerised storyteller has the following characteristics. It generates fictional narratives about pre-Columbian cultures. This seems just adequate because "The history of sequential art could be track to pre-Columbian picture

manuscripts since they were pictorial representations painted over strips that convey a story" (McCloud, 1993:9). The system includes 16 predefined characters, amongst them: *Tlatoani* (the ruler), *Jaguar Knight*, *Princess*, *Enemy*, *Fisherman*, and so on. It also includes 9 possible locations, e.g. *Chapultepec Forest*, *Popocateptl volcano*, *Tenochtitlan City*. The system generates a sequence of actions representing plots; the following lines are an example: *the Enemy kidnapped the Princess; Jaguar Knight found the Princess; Jaguar Knight and Enemy fought; Enemy ran away; Jaguar Knight rescued the Princess*; and so on. Once the narrative has been finished the system substitutes the sequence of actions with predefined texts. So, the action where the Knight and the Enemy fought in the previous example is substituted by "Suddenly, Jaguar Knight and Enemy were involved in a violent fight". The same happens for all actions.

One of the core characteristics of our computer model of creative writing is the idea that plots can be represented as groups of emotional links and tensions between characters that progress over time (Pérez y Pérez 2007). The current version of the storyteller system includes two emotional links, brotherly love and amorous love, and seven dramatic tensions: when a character is killed (Actor dead); when the life of a characters is at risk (Life at risk); when the health of a character is at risk (Health at risk); when a character is made a prisoner (Prisoner); when two characters are in love with a third character (Love competition); when a character hates and loves another character (Clashing emotions); when one character hates another character and both are positioned in the same location (Potential danger). Each time an action is performed within the story the current set of emotional links and tensions is updated. Thus, our storyteller not only produces plots but also generates detailed information about the emotional links and tensions between characters for each action in the tale.

The Visual Narrator recollects all these information to generate a visual narrative. As mentioned earlier, in this work a visual narrative is defined as a sequence of pictorial-scenes. A scene is a composition made up of images representing one or two characters, a location and a tension. So, the Visual Narrator transforms a plot into a Visual Narrative by converting each textual action into a pictorial scene. The process of composing a scene involves:

1) Building characters. We provide the system with a group of primitive graphic elements that represent different parts of the character's picture. Primitives include body parts, clothes, accessories and emotional and tensional facial expressions. We developed a grammar that drives the construction of the image.

2) Choosing a glyph that represents the core active tension in the story. We have defined a set of glyphs that represents each of the possible tensions within a story.

3) Choosing an image that represents the current location of the characters involved in the action. We provide the system with images representing all possible locations within a story.

4) Putting together all these elements in a scene.

These four points summarise the core characteristics of the Visual Narrator.

Following the theory of narrative from the structuralism point of view (Seymour, 1997), every narrative has two elements: a discourse (the means by which the content is communicated, the set of actual narrative statements) and a story or what is portrayed (the content, events, characters and context or setting). The text is produced by one person and is meant to be read by another; the proper understanding of the narration, requires that both share the same code (Acaso, 2009; Eco, 2005). However, "any text can be interpreted infinitely" (Eco, 1991). In Peirce's words, this is considered as the Dynamical Interpretant: "The Dynamical Interpretant is whatever interpretation any mind actually makes of a sign" (cited in Atkin, 2008:66). In this way, our main concern in this work is to evaluate if the code produced by our computer model satisfies (at least partially) this requirement of proper understanding of the narration. That is, if the visual composition produced by the computer model resembles similar ideas to those produced by its textual counterpart.

This paper is organised as follows: the second section describes the features of the pre-Columbian iconography relevant for this project; the third sections provides details of the composition process; the fourth section describes the evaluation of the system; the fifth section includes the discussion.

Characteristics of Pre-Columbian Iconography and graphic conventionalism

In the following lines we describe some of the graphic elements and conventions used in pre-Columbian codices that inspired the creation of our visual narrative. For this work we mainly considered the Boturini Codex (Galarza J. and Libura, 2004) and the following codices from the Borgia group: Vaticanus and Laud (Galarza, J. 1997), Borgia and Nuttall, (Mohar, L. 1997) and Moctezuma (Lopez A. 1999).

1) *Human figures and objects*. In the codices human figures are line drawn in black and white, in an abstract-graphical iconic level. Their body positions, either sitting or standing, are always represented in profile, facing right or left and never in a front view. Objects and animals are drawn employing the same graphic conventions.

2*) Social Hierarchy*. Pre-Hispanics had a very complex and hierarchical social structure, which was represented in their codices. A strict dressing code and the use of specific ornaments show their status within the group. Attires, ornaments, a stick held by human hands, or triangle tiara crowning a head are symbols representing a high rank within the hierarchy. Hairstyle is another sign of social status, gender and role distinction. Priests are always represented with the whole body or parts of their face painted in black and their hair tied back in a ponytail. A regular representation of women's attire is a *huipil* or skirt and a *quechquemetl* or bust over her waist. In certain codices it is

common to illustrate women showing their breasts. In this way, an image depicting a subject wearing a *tilma* or cape and sandals indicate that the person is a member of the nobility; by contrast, wearing a loincloth without sandals indicates that the person belongs to the group of common people.

3) *Locations*. Locations are represented by a chain of iconic images. For example, the representation of Chapultepec forest is composed by a stylised image of a hill with a grasshopper on the top and a wavy line coming out from its base. The pictogram could be read as "The big hill of the grasshoppers, where the water springs".

4) *Representing emotions and death.* Pre-Hispanic artists depicted emotions in their works. For example, in the Boturini codex there is a passage that shows a group of people crying, wearing dirty clothes. The action of crying is represented by stylised tears in the eyes of the characters. Such tears have an amorphous wavy shape that ends with a circle or oval. In this case, the weeping, reinforced by the dirty clothes, conveys a message of suffering. Other codices, like the Borgia and the Vaticanus, employ mouth and eyes gesticulations to produce richer facial expressions. Finally, a human image with her eyes closed represents death.

We would like to point out that our visual narrative is inspired by pre-Columbian iconography. However, we do not attempt to reproduce it or contribute to its understanding. We leave that to the experts in the area. We only employ such iconography to provide a framework to our research in computational creativity. Thus, some of the images presented in this work are free interpretations made by the authors.

Composition Process

The following lines describe the four process involved in the composition process.

Character Generation

Characters' portraits are the result of bringing together 9 layers of basic images or primitives (as we also refer to them). Each layer groups similar elements. Layer zero includes left arms; layer 1 includes bodies with legs; layer 2 includes heads; layer 3 includes eyes and emotional expressions; layer 4 includes hairstyles; layer 5 includes clothes; layer 6 includes right arms; layer 7 includes ornaments; layer 8 includes weapons and tools (see figure 1). Layers can include any number of primitives, although in some cases it is important to have at least one. The amount of different portrayals that the system can create depends on the number of available images. We have 4 types of characters: males standing, males sitting, females standing and females sitting. All figures can be painted facing east or west. In this work we only present males and females standing and employed 272 primitives.

We classify images in two types: universals and specifics. Universals are used in the construction of any character while specifics are only employed in the construction of those personae they were designed for. In other words, some features might appear in all portrayals while others are specific to a single character. For example, every person can use a shell but only the fisherman can have a fishing-net.

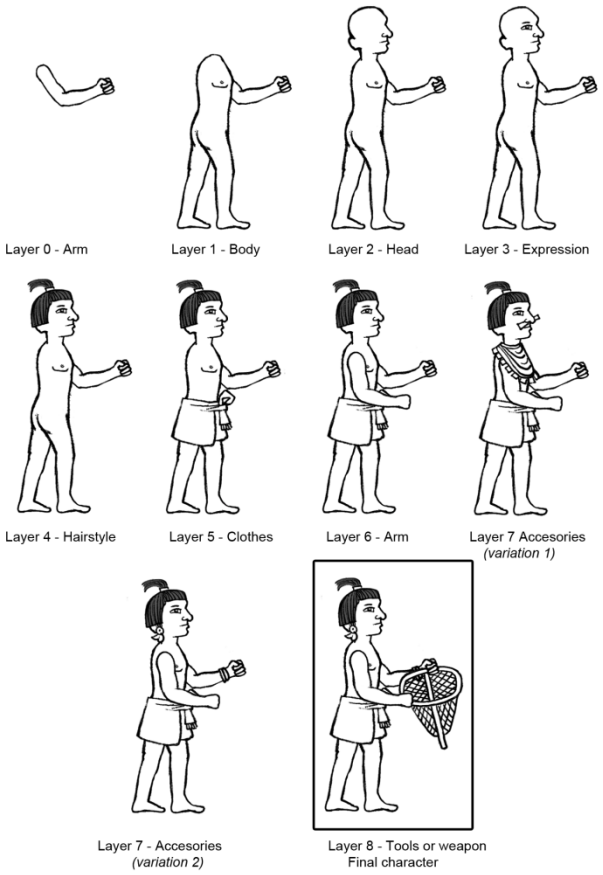


Figure 1. Construction of the character fisherman.

Thus, the Visual Narrator portrays a character by selecting one image from each layer and then painting all them in a canvas. For this work we implemented such a selection process as a random function to provide variety and surprise. In future versions we might include some constraints that help to select the images based on the necessity of the narrative. The user of the Visual Narrator can define in a text file a set of rules to associate one image in a concrete layer with others in different layers. In this way, it is possible to associate particular types of clothes with particular types of ornaments, or extended arms with specific weapons, and so on. For the experiments in this work we defined 25 rules. Figure 1 illustrates the process of characterising a fisherman; it shows two possible options for representing accessories. Figure 3 shows the portraits of a Jaguar Knight, a Princess and an Enemy.

The Visual Narrator depicts characters with emotional expressions. Our automatic storyteller generates information regarding emotional links and tensions between

characters. In this work we only represent tensions. Thus, a person can be responsible of triggering a tension or can be a victim. For example, if the enemy kidnaps the princess, the Enemy triggers the tension Prisoner, and the Princess is the captive. We refer to the former as the giver and to the latter as the receiver. As part of our work we have developed giver and receiver facial emotional expression for each of the possible tension that can be triggered during the development of a narrative. Examples of these expressions are wide open eyes, a wide open mouth and tears (e.g. figure 5a shows a princess crying).

The Visual Narrator analyses the narrative in order to determine which tensions should be represented in the scene. Deciding what tension to characterize is a complex task because the Visual Narrator must figure out which of the active tensions is the most appropriate to be represented in the current scene, for how long it should be visually represented, when it is necessary to reintroduce a tension, and so on. In this way, our implementation resembles those flip-a-face books or board books, where a person can create several different characters combining different predefined elements. Various videogames employ similar tools. Thus, the Visual Narrator models some of the decisions that humans takes to represent emotional characters when using flip-a-face like tools.

Building a Scene

Scenes are comprised of three elements: a location, characters and a glyph representing a tension.

We have 9 possible locations: Texcoco Lake, Popocateptl Volcano, Tlatelolco Market, Palace, Tenochtitlan City, Temple, Chapultepec Forest, Jail and Uncivilized Land. The representations of these locations are inspired by pre-Hispanic codices. For example, figure 5 shows the representation of Chapultepec Forest. Chapultepec means grasshopper's hill in Nahuatl; notice on the right of figure 5 the stylised image of a hill with a grasshopper on the top.

As a result of performing a story-action, one o mores tensions can be triggered or deactivated within a narrative. For example, if the Enemy wounds the Jaguar Knight the tension life at risk is triggered. We have designed glyphs to represent them. Figure 2 shows some examples. Currently, a scene only includes a single tension. When several tensions are active at the same time, the Visual Narrator needs to choose one to be painted in the composition. So, we have assigned them ranks. The following list shows the tensions ordered from the highest to the lowest rank: Health at Risk, Life at Risk, Actor Dead, Prisoner, Potential Danger, Clashing Emotions, Love Competition, Prisoner Free. In this way, the system includes in the scene the tension with the highest rank.

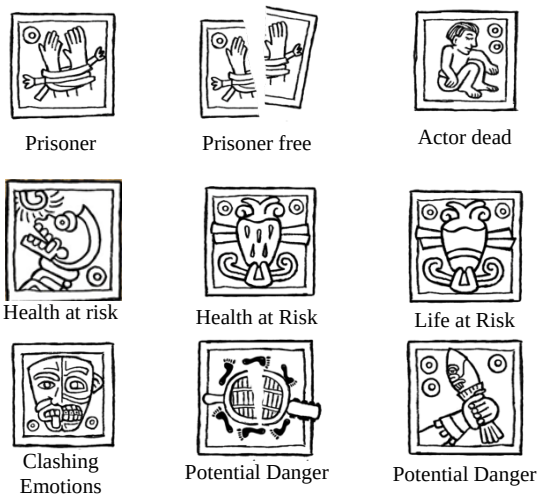


Figure 2. Glyphs representing tensions

Thus, a scene is comprised by a location on the back, two characters facing each other, and in the middle of them a glyph representing the core tension of the scene (see figure 5). The Visual Narrator employs the text to define which characters are participating in the scene as well as its location; it also employs internal representations to establish which tensions should be used.

Evaluation

We were interested in evaluating if the code produced by our system satisfied (at least partially) the requirement that both, author and reader, shared the same code. Thus, the goals of the survey were: a) To evaluate the degree of proper understanding of characters and scenes; b) To establish if the sequence of scenes communicate a clear and congruent narrative.

To perform the test the Visual Narrator illustrated a brief narrative generated by our plot generator; then, we asked a group of people to evaluate it. We developed a questionnaire that was answered by 44 persons: 91% Mexicans, 7% Spanish and 2% Guatemalans. 66% were females and 34% males. 5% had a PhD degree; 25% had a master degree; 52% had a bachelor degree; 18% had other types of degree. The questionnaire was elaborated and answered in Spanish. The questionnaire was divided in three sections. The first section showed three images of different characters developed by the Visual Narrator (see figurer 3). For each picture subjects were requested to perform the following tasks: 1) to answer if they recognized the character portrayed as pre-Hispanic ; 2) if they did, to select which of the following options described the best such a character: Tlatoani, Enemy, Jaguar Knight, Princess, Female Peasant; 3) if they did not recognize the character as pre-Hispanic, to briefly explain why.

a)  b)  c) 

Figure 3. Portraits of a) a Jaguar Knight, b) a Princess and c) an Enemy.

The second section showed two glyphs (see figure 4). Subjects were instructed that pre-Hispanics employed such glyphs to represent concepts. Then, they were asked to describe what they thought each glyph symbolized.

a)  b) 

Figure 4. Two glyphs representing a) Actor dead and b) Life at Risk

In the third section subjects were presented with an individual scene (see figure 5a) and then with the whole sequence of three scenes (see figure 5), all they developed by the Visual Narrator. Subjects were requested to describe what they thought the individual scene denoted; after that, they were asked to describe what the sequence of scenes denoted.

Evaluation of Characters' Representation

Figure 3a characterises a Jaguar Knight. To the question if they recognized the character portrayed as pre-Hispanic 95% of the subjects answered yes and 5% answered no. To the request of choosing the best description of the character depicted in the figure 3a, 93% selected the option Jaguar Knight 2% selected the option of *Tlatoani* and 5% did not answer the question.

In figure 3b the Visual Narrator characterised a Princess. To the question if they recognized the character portrayed as pre-Hispanic 89% of the subjects answered yes, 9% answered no, and 2% did not answer the question. To the request of choosing the best description of the character depicted in the figure 3b 63% selected the option Princess, 25% selected the option peasent, 2% selected Tlatoani and 10% did not answer.

In figure 3c the Visual Narrator attempted to characterise an Enemy. To the question if they recognized the character portrayed as pre-Hispanic75% of the subjects answered yes and 22% answered no. They did not identify the character as pre-Columbian due to the type of ornaments that the character was wearing and some of his characteristics like his baldness. In fact, some subjects identified this character as Egyptian. 2% did not answer the question. To the request of choosing the best description of the character depicted in the figure 3c 41% selected the option Tlatoani, 36% selected the option Enemy, and 23% did not answer.

a)

b)

c)

Figure 5. Three scenes generated by the Visual Narrator

Evaluation of Glyphs

In this section two glyphs were presented to the subjects and they were asked to describe what they thought each image symbolised.

Following the pre-Hispanic tradition, the Visual Narrator employed figure 4a to represent death. None of the subjects associated the picture with passing away. Partakers associated the following meanings to the glyph: 36% of the participants described a quiet man (meditating, praying or dreaming); 18% described a man that is thinking, which is related to the former description; 18% related this glyph to the idea of life (birth, harvest, vintage or a foetus in the

womb); 13% detected the presence of graphic elements associated to the numeric system of the pre-Hispanic civilization and therefore linked this glyph to a date; 15% gave different interpretations, such as someone seating, learning, a young person.

The Visual Narrator employed figure 4b to represent that the life of a character was at risk. As in the previous case, subjects associated different meanings to this graphic element. 48% related this glyph to the notion of death; interesting enough, 30% of this group elaborated their descriptions of death with notions like worship and rituals, the sun (which was a god between the pre-Columbian civilizations) and fights between life and death. 43% of the 44 participants associated the glyph with fear to the divinity, danger, fight between good and evil. These descriptions are closer to what the glyph intended to represent. 9% did not answer.

Evaluation of Scenes

In the last section subjects were presented with a complete scene (see figure 5a). The Visual Narrator built it from the action Enemy kidnapped the Princess, which triggers the tension Princess prisoner. 70% of the participants described the scene as representing lack of freedom, submission, slavery, capture, kidnapping and conquest; 13% related it to confrontation between two groups or tribes, and the conquest of territories; 13% linked it to concepts such as anger, macho, bulling; 4% related it to other themes, e.g. conversation between characters, or did not answer. In most cases the male (referred to mainly as the Enemy or the Tlatoani) was identified as the antagonist and the female (referred to as the Princess and sometimes as the Peasant) as the Victim. 36% of all descriptions involved an explicit interaction, mainly as a dialog, between the male and the female characters.

Finally, subjects were presented with a sequence of three scenes (see figure 5). The Visual Narrator developed this sequence from the following actions: The Enemy kidnapped the Princes; Jaguar Knight fought against the Enemy; Jaguar Knight rescued the Princess. 75% of the participants described the sequence with the same o similar group of events; 25% described the sequence with different accounts. 80% of the subjects made in their report an explicit reference to the main roles of the characters: the Enemy was the antagonist, the knight was the hero and the princess was the victim (9% made an implicit reference to the Enemy but an explicit reference to the other two personae). 20% did not make references to their roles.

Discussion

The first section of the questionnaire provides a feedback regarding the automatic construction of characters. Most of them are clearly identified as pre-Hispanics although a few characters' features seem to produce a degree of confusion. For example, a number of subjects commented that characters are portrayed with occidental facial features. In the case of the enemy, his baldness seems to puzzle some people.

Figure 3a is easily identified as a Jaguar Knight. In the case of figure 3b, most subjects identify the image with a princess, although one fourth of the participants identified it with a peasant. This situation suggests that participants are not aware of the meaning of some pre-Hispanic symbols, e.g. the use of two hands holding a stick to represent social status. Figure 3c was the most confusing. Most people identify it as a Tlatoani and a slightly minor amount of people identify it as an Enemy. A possible cause of this situation is hinted by one of the participants. This person reports having difficulties in identifying the character in figure 3c. He explains that the image lacks majesty to be considered a Tlatoani and, at the same time, lacks aggressiveness to be considered an Enemy. Therefore, this person concludes that the image is closer to represent a priest.

Thus, these results suggest that the process employed by the Visual Narrator for building characters works adequately. The grammar allows constructing characters that clearly can be differentiated by people and which, in general, are associated with what they attempt to represent. That is, the system is capable of producing original images that satisfy the requirements associated to particular characters. Nevertheless, it is necessary to improve the graphic elements to solve problems like the ones just described. An important issue that arises from this analysis, and which will be repeated in the following lines, is the fact that people are not familiar with some important pre-Hispanic symbols. That is the case of the image of the princess and the stick held by her hands. This point will be discussed later.

The second section of the questionnaire provides a feedback regarding the interpretation of the glyphs. From the beginning, glyphs were designed to be part of a scene. However, we are interested in knowing the type of concepts that they evoke when they are not presented within a visual context. None of the subjects associate figure 4a with death. This situation is understandable because the use of the eyes closed as a symbol of death is particular of pre-Hispanic civilizations and a not well known fact between the population that answered the questionnaire. However, the glyph clearly triggers lost of associations that make sense to us.

On the other hand, figure 4b represents that the health of a character is at risk (i.e. the character is wounded or ill). Although an important number of subjects associated it with death, a similar number associated it with concepts close to health at risk. In fact, we can think that death is also a concept close to the intended meaning. The reason why the second glyph was better interpreted than the first one probably has to do with the fact that this second glyph was designed by the authors of this work. That is, it does not belong to the pre-Columbian tradition. Thus, glyphs seem to fulfil its function of evoking concepts and ideas

although, again, we have the problem of a lack of knowledge about the significance of pre-Columbian symbols.

The third section of the questionnaire provides a feedback regarding scenes. In the first case, the majority of subjects interpreted figure 5a as expected. Thus, the composition, i.e. the interaction between the characters, the glyph and the location, seems to be working appropriately.

The majority of the descriptions of the sequence of the three scenes (see figure 5) provided by the subjects are alike the text generated by our storyteller. That is, participants interpreted the sequence as expected. In the same way, the role of the three characters in most of the participants' narrations equals the intended role.

Three glyphs are employed in such a sequence: the first represents that a character is a prisoner (a tied pair of hands); the second represents a Potential Danger, i.e. one characters hates other (a hand holding a flint knife); and the last represents that a character that was a prisoner has been released (a broken prisoner's glyph). The three glyphs were designed by the authors of this work. From our point of view, the first and third glyphs can easily be interpreted. However, the intended meaning of the second one is at least not obvious if not obscure. Nevertheless, the context provided by the three scenes seems to give to the user enough information to correctly interpret it. This is an interesting result that illustrates the importance of the context.

This brings us back to issue about the lack of knowledge of pre-Hispanic symbolisms. We are interested in achieving a good communication with those interested in the Visual Narrator. At the same, we would like to be as faithful as possible to the pre-Hispanic traditions, which were the inspiration of this work. So, we need to find an adequate balance. The first step is to keep on researching about the role of the context in the interpretation of visual narratives. The results in this work seem to suggest that we might be able to employ unknown symbols that, with the help of an adequate context, can be interpreted as intended by people.

One of the most interesting characteristic of the whole project is the use of emotional links and tensions between characters. The Visual Narrator employs this information to depict emotions in its characters. It is interesting to notice that 23% of the subjects make comments about the emotional states of characters with descriptions that include words like anger, surprise and sadness (or crying). This result seems to suggest that characters' portraits express emotional states. However, we need to perform a deeper evaluation of this aspect.

In this way, the Visual Narrator is capable of constructing a short visual narrative (three scenes) that, in general terms, is understood by a group of human evaluators. The primitives employed to build characters' portraits, and the composition process seems to be satisfactory. This result suggests that we are walking in the right direction. There are several challenges in front of us. The most important is to provide the Visual Narrator with mechanisms that allow more freedom during the composition process.

Most people that answered the questionnaire are unfamiliar with the meaning of pre-Hispanic iconography. Therefore, for them it might be difficult to comprehend this kind of visual narratives. However, it is this lack of prior experience that provides a great opportunity to better understand the mechanisms required to generate satisfactory visual narratives.

In summary, we have a plot generator system called MEXICA which is based on the E-R Model of creative writing. MEXICA is capable of illustrating its own narratives. The illustration process consists in analysing the dramatic tensions of the narrative and, employing a grammar, composing a sequence of images that represents such a narrative. This paper reports on how the grammar is employed to create the images. It is worth noticing that any system based on the E-R Model can employ the Visual Narrator. It might be necessary to modify the image-base; but the grammar and the process for analysing active tensions can be used.

We expect that this work will contribute to a better comprehension of this fascinating area.

References

Acaso, M. 2009. *El lenguaje visual.* Editorial Paidós. México.

Álvarez, M.; Pérez y Pérez, R.; Aliseda, A. 2007. A Generative Grammar for Pre-Hispanic Production: The Case of El Tajín Style. In Proceedings of the 4th International Joint Workshop in Computational Creativity, Goldsmiths, University of London, pp. 39-46.

Arnheim, R. 1969. *Visual Thinking*. University of California Press. Berkeley, California.

Atkin, A. 2008. Peirce's final account of signs and the Philosophy of Language, in *Transactions of the Charles S. Peirce Society*, Vol. 44, No. 1 Winter. pp. 63-85. Indiana University Press.

Blair, D.; Meyer T. 1997 Tools for an Interactive Virtual Cinema. In *Creating Personalities for Synthetic Actors: Towards Autonomous Personality Agents*. Ed. Robert Trappl and Paolo Petta. Berlin: Springer Verlag.

Cassell, J. 2001. Embodied Conversational Agents: Representation and Intelligence in User Interface. AI Magazine, Vol. 22, No. 3.

Colton, S. 2011. The Painting Fool: Stories from Building an Automated Painter. In J. McCormack and M. d'Inverno, eds., *Computers and Creativity*. Springer-Verlag, forthcoming.

Dautenhahn, K. 2001. The Narrative Intelligence Hypothesis: In Search of the Transactional Format of Narratives in Humans and Other Social Animals. In M. Beynon, C.L. Nehaniv, and K. Dautenhahn (Eds.): In *Cognitive Tech-*

nology: Instruments of Mind, 4th International Conference, CT 2001, Warwick, UK, pp. 248-266.
Eco, U. 2005. *Tratado de semiótica general*. Editorial Lumen. Barcelona.
Eco, U. 1991. *The Limits of Interpretation (Advances in Semiotics)*. Indiana University Press.
Eubanks, P. 2004. Poetics and Narrativity: How texts tell stories. In C. Bazerman & P Prior (eds.) What writing Does and how it does it. Mahwah, New Jersey: LEA.
Galarza, J. 1997. Los codices mexicanos. Arqueología Mexicana. 4:6-24.
Galarza, J.; Libura M. K. 2004. Para Leer La Tira de la Peregrinación. Ediciones Tecolote. México.
Lopez, A. 1999. Misterios de la vida y la muerte Arqueología Mexicana. 7:4-12.
López, A. 2004. La composición de la Persona tradicional Mesoamericana. Arqueología Mexicana. 11:30-41.
Mateas, M.; Sengers, P, 1999. Narrative Intelligence. *AAAI Fall Symposium on Narrative Intelligence*, pp. 1-10. North Falmouth, MA, USA.
McCloud, S. 1993. Understanding Comics The invisible Art. N.Y.: HarperCollins.
Mohar, L. 1997. Tres Códices Nahuas de México Antiguo. Arqueología Mexicana. 4:56-63.
Norton, D.; Heath, D.; Ventura, D. 2011. An Artistic Dialogue with the Artificial. In *Proceedings of the Eighth ACM Conference on Creativity and Cognition*, 31-40.
Pérez y Pérez, R. 2007. Employing Emotions to Drive Plot Generation in a Computer-Based Storyteller. *Cognitive Systems Research*. Vol. 8, number 2, pp. 89-109.
Pérez y Pérez, R.; Sosa, R.; Lemaitre, C. 2007. A computer Model of Visual Daydreaming. In *Proceedings of the AAAI 2007 Fall Symposia in Intelligent Narrative Technologies*, Arlington, Virginia, pp. 102-109.
Pérez y Pérez, R.; Sharples, M. 2001. MEXICA: a computer model of a cognitive account of creative writing. *Journal of Experimental and Theoretical Artificial Intelligence*. Volume 13, number 2, pp. 119-139.
Riedl, M.; Stern, A.; Dini, D. M.; and Alderman, J. M. 2008. Dynamic experience management in virtual worlds for entertainment, education, and training. *International Transactions on System Science and Applications*, 3:23–42.
Rickel, J.; Johnson, W L. 1999. Animated Agents for Procedural Training in Virtual Reality: Perception, Cognition, and Motor Control. Applied Artificial Intelligence, Vol. 13.
Rieff, P. 1996. Atuendos del México Antiguo. Arqueología Mexicana. 3:6-16.
Rose, G. 2001. *Visual Methodologies*. SAGE Publications Ltd. London.
Seymour, C. 1997. *Story and Discourse: Narrative Structure in Fiction and Film*. Ithaca, New York.

Generating a Complete Multipart Musical Composition from a Single Monophonic Melody with Functional Scaffolding

Amy K. Hoover, Paul A. Szerlip, Marie E. Norton, Trevor A. Brindle, Zachary Merritt, and Kenneth O. Stanley
Department of Electrical Engineering and Computer Science
University of Central Florida
Orlando, FL 32816-2362 USA
{ahoover@eecs.ucf.edu, paul.szerlip@gmail.com, marie.norton@knights.ucf.edu, tabrindle@gmail.com, zbmerritt@gmail.com, kstanley@eecs.ucf.edu }

Abstract

This paper advances the state of the art for a computer-assisted approach to music generation called *functional scaffolding for musical composition* (FSMC), whose representation facilitates creative combination, exploration, and transformation of musical concepts. Music in FSMC is represented as a functional relationship between an existing human composition, or *scaffold*, and a generated accompaniment. This relationship is encoded by a type of artificial neural network called a compositional pattern producing network (CPPN). A human user without any musical expertise can then explore how accompaniment should relate to the scaffold through an interactive evolutionary process akin to animal breeding. While the power of such a functional representation has previously been shown to constrain the search to plausible accompaniments, this study goes further by showing that the user can tailor complete multipart arrangements from only a single original monophonic track provided by the user, thus enabling creativity without the need for musical expertise.

Introduction

Among the most important functions of any approach to enhancing human creativity is what Boden (2004) terms *transformational creativity*. That is, key creative obstacles faced by human artists and musicians are the implicit constraints acquired over a lifetime that shape search space structure. By offering an instance of the search space (e.g. of musical accompaniments) with a radically different structure, a creativity-enhancing program can potentially liberate the human to discover unrealized possibilities. In effect, the familiar space of the human artist is *transformed* into a new structure intrinsic to the program. Once the user is exposed to this new world, as a practical matter the program must provide to the user the ability to *explore* and *combine* concepts within the newly-conceived search space, which corresponds to Boden's *combinatorial* and *exploratory* classes of creativity (Boden, 2004). That way, the user experiences a rich and complete creative process within a space that was heretofore inconceivable.

The danger with transformational creativity in computational settings is that breaking hard-learned rules may feel unnatural and thereby unsatisfying (Boden, 2007). Any attempt to facilitate transformational creativity should respect the relationships between key artistic elements even as they are presented in a new light. Thus for a given domain, such as musical accompaniment, a delicate balance must be struck between unfettered novelty and respect for essential structure.

Many approaches to generating music focus on producing a natural sound at the cost of restricting creative exploration. Because structure is emphasized, the musical space is defined by rules that constrain the results to different styles and genres (Todd and Werner, 1999; Chuan, 2009; Cope, 1987). The necessity for a priori rules potentially facilitates the combination of musical structures or exploration of the defined space, but precludes transformational outcomes.

In contrast, musical structures in the approach examined in this paper, *functional scaffolding for musical composition* (FSMC), are defined as the very functions that relate one part of a piece to another, thereby enabling satisfying transformational creativity (Hoover, Szerlip, and Stanley, 2011a,b). Based on the idea that music can be represented as a function of time, FSMC inputs a simple, isolated musical idea into a function that outputs accompaniment that respects the structure of the original piece. The function is represented as a special type of artificial neural network called a compositional pattern producing network (CPPN). In practice, the CPPN inputs existing music and outputs accompaniment. The user-guided creative exploration itself is facilitated by an interactive evolutionary technique that in effect allows the user to *breed* the key functional relationships that yield accompaniment, which supports both combinatorial and exploratory creativity (Boden, 2004) through the crossover and mutation operators present in evolutionary algorithms. By representing music as *relationships* between parts of a multipart composition, FSMC creates a new formalism for a musical space that transforms its structure for the user while still respecting its fundamental constraints.

Hoover, Szerlip, and Stanley (2011a,b) showed that FSMC can produce accompaniments that are indis-

tinguishable by listeners from fully human-composed pieces. However, the accompaniment in these studies was only a single monophonic instrument, leaving open the key question of whether a user with little or no musical expertise can perhaps generate an entire multipart arrangement with this technology from just a single-instrument monophonic starting melody. If that were possible, then anyone with only the ability to conceive a single, monophonic melody could in principle expand it into a complete multilayered musical product, thereby enhancing the creative potential of millions of amateur musicians who possess inspiration but not the expertise to realize it. This paper demonstrates that FSMC indeed makes such achievement possible.

Background

This section relates FSMC to traditional approaches to automated composition and previous creativity-enhancing techniques.

Automatic Music Generation

Many musical representations have been proposed before FSMC, although their focus is not necessarily on representing the functional relationship between parts. For example, from long before FSMC, Holtzman (1980) creates a musical grammar that generates harp solos based on the physical limitations imposed on harp performers. Similarly, Cope (1987) derives grammars from the linguistic principles of haiku to generate music in a particular style. These examples and other grammar-based systems are predicated on the idea that music follows grammatical rules and thus by modeling musical relationships as grammars, they are representing the important structures of music (Roads, 1979; McCormack, 1996). While grammars can produce a natural sound, deciding which aspects of musical structure should be represented by them is often difficult and ad hoc (Kippen and Bel, 1992; Marsden, 2000).

Impro-Visor helps users create monophonic jazz solos by automatically composing any number of measures in the style of famous jazz artists (Keller et al., 2006). Styles are represented as grammars that the user can invoke to complete compositions. Creativity-enhancement in Impro-Visor occurs through the interaction of the user's own writing and the program's suggestions. When users have difficulty elaborating musical concepts, they can access predictions of how famous musicians would approach the problem within the context of the current composition. By first learning different professional compositional techniques, students can then begin developing their own personal styles. While Impro-Visor is an innovative tool for teaching jazz styles to experienced musicians, it focuses on emulating prior musicians over exploration.

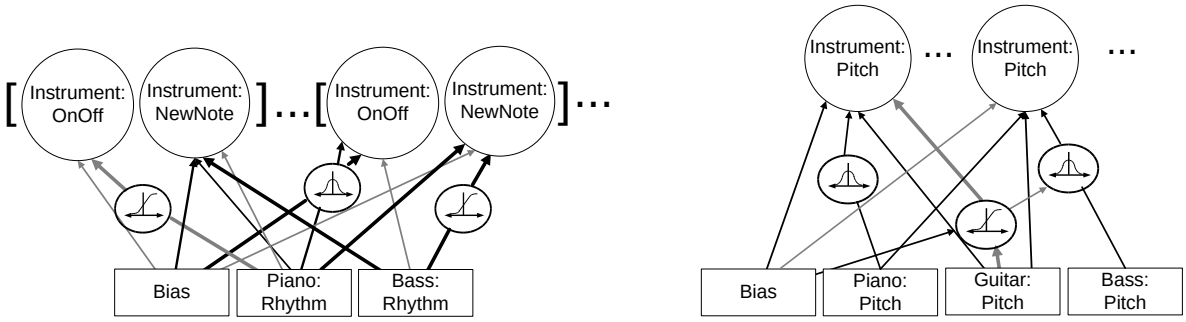
Enhancing Creativity in Music Composition

A problem with traditional approaches to music composition is that standard representations can potentially limit creative exploration. For instance, MySong generates chord-based accompaniment for a vocal piece from hidden Markov models (Simon, Morris, and Basu, 2008). Users select any vocal piece and MySong outputs accompaniment based on a transition table, a weighting factor that permits greater deviation from the table, and musical style (e.g. rock, big band). MySong thus allows users to create accompaniment for their own melodies in a variety of different predefined styles from which users cannot deviate. Zicarelli (1987) describes an early interactive composition program, Jam Factory, that improvises on human-provided MIDI inputs from rules represented in transition tables. Users manipulate the output in several ways including the probability distributions of eight different transition tables; there are four each for both rhythm and pitch. Users are provided more creative control in designing and consulting the transition tables, but the increased flexibility results in unnatural outputs that thereby limit the utility of the main algorithms (Zicarelli, 2002). The approach described by Chuan (2009) balances user control by training transition tables based on only a few user-provided examples. The tables then reflect the "style" inherent in the examples and can generate chord-based accompaniment for a user's own piece. While each of these systems offers users varied levels of control, rule manipulation alone may not be sufficient to access all three forms of creativity described by Boden (2004). For example, the representations cannot easily combine musical ideas or transform the musical space (due to inherent rule restrictions).

Alternatively, most interactive evolutionary computation (IEC) (Takagi, 2001) approaches facilitate creativity through the evolutionary operators of crossover and mutation, and require human involvement in the creative process. In GenJam a human player and computer "trade fours," a process whereby the human plays four measures and the computer "answers" them with four measures of its own (Biles, 1998). Musical propositions are mutated and combined into candidates that the user rates as good or bad. Similarly, Jacob (1995) introduces a system in which human users rate, combine, and explore musical candidates at three different levels of the composition process and Ralley (1995) generates melodies by creating a population from mutations of a provided user input. Finally, CACIE creates atonal pieces by concatenating musical phrases as they are generated over time (Ando and Iba, 2007). Each phrase is represented as a tree structure that users can interactively evolve or directly manipulate. However, most such systems impose explicit musical rules conceived by the developer to constrain the search spaces of possible accompaniment, thus narrowing the potential for discovery.

Previous Work in FSMC

The FSMC approach in this paper is based on previous work by Hoover, Szerlip, and Stanley (2011a,b), who focused on evolving a single monophonic accompa-



(a) Rhythm (b) Pitch

Figure 1: **How CPPNs Compute a Function of the Input Scaffold.** The rhythm CPPN in (a) and pitch CPPN in (b) together form the accompaniments of FSMC. The inputs to the CPPNs are the scaffold rhythms and pitches for the respective networks and the outputs indicate the accompaniment rhythms and pitches. Each rhythm network has two outputs: OnOff and NewNote. The OnOff node controls volume and whether or not a note is played. The NewNote node indicates whether a note is re-voiced or sustained at the current tick. If OnOff indicates a rest, the NewNote node is ignored. The pitch CPPN output decides what pitch the accompaniment should play at that particular tick. The internal topologies of these networks, which encode the functions they perform, change over evolution. The functions within each node depict that a CPPN can include more than one activation function, such as Gaussian and sigmoid functions. Two monophonic accompaniment outputs are depicted, but the number of instruments a CPPN can output is unlimited. The number of input instruments also can vary.

niment for a multipart MIDI. These accompaniments are generated through two functions, one each for pitch and rhythm, that are represented as compositional pattern producing network (CPPNs), a special type of artificial neural network (ANN). CPPNs can evolve to assume an arbitrary topology wherein each neuron is assigned one of several activation functions. Through IEC, users explore the range of accompaniments with NeuroEvolution of Augmenting Topologies (NEAT), a method for growing and mutating CPPNs (Stanley and Miikkulainen, 2002). Unlike traditional ANN learning, NEAT is a policy search method, i.e. it explores accompaniment possibilities rather than optimizing toward a target. While existing songs with generated accompaniments were indistinguishable in a listener study from fully-composed human pieces, the real achievement for this approach would be to help the user generate entire polyphonic and multi-instrument accompaniment from just a single voice of melody (Hoover, Szerlip, and Stanley, 2011a). This paper realizes this vision.

Approach: Extending Functional Scaffolding for Music Composition

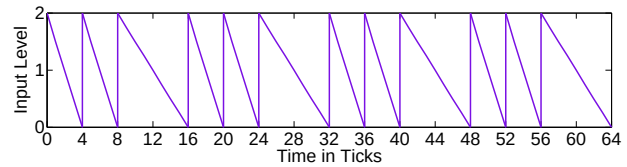
This section extends the original FSMC approach, which only evolved a single monophonic accompaniment (Hoover, Szerlip, and Stanley, 2011a,b). It explains the core principles of the approach and how they are applied to producing multipart accompaniments.

Defining the Musical Space

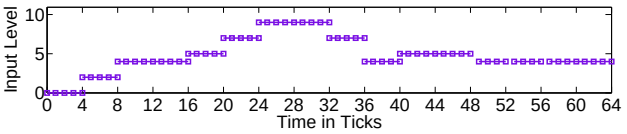
A crucial aspect of any creativity-enhancing approach for music composition is first to define the musical space. Users can help define this space in FSMC by first selecting a musical starting point, i.e. the monophonic melody or *scaffold*. Initial scaffolds can be composed in any style and if they are only single monophonic parts as in this paper, they can be composed by users within a wide range of musical skill and expertise. The main insight behind the representation in FSMC is that a robust space of accompaniments can be created with only this initial scaffold. Because of the relationship of different accompaniment parts to the scaffold and therefore to each other, the space is easily created and explored.

Each instrument part in the accompaniment is the result of two separate functions that independently relate rhythmic and pitch information in the scaffold (i.e. the inputs) to the generated accompaniment. Depicted in figure 1, these functions are represented as CPPNs, the special type of ANN described in the background (Stanley, 2007). As figure 1 shows, multiple inputs can be sent to the output and many different instruments can be represented by the same CPPN. CPPNs incrementally grow through the NEAT method, which means they can in principle evolve to represent *any* function (Stanley and Miikkulainen, 2002; Cybenko, 1989). Together, the rhythmic and pitch CPPNs that will be evolved through NEAT define the musical space that the user can manipulate. In effect, pitch information from the scaffold is fed into the pitch CPPN at the same time as rhythmic information is fed into the rhythm CPPN. Both CPPNs then output how the accompaniment should behave in response. That way, they compute a function of the scaffold.

Accompaniments are divided into a series of discrete time intervals called *ticks* that are concatenated together to form an entire piece. Each tick typically represents the length of an eighth note, but this division can



(a) Rhythm



(b) Pitch

Figure 2: **Input Representation.** The spike-decay representation for rhythmic inputs is shown in (a) and the pitch representation is in (b). Rhythm is encoded as a set of decaying spikes that convey the duration of each note. Because the CPPN sees *where* within each spike it is at any given tick, in effect it can synchronize its outputs with the timing of the input notes. Pitch on the other hand is input simply as the current note at the present tick.

be altered through an interface. Outputs are gathered from both the rhythmic and pitch CPPNs at each tick that are combined to determine the accompaniment at that tick. As shown in figure 1a, the two outputs of the rhythm network for each line of accompaniment are *OnOff*, which indicates whether a note or rest is played and its volume, and *NewNote*, which indicates whether or not to sustain the previous note. The single *pitch* output for each line of accompaniment in figure 1b determines instrument pitch at the current tick relative to a user-specified key.

To produce the outputs, rhythmic and pitch information from the scaffold is sent to the CPPN at each tick. The continuous-time graph in figure 2a illustrates how rhythmic information in the scaffold is sent to the CPPN. When a note strikes, it is represented as a maximum input level that decays linearly over time (i.e. over ticks) until the note ends. At the same tick, pitch information on the current note is input as a pitch class into the pitch CPPN (figure 2b). That is, two C notes in different octaves (e.g. C4 and C5) are not distinguished.

The sound of instruments in FSMC can be altered through instrument choice or key. A user can pick any of 128 pitched MIDI instruments and can request any key. Once a user decides from what preexisting piece the scaffold is provided and the output instruments most appropriate for the piece, candidate CPPNs can be generated, thus establishing the musical space of accompaniments. The theory behind this approach is that by exploring the potential *relationships* between scaffolds and their accompaniments (as opposed to exploring direct representations of the accompaniment itself), the user is constrained to a space in which candidate accompaniments are almost all likely *coherent* with respect to the scaffold. The next section describes how users can combine, explore, and transform this space to harness their own musical creativity.

Navigating the Musical Space

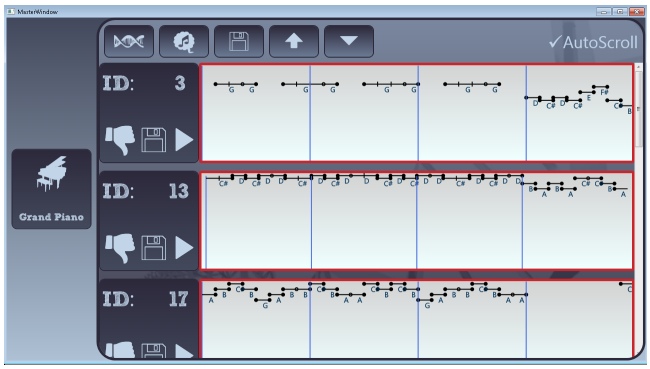


Figure 3: **Program Interface.** This screenshot of the program (called MaestroGenesis) that implements FSMC shows accompaniments for a melody input by the user. The instrument output is currently set to Grand Piano on the left-hand side, but can be changed through a menu. Accompaniments are represented as horizontal bars and are named by their ID. The user selects his or her favorite and then requests a new generation of candidates

Exploration of musical space in FSMC begins with the presentation to the user of the output of ten randomly-generated CPPN pairs, each defining the key musical relationships between the scaffold and output accompaniment. These accompaniments can be viewed in a graphical depiction (as shown in the screenshot in figure 3) or in standard musical notation. They can be played and heard through either MIDI or MP3 formats. The user-guided process of exploration that combines and mutates these candidates is called *interactive evolutionary computation* (IEC) (Takagi, 2001). Because each accompaniment is encoded by two CPPNs, evolution can alter both the pitch and rhythmic CPPNs or adjust them individually.

The user combines and explores accompaniments in this space by selecting and rating one or more accompaniments from one generation to parent the individuals of the next generation. The idea is that the good musical ideas from both the rhythmic and pitch functions are preserved with slight alterations or combined to create a variety of new but related functions, some of which may be more appealing than their parents. The space can also be explored without combination by selecting only a single accompaniment. The next generation then contains slight mutations of the original functions.

While IEC inherently facilitates these types of creativity, the approach in this paper extends the reach of transformational creativity offered by FSMC. Previously, FSMC generated single-voice accompaniments to be played with a fully-composed, preexisting human

piece (Hoover, Szerlip, and Stanley, 2011a,b). This paper introduces a new layering technique whereby generated accompaniment from previous generations can serve as inputs to new CPPNs that then generate more layers of harmony. The result is the ability to spawn an entire multi-layered piece from a single monophonic starting melody.

One such layering approach is performed by generating one new monophonic accompaniment at a time. The first layer is the monophonic melody composed by the human user. The second layer is generated through FSMC from the first. The third layer is then generated through FSMC by now inputting into the CPPNs the first *and* second layers, and so on. All of the layers are finally combined to create an entire accompaniment, resulting in accompaniments that are functionally related to both the initial melody and previous accompaniment lines. In this way, each accompaniment line is slightly more removed from the original melody and subsequent accompaniment lines are based functionally on both the scaffold and previously-generated lines.

To create accompaniments more closely related to the original melody, another layering technique is for users to generate *all* accompaniment layers from only the single monophonic starting point. For this purpose, the CPPNs are given enough outputs to represent all the instruments in the accompaniment at the same time. Because the melody and the accompaniments are functionally related, any accompaniment will follow the contours of the melodic starting point. However, in this case, the only influence on each accompaniment is this starting point itself, yielding a subtly different feel.

With either of these approaches or a combination of them users can further influence their accompaniments by holding constant the rhythm CPPN or pitch CPPN while letting the other evolve. Interestingly, when two accompaniments share the same rhythm network but differ in the pitch network slightly, the two monophonic instruments effectively combine to create the sound of a polyphonic instrument. Similarly, the pitch networks can be shared while the rhythm networks are evolved separately, creating a different sound. Notice that this approach requires no musical expertise to generate multiple lines of accompaniment.

Experiments

The experiments in this paper are designed to show how users can generate multipart pieces from a single monophonic melody with FSMC. They are divided into accompaniment generation and a listener study that establishes the quality of the compositions.

Accompaniment Generation

For this experiment, three members of our team composed in total three monophonic melodies. From each of these user-composed melodies, a multipart accompaniment was generated through FSMC by the author of the originating melody. Two other multipart accompaniments were generated for the folk song, *Early One Morning*. We chose to include each of these FSMC composers, who were undergraduate independent study students at the University of Central Florida, as authors of this paper to recognize their pioneering efforts with a new medium. The most important point is that no musical expertise need be applied to the final creations beyond that necessary to compose the initial monophonic melody in MIDI format. Thus, although results may sound consciously arranged it is important to bear in mind that all the polyphony you hear is entirely the output of FSMC. The original melodies, accompaniments, and CPPNs are available at `http://eplex.cs.ucf.edu/fsmc/iccc2012`. The program, called MaestroGenesis, is available at `http://maestrogenesis.org`.

As noted in the approach, FSMC provides significant freedom to the user in how to accumulate the layers of a multipart piece. In general, the user has the ability to decide from which parts to generate other parts. For example, from the original melody, five additional parts could be generated at once. Or, instead, the user might accumulate layers incrementally, feeding each new part into a new CPPN to evolve yet another layer. Some layers might depend on one previous layer, while other might depend on multiple previous layers. In effect, such decisions shape the subtle structural relationships and hence aesthetic of the final composition. For example, evolving all of the new parts from just the melody gives the melody a commanding influence over all the accompaniment, while incrementally training each layer from the last induces a more delicate and complex set of harmonious partnerships. As the remainder of this section describes, the student composers took advantage of this latitude in a variety of ways

Early One Morning, (Song 1) versions 1 and 2 with four- and five-part accompaniments began from an initial monophonic melody transcribed from the traditional, human-composed folk song. The second layer is identical in both versions and was evolved from Early One Morning itself. The third, fourth, and fifth parts of version 1 were all evolved from the second layer. The third, fourth, fifth, and sixth parts of version 2 were evolved from the pitch network of the second layer of version 1, and the rhythm network from the original Early One Morning monophonic melody. This experiment illustrates that the results with FSMC given the same starting melody are not deterministic and in fact do provide creative latitude to the user even without the need for traditional composition techniques.

Song 2 started from an original monophonic melody composed by undergraduate Marie E. Norton. The second layer was added by inputting this melody into the rhythm and pitch networks of the subsequent accompaniment populations. This second layer then served as input to the pitch and rhythmic CPPNs for layers 3 and 4. The pitch CPPN for layer 5 consisted of layer 2, but the rhythm network only had a bias input. Finally, the inputs for the pitch network for layer 6 were layers 3, 4, and 5, while the inputs to the rhythm CPPN were

layer 4 and a measure timing signal first introduced for FSMC by Hoover and Stanley (2009) that gives the network a sense of where the song is within the measure. All of the layers finally combined to create a single, multipart piece in which each line is functionally related to the others. Each layer took as few as three to as many as five generations to evolve.

For Song 3, Zachary Merritt first created a layer that influences most of the other layers, but is not heard in the final track. The fourth layer was generated from the third, which is influenced by the monophonic melody and the unheard layer. The fifth layer was generated from the population of the fourth layer with the rhythm network held constant to create a chordal feel. The sixth layer was generated from only the initial starting melody and a special timing signal that imparts a sense of the position in the overall piece (Hoover and Stanley, 2009). Similarly, the seventh layer is generated from only the initial starting melody, but adds a separate function input, $\sin(\pi x)$, where x is the time in measure. Although there are seven layers described in this experiment, only six were selected to be heard, meaning that there is a five-part accompaniment.

Trevor A. Brindle created an initial piece and evolved all five accompaniment lines for Song 4 directly from it. Instead of inputting results from previous generations, he started new runs for each voice from the same scaffold, giving a strong influence to the melody.

Notice that the key decisions made by the users are in general from which tracks to generate more tracks. Of course the users also performed the IEC selection operations to breed each new layer. Importantly, none such decisions require musical expertise.

Listener Study

The contribution of users to the quality of the generated works and accordingly the effectiveness of the creativity enhancement is evaluated through a listener study. The study consists of five surveys, one for each generated arrangement. The surveys present two MP3s to the listener, who is asked to rate the quality of both. The first MP3, called the *collaborative accompaniment*, is an arrangement resulting from the collaboration of the author with the program (i.e. the two versions from Early One Morning or Songs 2, 3, or 4). The second, called the *FSMC-alone accompaniment*, is generated by the program *alone*. That is, a random pitch CPPN and a random rhythm CPPN are provided the same monophonic starting melody as the collaborative accompaniment and their output is taken as the FSMC-alone accompaniment. Thus the factor that is isolated is the involvement of the *human user*, who is not involved in the FSMC-alone accompaniment. However, it is important to note that the FSMC-alone accompaniments do not actually *sound* random because even if the CPPNs are generated randomly, they are still functions of the same scaffold, which tends even in the random case to yield outputs that sound at least coherent (which is the motivation for FSMC in the first place). Thus this study investigates whether the human user is really able to make a creative contribution by leveraging FSMC.

A total of 129 students participated in the study. The full survey is available at `http://eplex.cs.ucf.edu/fsmc/iccc2012/survey`, but note that in the administered surveys, the order of the MP3s was random to avoid any bias. The users were asked to rate each piece with the following question:

> Rate MIDI i on a scale of one to ten. (1 is the worst and 10 is the best),

where i refers to one of the ten generated works. The idea is that if the user-created arrangements are rated higher than those generated by FSMC-alone, the user's own input likely positively influenced the outcome. While this study focuses on the quality of output, the degree to which FSMC *enhances* creativity will be addressed in future work.

Results

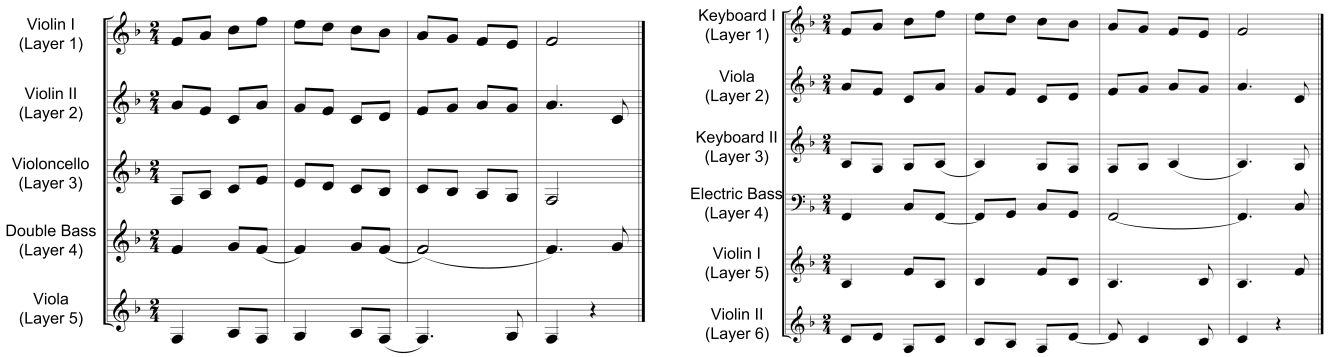
The generated accompaniments and original scaffold discussed in this section can be heard at `http://eplex.cs.ucf.edu/fsmc/iccc2012`.

Accompaniments

Samples of the scores for the two arrangements created to accompany Early One Morning are shown in figure 4. The layers are shown in order from top to bottom in both versions (layer 1 is the original melody). Layer 2, which is the same in both versions, is heard as violin II in version 1 and viola in version 2.

An important observation is that the violoncello part in version 1 follows the rhythm of the initial starting melody very closely while the pitch contour differs only slightly. While the viola and double-bass parts differ in both pitch and rhythm over the course of the song, both end phrases and subphrases on the tonic note, F, in many places over the course of the piece, including measure 4 in figure 4a. Version 2, on the other hand, contains many rhythmic similarities (i.e. the eighth note patterns contained in the keyboard I, viola, keyboard II, and the violin II parts), but illustrates distinct pitch contours. Together, the two versions illustrate how a single user can generate different accompaniment from the same initial monophonic starting melody and how the initial melody exerts its influence both rhythmically and harmonically.

Songs 2, 3, and 4 exhibit a similar effect: rhythmic and harmonic influence from the original melody, yet distinctive and original accompaniment nevertheless. The result is that the overall arrangements sound *composed* even though they are evolved through a breeding process. The next section provides evidence that impartial listeners also appreciate the contribution of the human user.



(a) Early One Morning Version 1

(b) Early One Morning Version 2

Figure 4: **Early One Morning.** The first four measures of versions 1 and 2 of Early One Morning illustrate how a single user with the same monophonic starting melody can direct the accompaniment in two different ways that nevertheless both relate to the initial melody. Because the accompaniments share two of their layers, they sound related. However, through timbre selection and the evolution of two and three distinct layers in versions 1 and 2 respectively, the user imparts a different feel.

Listener Study Results

The results of the listener study in figure 5 indicate that all of the collaborative accompaniments are rated higher than those generated with FSMC alone, with three out of five (Song 1 version 2, Song 4, and Song 5) displaying significant difference ($p < 0.05$; Student's paired t-test). Taken all together, the collaborative accompaniments sound highly significantly more appealing than those generated with FSMC alone ($p < 0.001$; Student's paired t-test). These results indicate that not only does FSMC provide a structurally plausible search space, but that it is possible to explore such a space without applying musical expertise. That is, the results suggest that the user input significantly improves the perceived quality of the generated compositions.

Discussion

A key feature of figure 4 is that the collaborative accompaniments generated by users with the assistance of FSMC follow the melodic and rhythmic contours of the original scaffold. Furthermore, the listener study suggests that FSMC helps the user establish and explore musical search spaces that may otherwise have been inaccessible.

While the users search this space through IEC, which facilitates the combination of musical ideas and the exploration of the space itself, an interesting property of this search space is its robustness; even FSMC-alone accompaniments, which are created without the benefit of human, subjective evaluation, can sound plausible. However, when coupled with the human user, this approach in effect *transforms* the user's own internal search space of possible accompaniments to one constrained by functional scaffolding.

While the quantitative data suggests the merit of collaborative accompaniments, music is inherently subjective. Therefore, it is important for the reader to judge the results for his or herself at `http://eplex.cs.ucf.edu/fsmc/iccc2012` to fully appreciate the potential of the FSMC method.

One interesting direction for future work is to explore new interpretations for the output of the pitch functions. Currently, accompaniment pitches are interpreted as discrete note values, a process that limits the instrument to playing the same note each time a given combination of notes occurs in the scaffold. However, by interpreting the output as a change in pitch (i.e. horizontal interval) rather than an absolute pitch, instruments can select any note to correspond to a particular combination depending on where in the piece it is occurring. In this way, an even larger space of musical possibilities could be created.

Perhaps most importantly, with only a single, monophonic melody, users could compose entire multipart pieces without the need for musical expertise. Even if not at the master level, such a capability opens to the novice an entirely new realm of exploration.

Conclusion

This paper presented an extension to functional scaffolding for musical composition (FSMC) that facilitates a human user's creativity by generating polyphonic compositions from a single, human-composed monophonic starting track. The technique enables creative exploration by helping the user construct and then navigate a search space of candidate accompaniments through a breeding process akin to animal breeding called interactive evolutionary computation (IEC). These collaborative accompaniments bred by users were judged by listeners against those composed only through FSMC. Overall, listeners liked collaborative accompaniments more than the FSMC-alone accompaniments. Most importantly, a promising poten-

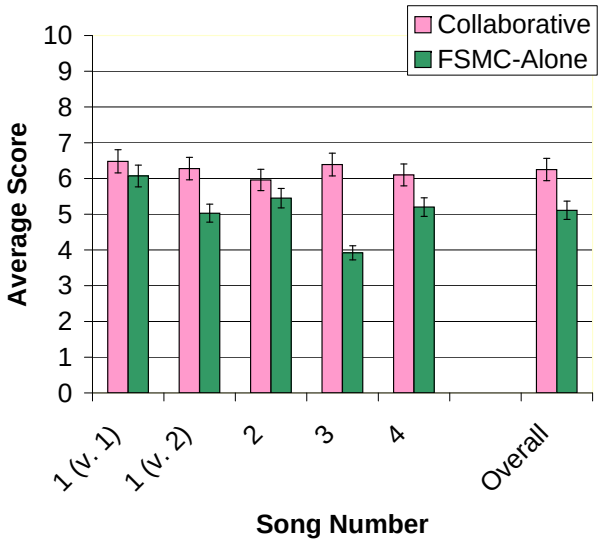


Figure 5: **Listener Study Results.** The average rating (by 129 participants) from one to ten of both the collaborative and FSMC-alone accompaniments are shown side-by-side with the lines indicating a 5% error bound. The overall results for the listener study indicate that on average the collaborative accompaniments are of significantly higher perceived quality than FSMC-alone.

tial for creativity enhancement in AI is to open up the world of the amateur to the domain once only accessible to the expert. The approach in this paper is a step in this direction.

Acknowledgements

This work was supported in part by the National Science Foundation under grant no. IIS-1002507 and also by a NSF Graduate Research Fellowship.

References

Ando, D., and Iba, H. 2007. Interactive composition aid system by means of tree representation of musical phrase. In *IEEE Congress on Evolutionary Computation (CEC)*, 4258–4265. IEEE.

Biles, J. 1998. Interactive GenJam: Integrating real-time performance with a genetic algorithm. In *Int. Computer Music Conf.(ICMC 98)*, 232–235.

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. Routledge, second edition.

Boden, M. A. 2007. *Creativity and Conceptual Art*. Oxford:Oxford University Press.

Chuan, C.-H. 2009. Supporting compositional creativity using automatic style-specific accompaniment. In *Proc. of the CHI Computational Creativity Support Workshop*.

Cope, D. 1987. An expert system for computer-assisted composition. *Computer Music Journal* 11(4):30–46.

Cybenko, G. 1989. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)* 2(4):303–314.

Holtzman, S. R. 1980. A generative grammar definition language for music. *Interface* 9(1):1–48.

Hoover, A. K., and Stanley, K. O. 2009. Exploiting functional relationships in musical composition. *Connection Science Special Issue on Music, Brain, & Cognition* 21(2):227–251.

Hoover, A. K.; Szerlip, P. A.; and Stanley, K. O. 2011a. Generating musical accompaniment through functional scaffolding. In *Proceedings of the Eighth Sound and Music Computing Conference (SMC 2011)*.

Hoover, A. K.; Szerlip, P. A.; and Stanley, K. O. 2011b. Interactively evolving harmonies through functional scaffolding. In *Proceedings of the Genectic and Evolutionary Computation Conference (GECCO-2011)*. New York, NY: The Association for Computing Machinery.

Jacob, B. L. 1995. Composing with genetic algorithms. In *Proc. of the 1995 International Computer Music Conference*, 425–455. Intl. Computer Music Association.

Keller, R. M.; Morrison, D.; Jones, S.; Thom, B.; and Wolin, A. 2006. A computational framework for enhancing jazz creativity. In *Proceedings of the Third Workshop on Computational Creativity, ECAI 2006*.

Kippen, J., and Bel, B. 1992. *Modeling Music with Grammars: Formal Language Representation in the Bol Processor*. Academic Press London. 207–238.

Marsden, A. 2000. *Readings in Music and Artificial Intelligence*. Harwood Academic Publishers. chapter Music, Intelligence, and Artificiality, 18.

McCormack, J. 1996. Grammar based music composition. *Complex Systems* 96:321–336.

Ralley, D. 1995. Genetic algorithms as a tool for melodic development. In *Proc. of the 1995 Intl. Computer Music Conf.*, 501–502. Intl. Computer Music Assoc.

Roads, C. 1979. Grammars as representations for music. *Computer Music Journal* 3(1):48–55.

Simon, I.; Morris, D.; and Basu, S. 2008. Mysong: Automatic accompaniment generation for vocal melodies. In *Proc. of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems*, 725–734. ACM.

Stanley, K. O., and Miikkulainen, R. 2002. Evolving neural networks through augmenting topologies. *Evolutionary Computation* 10:99–127.

Stanley, K. O. 2007. Compositional pattern producing networks: A novel abstraction of development. *Genetic Programming and Evolvable Machines Special Issue on Developmental Systems* 8(2):131–162.

Takagi, H. 2001. Interactive evolutionary computation: fusion of the capabilities of ec optimization and human evaluation. *Proceedings of the IEEE* 89(9):1275–1296.

Todd, P. M., and Werner, G. M. 1999. Frankensteinian methods for evolutionary music. *Musical Networks: Parallel Distributed Perception and Performace* 313–340.

Zicarelli, D. 1987. M and jam factory. *Computer Music Journal* 11(4):13–29.

Zicarelli, D. 2002. How I learned to love a program that does nothing. *Computer Music Journal* 26(4):44–51.

Soup Over Bean of Pure Joy: Culinary Ruminations of an Artificial Chef

Richard G. Morris, Scott H. Burton, Paul M. Bodily, and Dan Ventura
Computer Science Department
Brigham Young University
rmorris@axon.cs.byu.edu, sburton@byu.edu, norkish@gmail.com, ventura@cs.byu.edu

Abstract

We introduce a system for generating novel recipes and use that context to examine some current theoretical ideas for computational creativity. Specifically, we have found that the notion of a single inspiring set can be generalized into two separate sets used for generation and evaluation, respectively, with the result being greater novelty as well as system flexibility (and the potential for natural meta-level creativity), that explicitly measuring artefact typicality is not always necessary, and that explicitly defining an artefact hierarchically results in greater novelty.

1 Introduction

As a relatively new sub-field of artificial intelligence (AI), computational creativity is currently wrestling with many issues similar to those with which AI struggled several decades ago. Many questions similar to those originally asked of AI are now being asked in the context of computational creativity, including foundational questions such as "What is creativity?" Within computational creativity, there is an ongoing movement to define a theoretical foundation that can provide a level of maturity to the field. For example, Wiggins gives the following definition of computational creativity that closely mirrors definitions of intelligence accepted by many AI researchers (Wiggins 2006):

> *The study and support, through computational means and methods, of behaviour exhibited by natural and artificial systems, which would be deemed creative if exhibited by humans.*

As another example, Ritchie provides a level of formalism by supplying a framework for evaluating a creative system (Ritchie 2007). Assuming that a creative system's purpose is to produce creative artefacts, Ritchie's framework evaluates the creativity of the system in terms of the typicality and quality of generated artefacts in relation to some inspiring set of known artefacts.

Taking some of Ritchie's ideas one step further, Gervás proposes that creative systems must be able to consistently generate creative artefacts—producing artefacts that are also novel with respect to its own previous work (Gervás 2011). Gervás shows this can be accomplished by splitting the inspiring set (as discussed by Ritchie) into a reference set (used to determine the novelty of generated artefacts) and a learning set (used in the generation of artefacts). We modify this idea by splitting the inspiring set into a set used in the generation of artefacts and one for evaluating generated artefact quality. Note that this does not address the idea of a reference set at all, but it also does not preclude the use of one either (let us say the two ideas are orthogonal and likely complementary).

Evaluation of a creative system is both clearly important and inherently difficult. In a recent comprehensive survey of published creative systems, Jordanous found that only half of the papers give details on an evaluation of their system (Jordanous 2011). Despite the difficulty in measuring creativity, quality, and typicality, greater attempts must be made to evaluate them if the field is to gain maturity.

In an attempt to do so, we provide an explicit measure of quality used during the artefact generation process. We also show that an explicit measure of typicality is not necessary if it is built in to the generation process. In addition, we present an explicit measure of novelty (rare n-grams). We also show that explicitly defining a hierarchy for elements of our artefacts is beneficial to the creative system. We compare a hierarchical version of our system with one that is lacking any hierarchy and demonstrate greater novelty in the artefacts produced. The hierarchical version also gives a natural method to implicitly model typicality in the system without inhibiting novelty.

Novel perspectives on the developing theory of computational creativity are provided by concrete applications of the theory in diverse areas. Creative systems have been produced for a wide variety of artefacts, including poetry (Gervás 2000; Gervás 2001), literature (Pérez y Pérez and Sharples 2001; Pérez y Pérez 2007), music (Jordanous 2010; Lewis 2000; Monteith et al. 2011), theorem proving (Ritchie and Hanna 1984; Colton 2002), humor (Stock and Strapparava 2005; Binsted and Ritchie 1997), metaphor (Veale and Hao 2007), and art (Cohen 1999; Colton 2008; Norton, Heath, and Ventura 2011). The distinctive context of each of these concrete applications provides a novel perspective on the developing field of computational creativity. Further exploration of new domains provides additional viewpoints to help the theory mature. To this end, we present a creative system for recipe generation.

While work on recipes has been done in the field of artificial intelligence, to our knowledge, a recipe generation sys-

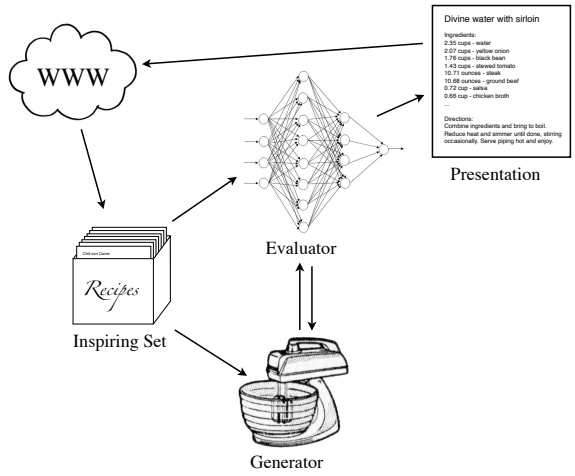


Figure 1: High-level view of the system architecture. Inspiring set recipes are taken from online sources and inform the evaluator and generator. Recipes are created through an iterative process involving both generation and evaluation. Eventually, generated recipes with the highest evaluation are fed to the presentation module for rendering and may be published online.

tem whose focus is *creativity* has not yet been developed (or even attempted). These other AI recipe generators use case-based reasoning to plan out a recipe, in the case of CHEF (Hammond 1986), or a meal, in the case of Julia (Hinrichs 1992). These approaches maximize the quality of a presented recipe without considering novelty, often preferring prior success to exploring new possibilities. The goal of our system is not only to produce a good recipe, but also to produce a *creative* one. This requires high quality as well as the development of *novel* artefacts.

2 PIERRE

Recipe generation is a complicated task that requires not only precise amounts of ingredients, but also explicit directions for preparing, combining, and cooking the ingredients. To focus on the foundational task of the type and amount of ingredients, we restrict our focus to recipes (specifically soups, stews, and chilis) that can be cooked in a crockpot. Crockpot recipes simplify the cooking process to essentially determining a set of ingredients to be cooked together.

We introduce a novel recipe generation system, PIERRE (Pseudo-Intelligent Evolutionary Real-time Recipe Engine), which, given access to existing recipes, learns to produce new crockpot recipes. PIERRE is composed primarily of two modules, for handling evaluation and generation, respectively. Each of these components takes input from an inspiring set and each is involved in producing recipes to send to the presentation module, as shown in Figure 1. In addition, the system interacts with the web, both acquiring knowledge from online databases and (potentially) publishing created recipes.

2.1 Inspiring Set

The inspiring set contains 4,748 soup, stew, and chili recipes gathered from popular online recipe websites[1]. From these recipes we manually created both a list of measurements and ingredients in order to parse recipes into a consistent format. This parsing enabled 1) grouping identical ingredients under a common name, 2) grouping similar ingredients at several levels, and 3) gathering statistics (including min, max, mean, variance, and frequency) about ingredients and ingredient groups across the inspiring set. Recipes in the inspiring set are normalized to 100 ounces.

The database of ingredients was explicitly partitioned into a hierarchy in which similar ingredients were grouped at a *sub*-level and these ingredient groups were further grouped at a *super*-level. For example, as shown in Figure 2, the super-group *Fruits and Vegetables* is composed of the sub-groups *Beans*, *Fruits*, *Leafy Vegetables*, and others. The sub-group of *Beans* includes many different types of beans including *Butter Beans*, *Red Kidney Beans*, *Garbanzo Beans*, and others.

Statistics are kept for each ingredient, including minimum, maximum, average, and standard deviation for the amount of the ingredient, as well as the probability of the ingredient occurring in an inspiring set recipe. These statistics are also aggregated at the sub- and super-group level, enabling comparison and evaluation of recipes at different levels of abstraction. In addition, gathering statistics at the group level provides for smoothing amounts for rare ingredients. Each statistic ω (min, max, mean, standard deviation, or frequency) for ingredients occurring less than a threshold in the set is linearly interpolated with the corresponding statistic of the sub-group, according to the following:

$$\omega = \begin{cases} \left(\frac{\alpha}{\alpha+\beta}\right) x + \left(\frac{\beta}{\alpha+\beta}\right) \xi & \text{if } \alpha < \theta \\ x & \text{if } \alpha \geq \theta \end{cases}$$

where x is the statistic of the ingredient, ξ is the statistic of the sub-group, α is the number of times the ingredient occurs in the inspiring set, β is the number of times any of the sub-group ingredients occur in the inspiring set, and the threshold θ is set to 100.

The inspiring set is used differently for generation than it is for evaluation. During artefact generation (Section 2.2) the inspiring set determines the initial population used for the genetic algorithm. During artefact evaluation (Section 2.3) the inspiring set determines which recipes and ratings are used as training examples for the multi-layer perceptron (MLP). Since the inspiring set is used in multiple ways, employing a different inspiring set for generating artefacts than the one used to evaluate artefacts can have useful effects.

2.2 Generation

PIERRE generates new recipes using a genetic algorithm acting on a population of recipes, each composed of a list of ingredients. The population is initialized by choosing recipes uniformly at random from the inspiring set, and the

[1] www.foodnetwork.com and www.allrecipes.com

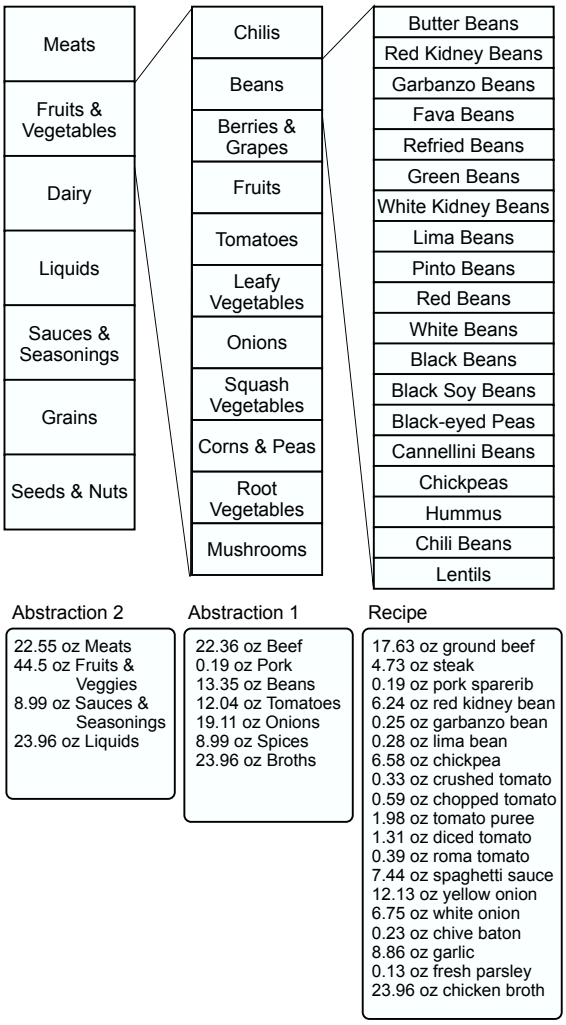


Figure 2: Above, a view of the ingredient hierarchy, showing the super-group (left), sub-group (middle), and ingredient (right) levels of abstraction. The *Fruits & Vegetables* super-group is expanded to show its sub-groups, including *Beans*, which is expanded to show its ingredients. Below, an example recipe is shown as it would appear at each level of abstraction.

fitness of each recipe is evaluated using the MLP evaluator described in Section 2.3. To produce each generation, a number of new recipes are generated equal to the number of recipes in the population. For each new recipe, two recipes are selected, with probability proportional to their fitness, for genetic crossover. The crossover is performed by randomly selecting a pivot index in the ingredient list of each recipe, thus dividing each recipe into two sub-lists of ingredients. A new recipe is then created by combining the first sub-list of the first recipe with the second sub-list of the second recipe.

After crossover, each recipe is subject to some probability of mutation. If a mutation occurs, the type of mutation is selected uniformly from the following choices:

- *Change of ingredient amount.* An ingredient is selected uniformly at random from the recipe, and its quantity is set to a new value drawn from a normal distribution that is parameterized by the mean and standard deviation of that ingredient's amount as determined from the inspiring set.
- *Change of one ingredient to another.* An ingredient is selected uniformly at random from the recipe, and is changed to another ingredient from the same super-group, chosen uniformly at random. The amount of the ingredient does not change.
- *Addition of ingredient.* An ingredient is selected uniformly at random from the database and inserted into a random location (chosen uniformly) in the recipe's ingredient list. The amount of the new ingredient is determined by a draw from a normal distribution parameterized by the mean and standard deviation of the ingredient amount as determined from the inspiring set.
- *Deletion of ingredient.* An ingredient is selected uniformly at random and removed from the recipe.

At the completion of each iteration, evolved recipes are re-normalized to 100 ounces for equal comparison to other recipes. The next generation is then selected by taking the top 50% (highest fitness) of the previous generation and the top 50% of the newly generated recipes. The rest of the recipes are discarded, keeping the population size constant.

Recipes 1 and 2 were generated using this process and were among those prepared, cooked, and fed to others by the authors. To produce these recipes, a population size of 150 recipes was allowed to evolve for 50 generations with a mutation rate of 40%.

2.3 Evaluation

To assess the quality of recipes, PIERRE uses an interpolation of two MLPs. Taking advantage of the (online) public user ratings of the recipes in the inspiring set, these MLPs perform a regression of the user rating based on the amount of different ingredients. The two MLPs are trained at different levels of abstraction within our ingredient hierarchy, with one operating at the super-group level and the other at the sub-group level. Thus, the model at the higher level of abstraction attempts to learn the proper relationship of major groups (meats, liquid, spices, etc), and the other model works to model the correct amounts of divisions within those groups.

Because we assume any recipe from the online websites is of relatively good quality, regardless of its user rating, we supplemented the training set with randomly constructed recipes given a rating of 0. These negative examples enabled the learner to discriminate between invalid random recipes and the valid ones, created by actual people.

Each MLP has an input layer consisting of real-valued nodes that encode the amount (in ounces) of each super-group (sub-group), a hidden layer consisting of 16 hidden nodes and a single real-valued output node that encodes the rating (between 0 and 1). The MLP weights are trained (with a learning rate of 0.01) until there is no measurable improvement in accuracy on a held out validation data set (consisting

Recipe 1 Divine water with sirloin

Ingredients:

2.35 cups - water
2.07 cups - yellow onion
1.76 cups - black bean
1.43 cups - stewed tomato
10.71 ounces - steak
10.68 ounces - ground beef
0.72 cup - salsa
0.66 cup - chicken broth
3.01 tablespoons - emeril's southwest essence
0.87 ounce - veal
1.22 tablespoons - white onion
1.22 tablespoons - diced tomato
1.17 tablespoons - red kidney bean
2.79 teaspoons - sambal oelek
0.22 clove - garlic
2.28 teaspoons - white bean
1.83 teaspoons - corn oil
0.29 ounce - pancetta
1.67 teaspoons - mirin
1.51 dashes - tom yam hot and sour paste
1.46 dashes - worcestershire
0.12 ounce - bologna

Directions: Combine ingredients and bring to boil. Reduce heat and simmer until done, stirring occasionally. Serve piping hot and enjoy.

Recipe 2 Exotic beefy bean

Ingredients:

2.2 cups - pinto bean
1.09 pounds - ground beef
1.6 cups - white onion
1.16 cups - diced tomato
1.13 cups - water
1.11 cups - chicken broth
0.77 cup - vegetable broth
0.63 cup - chile sauce
2.74 ounces - pork sausage
4.51 tablespoons - salsa
3.39 tablespoons - stewed tomato
1.43 ounces - chicken thigh
2.5 tablespoons - olive oil
1.09 ounces - hen
0.34 whole - red bell pepper
1.25 tablespoons - lentil
1.16 tablespoons - chopped tomato
2.87 teaspoons - red onion
2.03 teaspoons - garbanzo bean
1.65 teaspoons - cannellini bean
0.26 slice - bacon

Directions: Combine ingredients and bring to boil. Reduce heat and simmer until done, stirring occasionally. Serve piping hot and enjoy.

of 20% of the recipes) for 50 epochs. The set of weights used for evaluating generated recipes are those that performed the best on the validation data set.

2.4 Presentation

Colton (2008) has suggested that *perception* plays a critical role in the attribution of creativity. In other words, a computationally creative system could (and possibly must) take some responsibility to engender a perception of creativity.

In an attempt to help facilitate such a perception of its artefacts, PIERRE contains a module for recipe presentation. First, the module formats the recipe for human readability. Ingredient quantities are stored internally in ounces, but when rendering recipes for presentation, the ingredients are sorted by amount and then formatted using more traditional measurements, such as cups, teaspoons, dashes, and drops. Recipes are presented in a familiar way, just as they might appear in a common cookbook.

Second, the presentation module generates a recipe name. Standard recipes always have a name of some sort. While this task could be a complete work by itself, we implemented a simple name generation routine that produces names in the following the format: *[prefix] [ingredients] [suffix]*. This simple generation scheme produces names such as "Homestyle broccoli over beef blend" or "Spicy chicken with carrots surprise." The components of the name are based on prominent recipe ingredients and the presence of spicy or sweet ingredients. This simple approach creates names that range from reasonable to humorous.

3 EmPIERREical Results

To our knowledge, no other creative system has been designed to work in the recipe domain. As such, traditional concepts are highlighted in a new context. This new perspective admits additional analysis of the merits and nuances of theoretical ideas that have become generally accepted by the community. Here we evaluate the system with different combinations of inspiring sets, with and without a direct measure for typicality, and with and without the hierarchical definition of an artefact.

We measure novelty in a recipe by counting new combinations of (known) ingredients, n-grams. An n-gram is a combination of n ingredients. For example, a 2-gram would be *water-garlic*. A *rare* n-gram is an n-gram that does not occur in the inspiring set and does not contain a rare $(n-1)$-gram as a sub-combination (e.g., 4-grams containing rare 3-grams or, recursively, rare 2-grams are not included in the count of rare 4-grams). We define the *rare n-gram ratio* ρ_r^n for a specific recipe r as

$$\rho_r^n = \frac{\lambda_r^n}{\tau_r^n}$$

where τ_r^n is the total number of n-grams in r and λ_r^n is the number of those n-grams that are rare.

As another view of novelty, we consider a graph of ingredient amounts, which creates a visual profile of the type of recipes generated by the system. This comparison of visual

profiles was inspired by Faria and de Oliveira's use of a similar method in measuring aesthetic distances between document templates and generated document artefacts (Faria and de Oliveira 2006), and we found that it was easy to compare the outputs of the system based on the profiles that it generated.

3.1 Different Inspiring Sets for Evaluation

As mentioned, PIERRE can have different inspiring sets for both artefact generation and artefact evaluation. Thus the artefact initially generated would be inspired by one set of artefacts, but fitness would be determined by a fitness function inspired by a different set of artefacts. Using a combination of inspiring sets in the generative process hints at an idea which Buchanan identifies as "transfer" or knowledge sharing (Buchanan 2001), which refers to the notion that where two problems have simple, heterogenous representations, greater creativity can be achieved by transferring knowledge from one problem area to another. Although developing recipes from different inspiring sets may not constitute different problems in the same way as intended by Buchanan, the concepts and methods used by humans to develop recipes in one inspiring set may differ greatly from the concepts and methods used to develop recipes in a different culinary genre. Thus the knowledge used in the composition of artefacts in one inspiring set is introduced in the generation of new artefacts in a different domain, resulting in potentially greater creativity.

We experimented with various combinations of two inspiring sets. The first inspiring set included 4,748 soup, stew, and chili recipes crawled from the web (referred to as the "full" inspiring set). The second set is a subset of the first, including only the 594 chili recipes. The chili recipes were longer on average than the full recipes (13.97 ingredients as compared to 11.88 ingredients). We found no significant results from varying the generator's inspiring set therefore all reported experiments were conducted with a generator trained with the full inspiring set. We found that the recipes produced using the chili inspiring set to train the evaluator (hereafter referred to as the "chili evaluator") had a higher ratio of rare 2-grams and 3-grams (see blue lines in Figure 3) than those produced using the full inspiring set to train the evaluator (hereafter referred to as the "full evaluator", see red lines in Figure 3), and a relatively lower ratio of rare 4-grams and 5-grams. Because the system is using different inspiring sets to generate and evaluate recipes, it alters the original recipes to look more like the recipes found in the evaluator's inspiring set. In this context, generic soups or stews are being modified to look more like chilis. The resulting chilis retain some of the characteristics of the generic soups and stews, resulting in more novel combinations of ingredients and flavors (for chilis).

Systems which trained the evaluator with chili recipes produced recipes with a "chili" profile, as evidenced by more meat and vegetables, and less dairy and liquids (see blue lines in Figure 4). Systems which trained the evaluator with full recipes produced recipes with a marked "full" profile (red lines). This discovery suggests that a system's creativity can be guided through the use of different inspiring sets. Combining the use of different inspiring sets could introduce different flavor profiles, and allow the system to explore new parts of the recipe space.

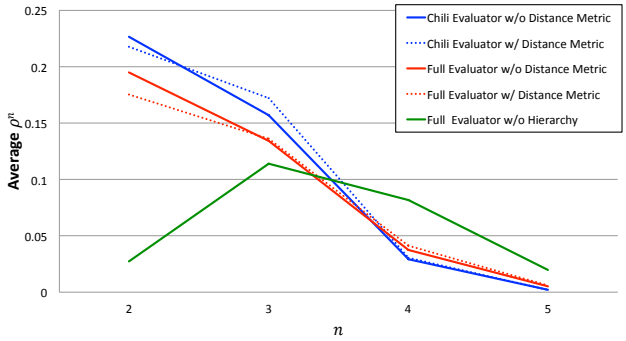


Figure 3: Average (over r) rare n-gram ratio for various values of n. Higher ratio values indicate increased novelty, with the chili evaluator producing the most novelty. Omitting the hierarchy noticeably reduces novelty, whereas including the distance metric has little effect.

3.2 Elimination of Explicit Typicality Metrics in the Fitness Function

We tested PIERRE with and without an explicit distance metric to essentially model a Wundt curve (Saunders and Gero 2001), promoting the generation of recipes that were neither too novel nor too typical. Although the theory can be interpreted to require an explicit evaluation of typicality (Ritchie 2007), in our experiments we found that removing the distance metric from our evaluation has no significant effect on the typicality or the novelty of our recipes (see the dotted lines in Figures 3 and 4). Explicitly measuring typicality is not necessary if typicality is implicitly modeled in the artefact generation process. In our system, ingredient quantities and ingredient counts were generated based on statistics found in the inspiring set. In addition, typicality is

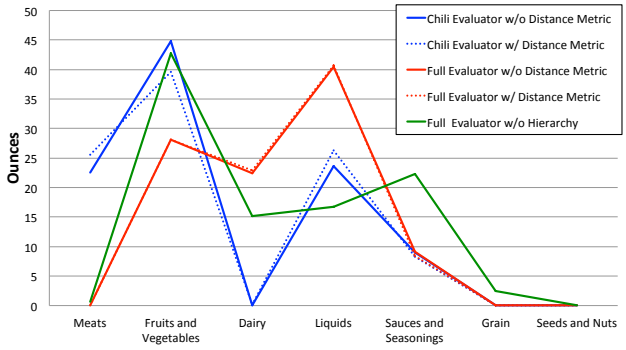


Figure 4: Ingredient amount (in ounces) of each supergroup. Different evaluators result in unique flavor profiles demonstrated by a visibly different recipe make-up. Omitting the hierarchy results in less extreme peaks and valleys and including the distance metric has little effect.

inherently imposed by both the generator and the evaluator. The generator selects new values based on normal distributions parameterized by statistics of the inspiring set, and if any recipe strays too far from what is typical, the evaluator assigns it a low fitness score.

3.3 Explicit modeling of the Ingredient Hierarchy

While modeling artefacts hierarchically perhaps seems like an obvious improvement for many creative systems, we compare our system with and without the hierarchy to validate that intuition. In terms of novelty (Figure 3), the recipes produced without the explicit hierarchical model (green line) exhibit fewer rare n-grams than the recipes produced using the hierarchy. Thus, the system is more capable of generating novel recipe combinations with the hierarchical model. This added novelty comes from the extra information that is introduced through the hierarchy. For example, when the system is generating a recipe, it can now know that, rather than a specific liquid, it really needs a type of liquid, thus allowing an interplay between typicality (maintaining the same sub/super-group) and novelty (trying a new member of the group). The system can then search for creative alternatives for the generic liquid that it needs to include in the recipe.

In addition, note that the ingredient amount profile of recipes created without the explicit hierarchy model (see the red and blue lines in Figure 4) exhibits less pronounced peaks or troughs than profiles for recipes generated using the hierarchy, suggesting that the hierarchy informs the system in generating recipes more characteristic or typical of a specific recipe genre. These results suggest that interpolated assessment of creative artefacts at multiple levels of abstraction is more effective at facilitating creativity than a unilateral assessment.

4 Discussion

One of the major contributions of this work is to provide a(nother) concrete implementation of many of the theoretical components called for in the literature, focusing on the area of computational recipe generation.

In doing so, we have presented a number of useful concepts, including the use of different inspiring sets for generation and evaluation, the implicit modeling of typicality (versus the notion of explicitly measuring it), and the abstraction of artefacts into a hierarchy for explicit use in both evaluation and generation.

The idea of evaluation-specific inspiring sets suggests a natural step towards meta-level creativity by providing a mechanism for changing evaluation criteria. In the context of recipe generation, this could include the ability for the system to change its "taste" over time, or, to use different inspiring sets to create different flavor profiles. Thus, just as the system has a more pronounced chili profile for its recipes when using the chili inspiring set for the evaluator, it could induce other types of profiles using other types of inspiring sets for evaluation and this affinity could vary over time. In general, the culinary arts provide a rich framework for varying preferences such as spice, sweetness, and texture, that could all be considered at the meta-level.

Table 1: Presentation survey results. Formatting recipes resulted in no significant difference.

Question	No Format	Format
Assuming I cooked on a regular basis, I would cook this recipe.	2.66	2.27
I think this recipe is creative.	3.59	3.28
I think this recipe is novel.	3.21	3.12
I think this recipe would be difficult to invent.	3.15	3.46
This recipe looks like it would taste different than anything I've previously tasted.	3.47	3.52
This recipe looks like it would taste good.	3.43	3.14

One significant area for future work is to incorporate the notion of goals and plans. As is, the system has a single goal: to create a high quality, high novelty chili. The different portions of the recipe space being explored by different parts of the population from the genetic algorithm could be seen as exploring different versions of that goal (for example, one part of the population would be predominantly chicken chili while another part would be predominantly vegetarian chili). However, the system would have more creative power if it could create and refine its own goal as it explores. Given user input, or even other factors (such as weather), the system could change its goal over time to be more appropriate to the given context.

In an attempt to assess the effectiveness of the presentation module, we hypothesized that the (admittedly simple) presentation effected by the system would make the artefacts more pleasing (than the raw, system versions of the recipes) to humans, and thus would increase the perceived creativity of the system. Thirty-eight participants were randomly given one of two surveys and asked to rate each of five recipes according to certain criteria. The first survey contained five unformatted recipes (no title, ingredients measured in ounces, without rounded quantities), whereas the second survey contained the formatted versions of the same five recipes (title, standard measurements, and rounded quantities). Contrary to our hypothesis, no statistically significant difference existed between the responses in the two groups (see Table 1).

Analysis of the survey lead us to an interesting (though perhaps retrospectively obvious) realization that the survey had not explicitly asked/forced the respondents to consider the creativity of the *written* recipe. Although the written recipes were quite different in each of the two cases, the *cooked* form of the recipes was the same for both formats. The effects of presentation on perceived creativity depend on which form of the artefact is being evaluated. In other words, were the survey participants evaluating the quality of the written recipe, or were they considering more what the cooked version would taste like? As another example of this idea, consider a system which generates musical scores. Is the creativity of artefacts produced by such a system determined from the written score or from the music that is heard

when a musician plays the score?

Though this question may seem trivial, consider that when 10 of PIERRE's recipes were posted on Food.com, the online community was outraged enough by some of the ingredient quantities (e.g., a dash of green beans)—which, though absurd by human standards, would not negatively affect taste—that even without considering the quality, or taste, of the cooked recipes, they removed the recipes from the site and suspended our account. The lack of typicality in the *written* recipe was condemned without considering that if cooked, the resulting chili would be considered typical (and possibly even tasty) by any reasonable measure.

We assert that many creative domains admit both a "written" and a "cooked" version of the artefact. Recognizing the distinction between the two may be essential to eliciting quantitative (or even qualitative) standards for evaluating creativity. Much of the work in computational creativity to date appears to have been on one level or the other—on the level of "written" artefacts perhaps because it is difficult to work at all at the "cooked" level or, possibly, because it is not obvious that there are, in fact, two distinct levels of artefact representation (e.g. visual art, perhaps?). Music is another good example of this phenomena. The sheet music is a written artefact, with instructions on how to produce the "cooked" artefact (or the actual melody). If there is no way to listen to the melody being played, then the artefact must be evaluated at the "written" level (using the sheet music alone). The idea of building modules which simulate forms of human perception has been explored to some extent in fields like computer vision, (audio) signal processing, haptic systems and the like, but in the context of computational creativity this sort of dual representation and evaluation is still largely unexplored at present, and this could represent a significant hurdle to establishing conventional tactics in evaluating computational creativity research (Jordanous 2011).

References

Binsted, K., and Ritchie, G. 1997. Computational rules for generating punning riddles. *HUMOR-International Journal of Humor Research* 10(1):25–76.

Buchanan, B. 2001. Creativity at the metalevel: AAAI-2000 presidential address. *AI Magazine* 22(3):13.

Cohen, H. 1999. Colouring without seeing: a problem in machine creativity. *AISB Quarterly* 102:26–35.

Colton, S. 2002. *Automated Theory Formation in Pure Mathematics*. Springer.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *Proceedings of the AAAI Spring Symposium on Creative Systems*.

Faria, A., and de Oliveira, J. 2006. Measuring aesthetic distance between document templates and instances. In *Proceedings of the ACM Symposium on Document Engineering*, 13–21. ACM.

Gervás, P. 2000. Wasp: Evaluation of different strategies for the automatic generation of spanish verse. In *Proceedings of the AISB-00 Symposium on Creative & Cultural Aspects of AI*, 93–100.

Gervás, P. 2001. An expert system for the composition of formal spanish poetry. *Knowledge-Based Systems* 14(3):181–188.

Gervás, P. 2011. Dynamic inspiring sets for sustained novelty in poetry generation. In *Proceedings of the Second International Conference on Computational Creativity*, 111–116.

Hammond, K. 1986. Chef: A model of case-based planning. In *Proceedings of the Fifth National Conference on Artificial Intelligence (AAAI-86)*, volume 1, 267–271.

Hinrichs, T. 1992. *Problem solving in open worlds: A case study in design*. Lawrence Erlbaum Associates.

Jordanous, A. 2010. A fitness function for creativity in jazz improvisation and beyond. In *Proceedings of the First International Conference on Computational Creativity*, 223–227.

Jordanous, A. 2011. Evaluating evaluation: Assessing progress in computational creativity research. In *Proceedings of the Second International Conference on Computational Creativity*, 102–107.

Lewis, G. 2000. Too many notes: Computers, complexity and culture in voyager. *Leonardo Music Journal* 10:33–39.

Monteith, K.; Francisco, V.; Martinez, T.; Gervás, P.; and Ventura, D. 2011. Automatic generation of emotionally-targeted soundtracks. In *Proceedings of the 2nd International Conference in Computational Creativity*, 60–62.

Norton, D.; Heath, D.; and Ventura, D. 2011. Autonomously creating quality images. In *Proceedings of the 2nd International Conference in Computational Creativity*, 10–15.

Pérez y Pérez, R., and Sharples, M. 2001. Mexica: A computer model of a cognitive account of creative writing. *Journal of Experimental & Theoretical Artificial Intelligence* 13(2):119–139.

Pérez y Pérez, R. 2007. Employing emotions to drive plot generation in a computer-based storyteller. *Cognitive Systems Research* 8(2):89–109.

Ritchie, G., and Hanna, F. 1984. AM: A case study in AI methodology. *Artificial Intelligence* 23(3):249–268.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17(1):67–99.

Saunders, R., and Gero, J. 2001. The digital clockwork muse: A computational model of aesthetic evolution. In *Proceedings of the AISB*, volume 1, 12–21.

Stock, O., and Strapparava, C. 2005. The act of creating humorous acronyms. *Applied Artificial Intelligence* 19(2):137–151.

Veale, T., and Hao, Y. 2007. Comprehending and generating apt metaphors: a web-driven, case-based approach to figurative language. In *Proceedings of the 22nd National Conference on Artificial Intelligence*, volume 2, 1471–1476. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Wiggins, G. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19(7):449–458.

Validation of Harmonic Progression Generator Using Classical Music

Adam Burnett
Cognitive Science
Simon Fraser University
Burnaby, BC Canada
ajb14@sfu.ca

Evon Khor
Cognitive Science
Simon Fraser University
Burnaby, BC Canada
ewk@sfu.ca

Philippe Pasquier
Interactive Arts and Technology
Simon Fraser University
Surrey, BC Canada
pasquier@sfu.ca

Arne Eigenfeldt
Contemporary Arts
Simon Fraser University
Vancouver, BC Canada
arne_e@sfu.ca

Abstract

We evaluate the output of a Markov model-based harmonic progression generator, a classic model for corpus-based computational creativity. 87 participants performed a discrimination task classifying 20 musical excerpts as either human-composed or computer-composed. Also recorded was each participant's level of confidence in their choice. Results indicated that while overall performance was above what would be expected from random guessing, further analysis revealed this was due to the human-composed pieces being much easier to identify than computer-composed pieces. Assessed separately, participants were unable to identify computer-composed pieces above chance-levels. We suggest improvements to the experimental design that could be implement in future evaluations.

Introduction

The trouble with evaluating artistic creativity is that it is difficult to establish objective criteria with which to judge the resulting creative artifact. This problem is compounded when the source of the artifact is *itself* a creative software. Computational creativity results in creations of creations, or *metacreations* (Whitelaw 2004), that differ from the artifacts we are used to encountering. As they are produced by machines that vary in their level of autonomy and in the amount of user-interaction they require in order to function, we must keep in mind a different set of considerations when we evaluate the resulting pieces.

The evaluation of aesthetics in metacreations is a fixture in the computational creativity literature (Eigenfeldt and Pasquier 2010; Eigenfeldt and Pasquier 2011; Pease, Winterstein, and Colton 2001; Stamp, Isenberg, and Carpendale 2007; Wallraven, Cunningham, Rigau, Feixas, and Sbert 2009). Whereas grading intelligence is a straight-forward matter of assessing how closely or quickly an individual can reach the optimum solution to a formally specifiable problem, there is usually no clear "goal" or "problem" that needs to be solved in a creative artwork; it often exists for its own sake, and to be enjoyed. When judging creative works such as music, it often comes down to relying on the subjective impressions of experts in the relevant domain (music critics, musicologists) or by quantifying subjective impressions in some measurable way (album sales, concert attendance).

Relying on entirely subjective measures is undesirable because it is not sufficient to simply test whether human listeners find computer-generated music creative or enjoyable or emotional: the mind is capable of finding patterns, design, and intention in random noise, deriving pleasure in the beauty of living organisms and ecosystems which were "designed" by the unguided, unintelligent processes of evolution, and sometimes even in randomness itself. To fairly and accurately assess the quality of computer-generated music, we must devise some sort of objective means to do so, even if that means indirectly measuring an effect of that creativity rather than directly measuring the creativity itself.

In the following we will describe previous attempts to evaluate computer-generated music, and then present our evaluation of the harmonic progression generator developed by Arne Eigenfeldt and Philippe Pasquier (2010).

Background

A problem is inevitably encountered when one tries to merge a scientific discipline concerned with objectivity, like Artificial Intelligence, with the subjectivity inherent in the creation, appreciation, and evaluation of art. This is the challenge for anyone proposing ways to evaluate machine creativity. As noted by Spector and Alpern (1994), there is no universally agreed-upon theory of aesthetic value within the artistic community. How then do we know when we have a *computational* artist? Approaches to this problem generally fall into two main camps. The first advocates a reliance on human judgements, particularly of the art-world and museum-going public, by holding art shows and getting feedback. However, this requires a lot of time and resources and is not always practical nor reliable. The other approach recommends the creation and application of codified, formalized evaluation criteria with which to judge a computational artist's creations. This method is especially popular in computer music evaluation as many forms of music can be formulated to follow a rule-system. There are three problems with this, however: 1) many existing formulations are "dead forms", which would penalize works in more contemporary genres for which detailed formalizations have not yet been established, 2) it is not evident that strict adherence to rules of a particular art-form or genre indicate aesthetic value. Meeting this criterion might indicate nothing more than aesthetically mediocre and boring, formulaic work, and 3) many formulations are rigid and

once established may not lend themselves to generalization across genres, essentially punishing novel works for being too original, even if they are of high quality.

Alternatively, Colton (2008) suggests that, rather than a focus on the input and output, *how* a creative work is produced is critical to its being perceived as creative. Colton asks us to consider the question of whether we label works as "creative" based on their quality, or whether we determine the quality of works based on how creative we found the process that generated them to be. Colton notes that in painting audiences are concerned with the process that led to the final product, and that this affects their enjoyment of the piece. In fact, it is noted that often the actual aesthetic quality of an artwork has little to do with how creative the work is perceived to be (consider Duchamp's *Fountain*, which was nothing more than a urinal). One conundrum which follows from this approach is that when *too little* is known about the process, we cannot evaluate whether or not it is creative, and if we know *too much* about the process, it is regarded as too mechanical and deterministic, leaving no room left for "creativity" to be exercised.

Colton presents a model of art appreciation, proposing that there are three judgements that consumers make about the creative process when determining how much they like a piece. These are: 1) the perceived effort required during the process, 2) the ingenuity of the process, and 3) the skill needed to carry out the process. From these Colton derived *The Creative Tripod*, which defines the three properties a system must possess in order to be judged as being creative: *skill*, *appreciation*, and *imagination*. Without skill, nothing can be produced; without appreciation, nothing of *value* can be produced, and without imagination, nothing *original* can be produced. The tripod analogy highlights the need for all three properties to be present in order of the label of "creative" to stand.

Whereas Colton directs attention on the process, Ritchie (2007) de-emphasizes the internal processes and favours focusing solely on the output. Ritchie argues that when assessing creative artifacts, we should be faithful to the traditional use of the word "creativity", which is tied to subjective human judgements. This, combined with the stance that we should only evaluate observable creative behaviours, levels the playing field and allows us to assess both human-produced and machine-produced creative works fairly and without bias. Ritchie warns that considering *both* the artifact and the process would introduce a fatal circularity: we would be left arguing that an artifact is creative because the process that produced it was creative, and that we know the process that produced it was creative because the artifact it produced was creative.

Discrimation Tasks

Pearce, Meredith, and Wiggins (2002) define four motivations for developing generative music systems: 1) to implement them as tools for personal use and/or 2) for general compositional use, 3) to provide theories of musical style, and 4) to provide cognitive theories of processes in compositional expertise. These four motivations can be collapsed into two general categories, the first of which is to use generative music systems as creative tools to produce original music, the other is to use these systems as a way to model theories of musical style and cognition. We will not be discussing the latter category any further here.

The problem of evaluating creativity mirrors a similar problem that befell early artificial intelligence researchers: *how do we evaluate machine intelligence*? It was difficult to say whether or not a machine could ever be said to think or demonstrate intelligence because there was little agreement on what those words would mean in the context. Today we face the same problem with machine creativity, unable to unanimously agree on what is meant by "creativity" in the question: *how do we evaluate machine creativity?*

Alan Turing (1950) famously suggested a way to tackle the problem. He had us consider a party game (the "imitation game") where a judge tries to determine which of two unseen players is pretending to be a woman; it is the job of the man to fool the judge by responding in the way he thinks a woman would, and it is the job of the woman to assist the judge in exposing him. With that in mind, Turing suggest that instead of trying to answer the impossible question of *can machines think?*, we should reformulate the question into something we can answer: *are there imaginable digital computers which would do well in the imitation game?* That is, could a computer program ever be designed that could successfully convince a judge that he was conversing with a human? This hypothetical procedure came to be known as the *Turing Test*. How this approach might be adapted for the problem of machine creativity is apparent: reformulate the question from whether or not a composition system is creative, and instead ask whether it does well in the "imitation game".

A popular method of evaluating generative music systems is to run a Turing-style test on the system's output (Boden 2010). This involves comparing computer-generated compositions to human-generated compositions through participant evaluation of the various pieces. Ariza (2009) cautions against the use of term "musical Turing Test" since intelligence of a generative music system cannot be determined by evaluating its output. The Turing Test has underlying assumptions on which it builds its criteria for machine intelligence: humans have minds, and natural language is sufficient to represent the mind; thus, if a machine is indistinguishable from a human in discourse, then it too has a mind. Joseph Weizenbaum's ELIZA is an early example of a system suited for the Turing Test. The ELIZA system was able to fool human interrogators; however, can we say that ELIZA is intelligent? John Searle's Chinese Room Argument suggests that a system is able to fool its interrogators without knowing anything about what it is doing; and so, having a façade of intelligence does not mean that the system is actually doing anything intelligent.

To test the outputs of generative music systems, it is possible to tweak the criteria of the Turing Test to accommodate the evaluation of musical outputs. Instead of having a text-based medium, sound symbols or forms are used. The

interrogator is replaced by a critic who may or may not interact with the system. Harnad (2000) labels tests of these sorts as toy Tests (tTs) instead of Turing Tests (TTs). In the Musical Output toy Test (MOtT), the critic is presented with musical pieces from two composer-agents. One of these agents is human, and the other, of course, is a machine. Based only on these works, the critic must attempt to distinguish the human from the machine.

Caution must be taken when interpreting the results of this type of test. For one, what criteria the listener uses when trying to discern between the machine- and human-composed pieces needs to be asked explicitly, and even so, listeners may fail to even be cognisant of their decision-making processes. Second, musical judgements are influenced by any combination of factors and can vary greatly from individual to individual. Furthermore, it is important to keep in mind that these tests are surveys of musical judgement and not of whether the system has thought or intelligence.

Boden (2010) cites David Cope's *Experiments in Musical Intelligence* (EMI) system, which generated music in the style of music contained in a supplied database, and notes how those listeners with some musical experience had difficulty determining whether the pieces it produced were human-composed or not. However, those with more extensive familiarity with, for example, Mozart were more readily able to distinguish true Mozart-composed pieces from the EMI-composed Mozart-esque pieces, though they there were unable to tell the difference between the EMI-composed pieces and human-composed pieces which were both meant to *mimic* Mozart's style. Even when EMI failed to perfectly mimic the intended style, it still was able to produce pieces that demonstrate proper compositional technique. Performance in a Turing Test thus varies heavily depending on exactly *what* is being tested (ability to mimic a particular composer? Ability to follow compositional conventions? Ability to produce interesting melodies?).

Another obstacle for Turing Tests is that they require the cooperation of the human judges. People have been known to retract their praise for computer-generated works upon learning of their synthetic source, protesting that it is requisite of art to have been produced by a human being, possessing all the facilities that enable one to express and communicate human emotion and experience. Some have refused to even give audience to a creative work knowing that it was produced by a machine, as happened to David Cope when debuting EMI. The reason cited is the belief that art requires creativity, and the belief that computers cannot be creative precludes computers from creating art. This prejudice will prevent some from ever accepting the results of a Turing Test, even if it is deemed internally successful.

Pearce and Wiggins (2001) proposed an objective framework for evaluating computer-generated musical compositions which, as they themselves point out, elicits comparisons to the Turing Test. The framework was developed in response to problems they identified with previous attempts to evaluate music composed by computer systems. They distinguished two kinds of evaluation: the *critic* and what we could call the *evaluation-proper*. The critic is part of the music-generating system itself and helps guide the development of the composition by evaluating the intermediate products of the system. The evaluation-proper is that which we are mainly concerned with here, and is unfortunately the more elusive of the two: it is the process and methods of establishing whether the compositions produced by the system satisfy the specified conditions of creativity.

Pearce and Wiggins highlight the necessity of objective measurements when evaluating machine creativity, as an objective approach to evaluation would be consistent with standard scientific investigation. Empirical science carries a respectable weight, and if a creative system could survive the sort of rigorous testing expected in scientific domains, then the results would be far more compelling than the wishy-washy, subjective evaluations seen elsewhere.

The existing evaluation methods Pearce and Wiggins reviewed are criticized for failing to confirm to these standards of objectivity, in part due to the presence of programmers' bias in the critic algorithms embedded in a number of the systems they discussed. They also note that subjective impressions are very imprecise and potentially unreliable: it is difficult to ensure that a group of human evaluators are following the same criteria.

In reaction to these shortcomings, Pearce and Wiggins layout a method of evaluating composition systems that maintain objective integrity. In order to be objective, a specific compositional aim must be explicitly established beforehand—if the goal is to mimic the style of a specific Baroque composer, the system should not receive a positive evaluation score because it is able to generate a realistic progression of 20th century jazz chords. To eliminate programmer bias, the critic should be derived from a pattern extracted from a data set of existing compositions using a machine learning algorithm. Once music is composed that is able to satisfy the critic, an evaluation using human participants is conducted. The participants should be played both music composed by the system and from the data set itself, and then tested to see if they are able to distinguish the two.

Having participants simply indicate whether they think a piece of music is computer-composed or not frees us from the subjective question of whether the computer-generated work is creative, enjoyable, or emotional, and instead allows us to home in on the *objective fact* about whether or not humans are able to tell whether the works are computer composed. Reformulating the question from *can this system produce creative works?* to *can this system produce works indistinguishable from the human composers in the data set?* creates a predictable, testable, and perhaps most importantly, a *clearly refutable* claim, as would be expected from a rigorous scientific experiment.

Experiment

In the framework of Pearce and Wiggins (2001), no at-

tempt was made to deceive the judges/participants about the nature of the experiment: participants were explicitly informed that they were comparing human- and machine-composed music. Along with this candid approach, our experiment resembled the framework described by Pearce and Wiggins in many additional ways. Our experiment differs however in that we compare the performance of both musicians and nonmusicians, and go beyond offering a simple binary choice between machine-composed and human-composed by providing a 4-point scale which will reflect both the participants' choice and their confidence in their choice.

In our study, we aim to evaluate the quality and particularly the robustness of the harmonic progression generator developed by Arne Eigenfeldt and Philippe Pasquier (2010). We sought to determine how successful the program is at generating harmonic progression of the same quality and style as human composers from traditional classical style periods. To do this, we had two groups of human participants, varying in their musical fluency, attempt to distinguish musical excerpts generated by the program from those written by human composers.

System Description

The system we are evaluating uses a third-order Markov model to derive harmonic progressions from a supplied corpus (Eigenfeldt and Pasquier 2010). This allows versatility as the particular rules from a style-period or genre do not have to be hard-coded into the program. Instead, the appearance of the rules emerge from the reliance on the corpus to guide the generation of progressions. By foregoing set rules, the system is not biased toward only producing progressions that follow traditional harmonic and voice leading conventions, but can just as competently function within the harmonic freedom of 20th century music if provided with a sufficiently rich corpus.

In contrast to many other music generators, the system was designed to function and respond to user request. The user is given the ability to specify the number of bars to be generated, a target bass line, the level and variation of harmonic complexity, and the voice-leading tension of the generated chords. These vectors help select the best candidate among the generated Markov conditional probability distributions of chord transitions. The system is written in MaxMSP and is available on its first author's webpage[1]. A full outline of the system is provided in Eigenfeldt and Pasquier (2010).

The corpus which serves as input to the system consists of pre-processed MIDI files: all musical content is reduced to a sequences of chords (and their durations) with controller data indicating the beginning of phrases and cadences.

Participants

The participants were recruited from Simon Fraser University and the University of British Columbia. Participation was incentivized by offering four $50 prizes to be distributed upon completion of the study.

To increase the resolution of our test of the harmonic progression generator, we compared the performance of two independent groups: musicians and nonmusicians. Much like a spoken language, well-written music is constrained by and emerges from conventions and rules and patterns. If one were to do a validation of a spoken or written language-generating program using human participants as judges, it would clearly be necessary to have the participants be fluent in the target language. It is for this reason that we found that in order to perform an accurate validation, it was critical to test the difference between musicians and nonmusicians in this task.

For the purpose of this experiment, only those with formal training in classical musical analysis were deemed "musicians"—mere proficiency with an instrument did not suffice. While instrumentalists are indisputably "musicians", we were exclusively interested in those students who have spent time studying and analyzing classical music scores and may have developed an ability to identify unusual harmonic choices and other errors that might arise in a machine-generated composition. Therefore, we decided that the musician group would consist of students who have received two or more years of classical musical training at a post-secondary institution. To ensure sufficient group-size, we also admitted those who have received at least 5th grade certification in the royal conservatory of music (or equivalent). The group consisting of laypeople (nonmusicians) was screened during the survey to ensure their musical naïvety.

Music Selection

Corpus For our study, we used harmonic progressions derived from a corpus of classical music (a mixture of Classical and Romantic style periods). Only chords already present in the pieces that made up the corpus found their way into the generated excerpts. We presented ten excerpts of music generated by the system and ten from classical pieces adapted to match the non-melodic presentation style of the computer-generated pieces.

The following is a list of the pieces that made up the corpus from which the computer-generated pieces were derived. The corpus is divided into five sections, each containing five to six pieces from the same composer to ensure consistency of style. We have tried to maintain consistence of form in our selections as well. Two of the pieces from each section were included in the survey as the human-composed pieces (marked in bold), and two progressions were generated from each section using the harmonic progression generator. As we are evaluating the quality of progressions produced by the system rather than determining the limits of its functionality, the setup described here corresponds to a typical use of the system.

Frédéric Chopin: **Nocturne in Eb Major Op. 9, No. 2**; Nocturne in F# Major Op. 15, No. 2; Nocturne in G minor, Op. 15, No. 3; Nocturne in Db Major, Op. 27, No. 2; **Nocturne in F major Op. 55, No. 1**.

[1] http://www.sfu.ca/~eigenfel/arne/main.html

*Antonín Dvoř*ák: **Humoresque**, Legend, **Slavonic Dance No. 1**, Slavonic Dance No. 2, Symphony No. 9 "From The New World" Second Movement, Valse Gracieuse.

Johannes Brahms: Symphony No. 1 In C Minor 3rd Movement, **Symphony No. 2 In D 3rd Movement**, **Symphony No. 3 in F 2nd Movement**, Symphony No. 3 in F 3rd Movement, Symphony No. 4 In E minor 3rd Movement, Hungarian Dance No. 5.

Felix Mendelssohn: **Consolation**, If With All Your Hearts, Spinning Song, O Rest In The Lord, **Scherzo in E Minor**, Venetian Boating Song (from Songs Without Words).

Robert Schumann: About Strange Lands And People, Träumerei, (from Scenes from Childhood), **The Happy Farmer** (from Album for the Young), **Piano Concerto in A Minor**, The Wild Horseman, Arabesque.

Processing The original Turing Test was not an assessment of a machine's ability to mimic speech, and neither was our experiment a test of the system's ability to creatively interpret and audibly produce music like a performer, but merely to compose it. Therefore, all pieces used in the experiment were "performed" and recorded using Kontakt Player (Native Instruments) and Cakewalk Sonar 4 (Cakewalk).

The system we evaluated requires pre-processing of the items in the data set: as the system is concerned only with analyzing and generating harmonic progressions, it was designed to receive as input MIDI files that conform to a specific format of block chords in closed position with the bass note separately specified. In the name of efficiency, rather than manually analyzing the harmony in our chosen classical pieces, we utilized "The Real Little Classical Fake Book" (Hal Leonard Corp. 1993), a large collection of classical themes transcribed for piano, and simply discard the melodic line and sequenced the harmonies and harmonic rhythms into MIDI files using the chord symbol realization plugin (which generates notes from chord symbols) for the Sibelius scorewriting software (Sibelius).

As we planned to test both musicians as well as nonmusicians, we recognized it was important that the human-composed pieces we chose be unfamiliar enough to reduce the likelihood that either groups would recognize their harmonic structure. Though we imagine that the elimination of the melodic line alone sufficiently obscured the identity of the pieces (as will transposition to a different key and changing the tempo), discretion was taken to ensure that pieces that obtain most of their notoriety from their harmonic sequences were excluded.

Determining which computer-generated pieces would make it into the final test was done by selecting those that most closely conformed to a pre-specified criteria. To ensure the feasibility of the study, it was decided to restrict the length of the harmonic sequences generated to around 8-bars and ensure that the progression ended on the tonic or dominant chord (regardless of the chord that preceded it), or in a cadence. The first four bars of one of the computer-generated pieces is presented in Figure 1. Note that the system encodes rhythm and can generate more than one chord per bar.

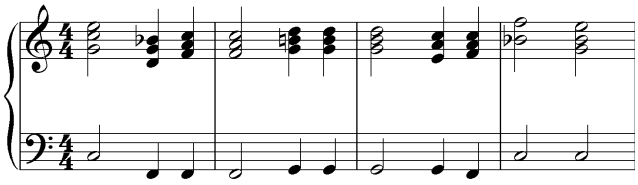


Figure 1. four-bar example of a "Brahms-inspired" excerpt.

Procedure

87 participants, a mix of students and faculty from Simon Fraser University and the University of British Columbia, were provided a URL to an online survey[1]. The survey was built using Drupal (Drupal) and a number of modules to enable audio playback and time tracking (to ensure that the participants were listening to the musical excerpts in full at least once). Participants were presented with a consent page indicating that their consent would be assumed by their completing the survey. They were then directed to a screen inquiring about their musical training (to enable us to assign them to the correct experimental group), followed by instructions detailing how to complete the survey. The instructions were upfront about the purpose and methods of the experiment; participants were informed that they would hear a mix of human-composed music and machine-composed music. No deceptive protocols were utilized.

Participants were presented one piece of music at a time, presented in a pre-established random order (limitations of the implementation prevented us from having each participant experience a unique ordering of questions). After listening, they were asked to rate the likely composer of each piece on a 4-point scale with 1 being "definitely human", 2: "probably human", 3: "probably computer" and 4: "definitely computer".

We decided to avoid using an odd-numbered scale for two reasons: we first wanted to discourage participants from disengaging from the task and choosing a neutral rating of 3 throughout. If participants lack certainty, this will be reflected in a greater proportion of "probably X" responses. We wanted to encourage participants to provide their best guess instead of defaulting to a safe "don't know" choice as it has been shown in other perceptual judgement tasks that participants underestimate their ability and that when forced to make a choice they often choose the appropriate response (Brown 1910). Gilljam and Granberg (1993) suggest that the presence of "don't know" options in questionnaires might encourage even those with definite opinions to chose the more cautious response. Poe et al. (1988) found that, when concerned with question testing factual knowledge, there is little difference in the responses

[1]The survey can be accessed at the following URL: http://magnum-interactive.com/metacreation

on questionnaires with and without a “don't know” option, but that excluding a neutral choice resulted in more usable data. A study by Alwin and Krosnick (1991) also found that including a “don't know” option did not improve the reliability of the results.

Secondly, the 4-point scale allowed us to reserve the ability to collapse the data into a binary human/computer choice, as well as compare the frequencies of “definitely X” to “probably X” selections between musicians and non-musicians later on for statistical analysis.

Following the questionnaire, participants were thanked for their participation and asked to indicate whether or not they recognized any of the progressions (and specify what they thought they were if they did), and what strategies (if any) they used to determine if the pieces were human- or computer-composed. Participants were then directed to a separate website where they provided their email address. Here they could indicate whether they wanted to be contacted about the results of the experience and/or be entered into the prize draw. As this section is separate from the survey-proper, it prevented us from matching survey answers to identifiable e-mail address, preserving anonymity.

We hypothesized that if the harmonic progression generator is capable of creating music of a quality and style similar to human composers, we should expect to see the performance of the two groups be similar to that which would result from random guessing (null hypothesis). If there *is* a detectable difference between the computer-generated and human-composed originals (that is: it is possible to distinguish the two), we should expect to see the nonmusicians perform either close to or slightly-above chance levels, and the musicians out-perform the nonmusicians with a performance even further from chance levels in the direction of correct classification.

We used one sample t-tests to compare musicians and nonmusicians to chance levels, and two sample t-tests to compare the mean scores of the groups. Tests were conducted using Bonferroni corrected alpha level of 0.005 (0.05/10). Comparisons were also made between these two groups' confidence with their choices as derived from their proportions of 1s and 4s compared to 2s and 3s on the 4-point scale.

Analysis of Data

Performance Participants were given four options when indicating their level of musical experience. They could specify that they had at least 2 years of a Bachelor's degree in music (Bachelor's), had achieved 5th grade certification in the royal conservatory of music or equivalent (Royal Cons.), had some unspecified formal musical training (Some), or no training (None). Our original “musician” category collapsed the data from the Bachelor's and Royal Cons. groups together, while the “nonmusicians” are made up of participants from the Some and None group. Table 1 shows the different groups' overall performance on the discrimination task.

Experience	*mean*	*t-score*	*p*	*df*
musicians	11.92 (3.27)	2.633	0.0164	19
nonmusicians	11.62 (2.46)	5.386	< .0001	66

Table 1. mean of correct answers out of 20, t-score (compared to chance), significance level, and degrees of freedom. Note. Standard deviations appear in paratheses.

As a test of our first hypothesis, one sample t-tests were used to compare performance to chance-levels (given that the questions were binary, 10 good guesses out of 20 is the mean for chance level). The results were not as anticipated: nonmusicians did significantly better than chance, leaving us unable to retain the null hypothesis (that the quality or style of the computer-generated pieces are indistinguishable from the quality and style of the human-composed pieces).

These data appear at first glance to be in the opposite directions of what we expected. Further analysis however revealed that by only looking at participants' total scores we had overlooked an interesting pattern buried in the data. Inspired by a comparable analysis conducted in Pearce and Wiggins (2001), when scores on identifying human-composed pieces were analyzed independently from scores identifying computer-composed pieces, a much different picture of the results emerged. Table 2 shows the results of this analysis.

Experience	*mean*	*t-score*	*p*	*df*
musicians (H)	6.550 (1.82)	3.808	0.0012	19
musicians (C)	5.300 (1.92)	0.698	0.4936	19
nonmusicians (H)	6.477 (1.51)	8.003	< .0001	66
nonmusicians (C)	5.089 (1.99)	0.369	0.7136	66

Table 2. results broken down by group and compositional source. (H) = human-composed. (C) = computer-composed.

When tested with a one sample t-test, this analysis shows that while participants were able to classify human-composed pieces (H) well above chance-levels (5 out of 10), their performance identifying computer-composed pieces (C) was *not* significantly different from chance-levels. Human-composed pieces were much more easily identifiable as human-composed than computer-composed pieces were identifiable as computer-composed. However, we still failed to see any statistically significant difference between the scores of the musician and nonmusician groups.

Confidence We also measure the level of confidence the

participants experienced for each question in the discrimination task. Confidence for each question was determined by assigning two points for "definitely" answers and one point for "probably" answers. A percentage was calculated using the score and the maximum possible score (thus a score of 100 would mean that the participant gave a "definitely" answer on every question). Average group confidence scores are indicated in Table 3.

Experience	*mean*	*n*
musicians (H)	67.25 (11.06)	20
musicians (C)	64.00 (12.84)	20
nonmusicians (H)	67.46 (12.83)	67
nonmusicians (C)	63.06 (11.18)	67

Table 3. confidence scores by group and compositional source.

While comparisons between groups' confidence are not statistically significant (there was no difference in confidence between experts and laymen), if the group means do in fact hint at a general tendency, they would indicate that participants are more confident about their answers when classifying human-composed pieces. This would be consistent with the analysis of performance that indicates that participants are likely to correctly identify these pieces as human-composed.

Strategies In the written response section of the survey, out of 87 participants, and out of only three who ventured guesses, only one correctly identified that they heard a progression taken from a Chopin piece (though they did not specific the piece's name). For the rest, participants seemed to rely on a number of factors to help them correctly identify the pieces' compositional source. Participants classifying human-composed pieces indicated that they listen for qualities such as *depth, clarity, complexity, feeling, life, regularity in rhythm, consonance, variety of dynamics, fluidity, subtly, repetition, pleasantness, simplicity*, and *logic*. When trying to identify computer-composed pieces, participants indicated that they listened for *repetition, simplicity, increased speed, odd resolutions, invariable rhythm, dissonance, lack of feeling, symmetry, rigidity, formality, awkwardness, logic, choppiness, static dynamics*, and *disorganization*. Interestingly, a number of these properties overlap: participants trying to identify both human and computer-composed pieces claimed to be listening for *simplicity* and *logic*, and participants within each condition often were looking for opposite properties to help identify the same source.

Conclusion

Participants listened to a series of 20 harmonic progressions and indicated whether they thought each was human- or computer-generated, along with a rating of their confidence for each choice. We hypothesized that participants would not perform significantly better than chance at this task.

Overall, participants *did* discriminate between the human-composed material and the progressions generated by the system. However, examining the results in more detail revealed something unexpected. When looking at participant responses to trials containing computer-composed progressions in isolation, it was found that participants were not capable of identifying the pieces generated by the harmonic progression system as computer-composed. Surprisingly, participants were nonetheless capable of identifying the human-composed pieces above chance levels. This results suggest that humans have a "natural" tendency to correctly recognize human-generated content. This would explain while our validation test failed. Further study would be needed to generalize this last finding. We believe that this tendency could be of interest to the computational creativity community as well as for cognitive sciences in general.

There are a number of changes to our experimental design that would be worth attempting in follow-up studies. The group sizes in the present experiment were quite heterogeneous, and the results seem to suggest that a larger number of participants qualifying for the Bachelor's group could provide us with valuable data.

We would also likely benefit from randomizing the presentation order of the musical excerpts or offering a more extensive "practice" section in future experiments. The collective results of all participants, plotted against time, gave a Pearson's correlation of $r = 0.54$, suggesting a significant practice effect.

Another concern is that we were not explicit enough when explaining our procedure. A number of participants tried to "outsmart" us and listen for superficial clues in the recordings, such as whether a real or synthesized piano was used, which evidently led them astray as both human-composed and computer-composed excerpts were created using the same equipment. We may also want to increase the duration of the excerpts as eight-bar phrases may be too short for listeners to be able to realistically gauge authorship.

We might consider abandoning the candid approach and instead employ an experimental paradigm that relies on deception, such as was done in Levisohn and Pasquier's evaluation of *BeatBender* (2008). This would rid us of the complications that arose from participants trying to over-dissect the musical excerpts for clues, and allow us to test for a larger range of properties. A limitation of our study was that it only asked participants to rate whether the pieces were human- or computer-generated; what, one could wonder, does this tell us about how successful the system was at being creative? If we modify the design and add additional criteria (e.g. ratings of *naturality*, *enjoyableness*, and *complexity* as was done in the evaluation of *BeatBender*) that parti-

cipants could listen for, it might tell us something more detailed about the differences between human-generated and computer-generated music.

Looking beyond the dichotomy of subjective judgements verses formalized criteria, there are arguably five levels of validation for artistic metacreations: the academic forum (whether the paper describing the creative system gets accepted or not), controlled evaluation (experiment such as those described in this paper), and feedback from journalists and critics, peers (artist from that community), and audiences. No evaluation study to our knowledge has attempted to cover all five of these levels. In future studies, we may consider rectifying this by adopting a methodology which would encompass all of these dimension, enhancing the validity of and confidence in our conclusions.

Acknowledgements

Thanks to Arne Eigenfeldt and David Mesiha for taking the time to give us a demonstration of the software and for providing troubleshooting correspondence.

References

Alwin, D. F., and J. A. Krosnick. 1991. *The reliability of survey attitude measurement:The influence of question and respondent attributes.* Sociological Methods & Research. 20 (1), 139–181.

Ariza, C. 2009. The Interrogator as Critic: The Turing Test and the Evaluation of Generative Music Systems. *Computer Music Journal*. 33(2), 48-70.Ariza, C. 2009. The Interrogator as Critic: The Turing Test and the Evaluation of Generative Music Systems. *Computer Music Journal*. 33(2), 48-70.

Boden, M. 2010. *The Turing test and artistic creativity*, Kybernetes, 39(3), 409–413.

Brown, W. 1910. *The Judgment of Distance*. Publications in Psychology. 1, 1–71.

Cakewalk. http://www.cakewalk.com.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *Creative Intelligent Systems: Papers from the AAAI Spring Symposium*, 2008. Menlo Park, CA: AAAI Press.14–20.

Drupal. http://drupal.org.

Eigenfeldt, A. and Pasquier, P. 2010. *Realtime Generation of Harmonic Progressions Using Controlled Markov Selection*. In *Proceedings of the First International Conference on Computational Creativity (ICCCX)*, ACM Press, Lisbon, Portugal, 16-25, 2010.

Eigenfeldt, A. and Pasquier, P. 2011. *Negotiated Content: Generative Soundscape Composition by Autonomous Musical Agents in Coming Together: Freesound*. In *Proceedings of the Second International Conference on Computational Creativity*, ACM Press, México City, México, 27-32, 2011.

Gilljam, M., and D. Granberg. 1993. Should we take "don't know" for an answer? *Public Opinion Quarterly.* 57, 348–357.

Harnad, S. 2000. Minds, Machines and Turing. *Journal of Logic, Language and Information.* 9(4), 425–445.

Hal Leonard Corp. 1993. *The Real Little Classical Fake Book*. Hal Leonard Publishing Corporation, Milwaukee, WI.

Levisohn, A. and Pasquier, P. 2008. *BeatBender: Subsumption Architecture for Rhythm Generation*. ACM International Conference on Advances in Computer Entertainment (ACE 2008), Yokohama, Japan, pages 51-58, ACM Press, 2008.

Native Instruments. http://www.native-instruments.com.

Pearce, M., Meredith, D., and Wiggins, G. 2002. Motivations and Methodologies for Automation of the Compositional Process. *Musicae Scientiae* 6(2), 119–147.

Pearce, M. and Wiggins, G. 2001. Towards a framework for the evaluation of machine compositions. In *Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in Arts and Science.* Brighton: SSAISB. 22–32.

Pease, A., Winterstein, D., and Colton, S. 2001. Evaluating machine creativity. In *Workshop on Creative Systems, 4th International Conference on Case Based Reasoning*. (ICCBR-01), Vancouver, British Columbia, Canada, 30 July-2 August. 56–61.

Poe, G. S., I. Seeman, J. McLauglin, E. Mehl, and M. Dietz. 1988. "Don't know" boxes in factual questions in a mail questionnaire: Effects on level and quality of response. *Public Opinion Quarterly.* 52, 212–222.

Ritchie, G. 2007. Some Empirical Criteria for Attributing Creativity to a Computer Program. *Minds & Machines.* 17, 67–99.

Sibelius. http://www.sibelius.com.

Spector, L. and Alpern, A. 1994. Criticism, Culture, and the Automatic Generation of Artworks. In *Proceedings of Twelfth National Conference on Artificial Intelligence* (Seattle, Washington, USA, 1994). 3–8. AAAI Press/MIT Press.

Stamp, A., Isenberg, T., and Carpendale, M.S.T. A Case Study from the Point of View of Aesthetics: A Dialogue Between an Artist and a Computer Scientist. In *Proceedings of Computational Aesthetics in Graphics, Visualization, and Imaging 2007* (CAe 2007, June 20–22, 2007, Banff, Alberta, Canada). (Aire-la-Ville, Switzerland), Eurographics Association. 129–134, 2007.

Turing, A. 1950. *Computing Machinery and Intelligence*. Mind. 59, 236 (Oct. 1950), 433–460.

Wallraven, C., Cunningham, D., Rigau, J., Feixas, M. and Sbert, M. 2009. Aesthetic appraisal of art - from eye movements to computers. *Computational Aesthetics 2009: Eurographics Workshop on Computational Aesthetics in Graphics, Visualization and Imaging*, 137–144.

Whitelaw, M. 2004. *Metacreation: Art and Artificial Life*. Cambridge, MA: MIT Press.

Automatic evaluation of punning riddle template extraction

Try Agustini and Ruli Manurung
Faculty of Computer Science
Universitas Indonesia
Depok 16424, Indonesia
try.agustini@gmail.com and maruli@cs.ui.ac.id

Abstract

This paper reports an empirical study to automatically evaluate the ability of T-PEG (Hong and Ong 2009) to extract joke templates by providing it with a corpus of punning riddles produced by another system, STANDUP (Manurung et al. 2008). This setup allows us to compare the extracted templates against the underlying data structures used by STANDUP in generating the corpus. In our setup, T-PEG is modified with a generalization component that clusters extracted templates based on structural similarity. These clusters are then compared against the underlying rules used by STANDUP to measure how well T-PEG is able to induce the schema used by STANDUP to generate the jokes. Whilst far from conclusive, an overall precision of 0.61 and recall of 0.763 suggests that T-PEG is able to extract some salient information regarding the underlying lexical relationships found within a punning riddle.

Background

The automatic construction of jokes, specifically punning riddles, as artefacts of linguistic creativity, has received quite a bit of attention in recent years. See (Ritchie 2005) for a good overview. Compared to other forms of linguistic creativity, such as stories and poems, punning riddles obey more formal structures, and hence are more amenable to automated construction.

Most of the existing systems such as JAPE (Binsted 1996) and STANDUP (Manurung et al. 2008) work from a predefined set of rules often called schemata, which forms the foundation of a joke, and which arguably encodes the humorous knowledge which makes the resulting text recognizably a joke. The schemata are typically handcrafted by the researchers based on analysis and observation of collections of exemplar jokes such as punning riddles.

T-PEG (Hong and Ong 2009) is a system which works along similar lines, but with a crucial difference: the schemata, or in T-PEG terms the 'templates', are automatically extracted from a corpus of exemplar punning riddles. Although (Hong and Ong 2009) report a manual evaluation of the extracted templates, it is subjective in nature and limited in scope. In this paper we present our work in automatically evaluating the template extraction functionality of T-PEG by providing it with example jokes produced by STANDUP, from which we can carry out a more extensive empirical study due to the fact that we have access to the underlying data structures used by STANDUP in generating the samples.

The rest of the paper is structured as follows. We first provide a brief overview of automatic punning riddle generation, describing the mechanisms of STANDUP and T-PEG, before discussing our proposed setup of how to carry out an automatic evaluation of the extracted templates. We then present the results and discussion thereof.

Punning riddle generation

STANDUP

Riddle generation in systems such as JAPE and STANDUP consists of several stages, where at each stage a particular kind of rule is selected and instantiated, i.e. schemas, description rules, and templates. Instantiation is performed by matching against a lexical database. In STANDUP, various lexical resources are utilized, among others Unisyn[1] for phonetic information and WordNet (Fellbaum 1998) for lexical semantic information.

A schema is composed of a header, which specifies a symbolic label and its input parameters, a set of lexical preconditions that specify phonetic, syntactic, and semantic relations that must hold between key lexical items, and a question and answer specification, which determines the output for the next stage (Manurung et al. 2008).

For example, the `newelan2` schema is defined as follows:

- Header: `newelan2(NP, A, B, HomB)`
- Lexical preconditions: `nouncompound(NP,A,B)`, `homophone(B,HomB)`, `noun(HomB)`
- Question specification: `shareproperties(NP, HomB)`
- Answer specification: `phrase(A,HomB)`

It states that a punning riddle can be generated if a set of four lexical items, i.e. `NP`, `A`, `B`, and `HomB`, can be found within a lexical database such that `A` and `B` form the compound noun `NP`, `B` is a homophone of `HomB`, and `HomB` is

[1] `http://www.cstr.ed.ac.uk/projects/unisyn`

a noun. For example[2], an instantiation in which NP = *computer screen*, A = *computer*, B = *screen*, and HomB = *scream*[3] could give rise (after the two further stages) to a riddle such as *"What do you call a shout with a window? A computer scream."*

Once a schema has been instantiated as above, the question and answer specifications are non-deterministically matched against a set of description rules, which encode linguistic variations that are warranted given the schema instantiation. These rules have a structure similar to schemas, in that they have a header, some preconditions, and an output expression called the template specifier. For example, one description rule is:

- Header: `shareproperties(X,Y)`
- Preconditions: `meronym(X, MerX), synonym(Y, SynY)`
- Template specifier: `merHyp(MerX, SynY)`

In the example joke given above, the question specification `shareproperties(computer screen, scream)` would match the header for this rule, where the values (`computer screen`, `scream`) would be bound to the local variables `X,Y` of the rule. Subsequently, the preconditions check further lexical relations to determine whether the rule is applicable. It may also instantiate additional variables, e.g. `MerX` and `SynY`, where in the example above `MerX` is bound to `window`, a meronym of `computer screen`, and `SynY` is bound to `shout`, a synonym of `scream`.

The answer specification `phrase(A,HomB)` trivially concatenates the instantiations, e.g. *computer scream*.

Finally, all these instantiations are passed on to the template-filling stage, where the template specifier `merHyp(MerX, SynY)` is non-deterministically matched against a set of canned text such as *"What do you call a * with a * ?"* where * indicates a slot to be filled with the instantiated words.

T-PEG

In the preceding section, we can see that the role of schemas, description rules, and templates is crucial in defining the humorous effect of the resulting riddle. In both JAPE and STANDUP, these rules were manually handcrafted. For instance, STANDUP has 11 schemas. (Hong and Ong 2009) present T-PEG (Template-based Pun Extractor and Generator), a system that generates punning riddles in a manner similar to JAPE and STANDUP, i.e. working from a set of symbolic rules that define a punning riddle, it instantiates certain key variables by selecting appropriate sets of words from a combination of lexical resources. The crucial difference is that whereas the symbolic rules used by JAPE and STANDUP were manually crafted, T-PEG relies on templates that are automatically extracted from a collection of exemplar jokes.

[2]The following worked example is taken from (Manurung et al. 2008).

[3]Homophony can be generalized to include pairs of words whose phonetic similarity is above a certain threshold.

Given a sample punning riddle, T-PEG constructs a template by firstly identifying nouns, verbs, and adjectives with the help of a part-of-speech tagger, and replacing them with what they term *regular variables*. Additionally, *similar-sound variables* are identified as words that are homophones of regular variables. Regular variables follow a naming convention where Xn indicates the n-th word in the question part and Yn indicates the n-th word in the answer part. Similar-sound variables are indicated by appending a dash and a number. All pairs of variables are then checked against a lexical resource to detect whether any semantic relations hold between them. In T-PEG, the lexical resources used are Unisyn, WordNet, and ConceptNet (Liu and Singh 2004).

For example, given the joke "How is a window like a headache? They are both panes", T-PEG can extract the template "How is a <X3> like a <X6>? They are both <Y3>", where the following word relations must hold:

- `Y3-0 SoundsLike Y3`
- `X3 ConceptuallyRelatedTo Y3`
- `Y3 ConceptuallyRelatedTo X3`
- `Y3 PartOf X3`
- `X6 ConceptallyRelatedTo Y3-0`
- `X6 IsA Y3-0`
- `Y3-0 ConceptuallyRelatedTo X6`

From the example above we can see that the notion of a template in T-PEG is equivalent to the conflation of a schema, description rule, and template in STANDUP.

The constructed templates are then filtered based on a graph-connectedness heuristic, i.e. if the variables are nodes and the word relationships are edges, a template must form a connected graph to be deemed a valid template.

Once the template has been extracted, generation proceeds in a similar fashion to STANDUP. In this paper we are less interested in the generation aspect of T-PEG, as it is largely similar to JAPE and STANDUP, and more interested in its ability to automatically learn or extract templates. Specifically, we are interested in assessing the ability of T-PEG to correctly identify the necessary and sufficient conditions for generating punning riddles. (Hong and Ong 2009) report a manual evaluation where a subset of the extracted templates was manually assessed by a linguist, whose job was to determine if the extracted templates were able to capture the 'essential word relationships in a pun'. The evaluation criteria are based on the number of incorrect relationships as identified by the linguist, and includes missing relationship, extra relationship, or incorrect word pairing. A scoring system from 1 to 5 is used, where 5 means there is no incorrect relationship, 4 means there is one incorrect relationship, and so on. They report an average score of 4.0 out of a maximum 5.

Automatic evaluation of template extraction

There are two issues concerning the manual evaluation of template extraction presented in (Hong and Ong 2009).

Firstly, we believe this evaluation is rather subjective. Although punning riddles are relatively simple and straightforward to analyse, the linguists were not the original authors of the jokes, and thus there is room for misinterpretation or incorrect emphasis. Furthermore, it is unquestionable that relying on the manual judgment of a linguist is both time-consuming and costly. The manual evaluation reported in (Hong and Ong 2009) was carried out on 27 templates generated from 27 jokes, which is a rather small sample from which to draw any conclusion.

Our observation is that if one had access to a large corpus of punning riddles that had somehow been annotated with the 'correct' word relationships that underlie the joke, one could assess T-PEG's template extraction functionality by comparing the resulting template against the reference word relationships. Unfortunately, we know of no such annotated resource that currently exists. However, we can use an existing punning riddle generator such as STANDUP to produce an approximation of such a resource, since we can access the underlying data structure of a generated punning riddle. In the joke generation module of STANDUP, the `JokeGraph` object of a generated punning riddle provides full access to the underlying lexical relationships.

Another issue we attempt to address is the fact that T-PEG makes no attempt at generalization of the extracted templates. Given fifty exemplar punning riddles, it will attempt to construct fifty templates. Hong and Ong state that this is beneficial 'to increase coverage'. However, we contend that if we are interested in building systems that computationally model the mechanisms of linguistic humor, coverage is not enough. A creative generative system should be able to generate artefacts from a limited set of symbolic rules. Thus, T-PEG should be able carry out some abstraction over the extracted templates, to yield a set of highly-productive patterns.

These two goals form the rationale of our proposed setup, which we discuss in the next section.

Proposed setup

As discussed above, the purpose of our experiment is to automatically evaluate the ability of T-PEG to correctly extract templates that underlie a collection of punning riddles. The proposed setup is as follows. Firstly, STANDUP is used to generate a large number of punning riddles. For each riddle, we note the actual rules used by STANDUP to generate them, which are used during the evaluation phase. The riddles are then given to T-PEG, which yields a template for each riddle. These templates are then organized into clusters using *agglomerative clustering* by calculating the similarity between templates using a structural similarity metric based on the semantic similarity evaluation function presented in (Manurung, Ritchie, and Thompson 2012). We then apply a simple majority rule to label the clusters, and then evaluate the template extraction process using the widely-used notions of precision and recall.

To achieve this, T-PEG first had to be modified by replacing its lexical and conceptual resources with those that were used in STANDUP, thus ensuring that the template extraction module would be able to identify the original lexical relationships in STANDUP.

Template clustering

The agglomerative clustering process starts with all templates belonging to singleton clusters. The distance of each cluster to all other clusters is then computed. The distance between two clusters is defined as the average distance of each pair of elements contained within the two clusters, also known as *average linkage clustering*. The two clusters with the shortest distance are then merged together. This process is repeated until k clusters remain, where k is provided as an input parameter.

In defining the notion of distance between two templates, we turn to the structure-mapping work of (Love 2000) and (Falkenhainer, Forbus, and Gentner 1989) that has defined a computational model of semantic similarity in terms of conceptual and structural similarity. Structural similarity measures the degree of isomorphism between two complex expressions. Conceptual similarity is a measure of relatedness between two concepts.

More specifically, we use the semantic similarity evaluation function used in (Manurung, Ritchie, and Thompson 2012), which implements a greedy algorithm based on Gentner's structure mapping theory (Falkenhainer, Forbus, and Gentner 1989). It takes two complex expressions, in our case two T-PEG extracted templates, and attempts to 'align' them in an optimal manner. It then applies a scoring function based on Love's computational model of similarity (Love 2000) to compute a score based on various aspects of the alignment. This function yields a distance of zero between two conceptually and structurally identical templates, and further distances for increasingly different template pairs.

Cluster labeling

We then automatically label the clusters using a simple majority rule. First, we define the *underlying schema* of a template to be the label of the STANDUP schema that was used to generate the punning riddle from which the template was extracted.

A cluster is then automatically labelled by identifying the underlying schemas of all its member templates, and simply choosing the schema that created the majority of templates within that cluster. If there are several underlying schemas that produced the same number of templates in a cluster, one is randomly selected. As an example, if a cluster contains 10 templates whose underlying schema is *lotus*, and 6 templates whose underlying schema is *newelan1*, then that cluster is labelled as representing the *lotus* schema.

Precision and recall

Using these cluster labels, we can compute measures which correspond to the widely-used notions of *precision* and *recall* in pattern recognition. In classification tasks, these measures are defined as follows:

$$Precision = \frac{tp}{tp + fp} \qquad Recall = \frac{tp}{tp + fn}$$

where in our case, given a cluster c with label l, tp (true positive) is the number of extracted templates that appear as members of c whose underlying schemas are l, fp (false positive) is the number of templates in c whose underlying schemas are not l, and fn (false negative) is the number of templates not in c but whose underlying schemas are l.

Precision and recall can be computed for each cluster, or as an aggregate measure over all resulting clusters.

Experimental setup

The experimental setup is as follows. Firstly, STANDUP is used to generate 20 jokes for each of 10 schemas, namely `bazaar`, `lotus`, `doublepun`, `gingernutpun`, `rhyminglotus`, `newelan1`, `newelan2`, `phonsub`, `poscomp`, and `negcomp`, resulting in a collection of 200 exemplar jokes. These jokes are then analysed by T-PEG, which yields 200 joke templates.

We then apply agglomerative clustering until 10 clusters are formed (since 10 STANDUP schemas are used). Our hypothesis is that for T-PEG to be deemed successful in extracting templates, it should be able to correctly organize the 200 templates into 10 clusters that correspond to the 10 STANDUP schemas. The precision and recall metrics should provide appropriate quantitative measures as to this goal.

Results and discussion

Table 1 shows the results of applying agglomerative clustering on the 200 templates into 10 clusters. The first column indicates the cluster number. The second and third columns specify the cluster membership, by indicating the number of templates with a given underlying STANDUP schema found within that cluster. The fourth column indicates the cluster size. The last column indicates the label assigned to that cluster using the majority rule described above. For example, cluster 1 contains 21 templates, 19 of which have `rhyminglotus` as their underlying schema, and 2 of which have `bazaar` as their underlying schema. Accordingly, it is assigned the label `rhyminglotus`.

Table 2 shows the precision and recall values computed for the clustering results. The first two columns indicate the cluster numbers and assigned labels, which correspond to the information in Table 1, and the last two columns indicate the precision and recall values computed for each cluster. Finally, the last row indicates the overall precision and recall. This aggregate result take into account the weighted averages given the different cluster sizes. Note that we collapsed clusters 1 and 6 because they were both labeled as `rhyminglotus`, and similarly, clusters 2 and 3 are collapsed due to the fact that they are both labeled as `bazaar`.

Finally, Table 3 shows the confusion matrix of how the templates are classified according to their underlying schema. The rows indicate the original underlying schemas, and the columns indicate the cluster labels. For example, of the 20 templates extracted from punning riddles generated using the `bazaar` schema, 14 are correctly found within a cluster labeled `bazaar`, 2 are found in a cluster labeled `rhyminglotus`, and 4 are found in a cluster labeled `lotus`.

No.	Schema	Amount	Total	Label
1	rhyminglotus bazaar	19 2	21	rhyminglotus
2	bazaar	11	11	bazaar
3	bazaar	3	3	bazaar
4	lotus newelan1 gingernutpun newelan2 doublepun bazaar	20 19 17 16 13 4	89	lotus
5	doublepun newelan2 gingernutpun	7 4 3	14	doublepun
6	rhyminglotus	1	1	rhyminglotus
7	newelan1	1	1	newelan1
8	phonsub	20	20	phonsub
9	poscomp	20	20	poscomp
10	negcomp	20	20	negcomp

Table 1: Results of agglomerative clustering

No.	Label	Precision	Recall
1 & 6	rhyminglotus	20/22=0.91	20/20=1
2 & 3	bazaar	14/14=1	14/20=0.7
4	lotus	20/89=0.225	20/20=1
5	doublepun	7/14=0.5	7/20=0.35
7	newelan1	1/1=1	1/20=0.05
8	phonsub	20/20=1	20/20=1
9	poscomp	20/20=1	20/20=1
10	negcomp	20/20=1	20/20=1
	Overall	0.61	0.763

Table 2: Precision and recall measures

From these results, we can see that only templates with `phonsub`, `poscomp`, and `negcomp` as their underlying schemas are perfectly identified. Templates with the underlying schemas `rhyminglotus` and `lotus` are correctly clustered together, but suffer some impurities with other templates also being deemed to belong to their clusters. Most notably, the cluster labeled `lotus` contains a very large number of templates from other schemas such as `bazaar`, `doublepun`, `gingernutpun`, `newelan1`, and `newelan2`. The cluster labeled `newelan1` contains only one template. No clusters were labeled as `gingernutpun` and `newelan2`.

A purely random baseline, in which the 200 extracted templates are randomly assigned to 10 different clusters, would yield an expected precision and recall of 0.1. Whilst this is an artificially low baseline, the results of an overall precision of 0.61 and recall of 0.763 suggests that T-PEG is able to extract some salient information regarding the underlying lexical relationships of a pun. However, certain underlying schemas are very difficult to distinguish from each other.

Upon further analysis, we can see that the problems arise from the fact that the templates extracted by T-PEG conflate

	bazaar	rhyminglotus	lotus	doublepun	gingernutpun	newelan1	newelan2	phonsub	poscomp	negcomp
bazaar	14	2	4							
rhyminglotus		20								
lotus			20							
doublepun			13	7						
gingernutpun			17	3						
newelan1			19			1				
newelan2			16	4						
phonsub								20		
poscomp									20	
negcomp										20

Table 3: Confusion matrix of clustering results

the various rules in STANDUP, i.e. schemas, description rules, and canned text templates. To illustrate the issues, observe the following two jokes, both generated using the `lotus` schema, and their resulting extracted templates:

Joke 1:

What do you call a cross between a firearm and a first step?
A piece initiative

The resulting template is:

What do you call a cross between a <X8> and a <X11> <X12>?
A <Y1> <Y2>

with the following word relationships:

- IsCompoundNoun(X11, X12)
- IsCompoundNoun(Y1-0, Y2)
- Synonym(X8, Y1)
- Synonym(Y2, X11:X12)
- Hypernym(X11:X12, Y1-0:Y2)
- Homophone(Y1-0, Y1)

From the joke we can see that the instantiations for the regular variables are X8=firearm, X11=first, X12=step, Y1=piece, and Y2=initiative. Furthermore, the similar-sound variable Y1-0 is bound to peace, because in WordNet, "peace initiative" is an instance of an initiative.

Thus, the word relationships state that "first step" and "peace initiative" are compound nouns, firearm is a synonym of piece, initiative is a synonym of "first step", which in turn is a hypernym of "peace initiative", and that lastly, peace and piece are homophones.

Joke 2:

What do you call a cross between an l and a correspondent?
A litre writer.

The resulting template is:

What do you call a cross between an <X8> and a <X11>?
A <Y1> <Y2>

with the following word relationships:

- IsCompoundNoun(Y1-0, Y2)
- Synonym(X8, Y1)
- Synonym(X11, Y1-0:Y2)
- Homophone(Y1-0, Y1)

From the joke, we can see that the instantiations for the regular variables are X8=l, X11=correspondent, Y1=litre, and Y2=writer. Furthermore, the similar-sound variable Y1-0 is bound to letter, because in WordNet, "letter writer" is a synonym for "correspondent".

Thus, the word relationships state that "letter writer" is a compound noun, l is a synonym of litre, correspondent is a synonym of letter writer, and litre and letter are deemed to be homophones.

From these two jokes and their resulting extracted templates, we can make several observations. Firstly, T-PEG correctly extracts the core lexical preconditions stated for the `lotus` schema, in that the 'punchline' must contain a compound noun Y1-0:Y2 ("peace initiative" and "letter writer", respectively), where the first word is replaced with a homophone, Y1 (in the first joke, piece, and in the second, litre).

However, the two jokes use different description rules for the question part. Whereas the former joke used a synonym (firearm) and a hypernym (first step) to describe the punchline, the latter used two synonyms, namely l and correspondent. Since T-PEG makes no distinction between word relationships arising from schemas or description rules, the choice of description rule, which is a somewhat trivial linguistic variation, leads the agglomerative clustering to falsely conclude that two jokes from different underlying schemas in fact use the same pattern.

Additionally, T-PEG extracted 'noisy' word relationships that play no part in the joke construction. Whereas the word relationships of the extracted template for the second joke capture exactly the necessary and sufficient conditions, in the former joke, nothing hinges on the fact that a "first step" is a compound noun, nor that initiative is a synonym of "first step". Such extraneous word relationships further pull the templates into wrong clusters.

From this experiment, we can conclude that although T-PEG may be successful at learning some joke templates given sample punning riddles, it is still making faulty assertions as to the constraints that specify what makes the riddle 'work' as a joke.

However, we speculate that much work can be done to repair such errors. The data redundancy contained within the specific templates that are clustered together, for instance, can be further analysed to form core relationships that are at the heart of the punning riddle structure. This is an avenue of future work that we intend to explore.

Summary

The manually constructed rules of STANDUP are specifically designed to maximize generative powers whilst retaining a fairly limited set of symbolic rules. T-PEG, on the other hand, tries to address the issue of having to handcraft rules by automatically extracting templates from example jokes. However, the evaluation of this functionality was

fairly limited given the difficulty and cost of manual evaluation. This paper has attempted to carry out a fairly novel methodology of automatically evaluating the performance of one creative system (T-PEG) using another creative system (STANDUP) to produce sample data with complete underlying annotations for comparing against. Although the results are far from conclusive, it corroborates the results of the manual evaluation in (Hong and Ong 2009) that claims T-PEG was successful in extracting templates from sample jokes. Furthermore, the experiment could shed more light on where T-PEG was still lacking in its ability to extract the underlying generative rules of the punning riddles.

Furthermore, the benefit is twofold: the template clustering process proposed in this work has attempted to address the generalizability of T-PEG. It is not sufficient to say that a huge number of templates will ensure maximum coverage. For a generative system to be deemed creative, it should be able to generate a high ratio of good quality artefacts from a limited set of symbolic rules.

By breaking down the patterns into schemas, description rules, and templates, and by stating the conditions when they can be composed together, STANDUP is able to produce a very wide range of different jokes, and in some sense can "explain" the craftsmanship behind its joke production facilities, as each individual component represents a specific function of the joke. T-PEG's templates, on the other hand, are monolithic structures that cannot be broken down into its functional components. This distinction is to be expected, given that the rules found within STANDUP were manually constructed, whereas T-PEG rules are automatically extracted. Nevertheless, this points towards a promising direction of future work in the automatic extraction of rules from creative artefacts.

As stated in the previous section, we believe that the template clustering process opens up the possibility of future work, i.e. by further leveraging the data redundancy contained with the resulting clusters. Further still, given that the clustering process makes use of the notion of distances between templates and clusters, one could imagine a technique that selects the template closest to the centroid of the cluster as being the most representative template for further generation. Finally, we would like to explore more sophisticated generalization techniques that would enable us to tease out the distinctions between component rules such as schemas, description rules, and canned text templates.

Acknowledgments

We would like to thank the creators of the STANDUP and T-PEG systems for making their software available for further experimentation and development. We would also like to thank the anonymous reviewers for their feedback and suggestions.

References

Binsted, K. 1996. *Machine Humour: An Implemented Model of Puns*. Ph.D. Dissertation, Department of Artificial Intelligence, University of Edinburgh, Edinburgh, UK.

Falkenhainer, B.; Forbus, K. D.; and Gentner, D. 1989. The structure-mapping engine: Algorithm and examples. *Artificial Intelligence* 41:1–63.

Fellbaum, C., ed. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Hong, B. A., and Ong, E. 2009. Automatically extracting word relationships as templates for pun generation. In *Proceedings of the 1st Workshop on Computational Approaches to Linguistic Creativity (CALC-09)*.

Liu, H., and Singh, P. 2004. Conceptnet: A practical commonsense reasoning toolkit. *BT Technology Journal* 22(4):221–226.

Love, B. C. 2000. A computational level theory of similarity. In *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*, 316–321.

Manurung, R.; Ritchie, G.; Pain, H.; Waller, A.; O'Mara, D.; and Black, R. 2008. The construction of a pun generator for language skills development. *Applied Artificial Intelligence* 22(9):841–869.

Manurung, R.; Ritchie, G.; and Thompson, H. 2012. Using genetic algorithms to create meaningful poetic text. *Journal of Experimental & Theoretical Artificial Intelligence* 24(1):43–64.

Ritchie, G. 2005. Computational mechanisms for pun generation. In *Proceedings of the 10th European Natural Language Generation Workshop*, 125–132.

Evaluating Musical Metacreation in a Live Performance Context

Arne Eigenfeldt
Contemporary Arts
Simon Fraser University
Vancouver, BC CANADA
arne_e@sfu.ca

Adam Burnett
Cognitive Science
Simon Fraser University
Burnaby, BC CANADA
ajb14@sfu.ca

Philippe Pasquier
Interactive Arts and Technology
Simon Fraser University
Surrey, BC CANADA
pasquier@sfu.ca

Abstract

We present an evaluation study of several musical metacreations. An audience that attended a public concert of music performed by string quartet, percussion, and Disklavier was asked to participate in a study to determine its success: 46 complete surveys were returned. Ten compositions, by two composer/programmers, were created by five different software systems. For purposes of validation, two of these works were human-composed, while a third was computer-assisted: the audience was not informed which compositions were human-composed. We briefly discuss the different systems, and present the artistic intent of each work, the methodology used in gathering audience responses, and the interpreted results of our analyses.

Introduction

The Musical Metacreation project[1] is an ongoing research collaboration between scientists and composer/musicians at Simon Fraser University that explores the theory and practice of metacreation – the notion of developing software that demonstrates creative behaviour (Whitelaw 2004). The objectives include not only developing software, but producing and presenting artistic works that use the software, and validating their musical success.

The research team includes a composer of acoustic and electroacoustic music who has created music composition and performance systems for over twenty years, an artificial intelligence researcher whose specialty includes multi-agent systems and cognitive modeling (and who is himself a creative artist in the field of computer music, sound design, audio and media arts), and several research assistants who are composers and/or scientists.

The fields of musical metacreation revolves around two central tasks:

- The composition task: the aim of this task is to produce music in the form of a symbolic representation, often a musical score. If the system takes existing compositions as input, it will be said to be corpus-based.
- The interpretation task: given some symbolic musical notation, this task consists of generating an acoustic signal.

Sometimes, these two tasks collide. For example, in electroacoustic music (in which we include electronica), an acoustic signal is directly generated as the output of the composition task. In the case of improvised music, composition and interpretation can be seen to happen simultaneously. The systems described in this paper, along with their evaluation, are all addressing the composition task.

[1] http://www.metacreation.net/

The creative systems produced by our research team have already been described in conference proceedings and journals, while the music produced has been presented in public concerts and festivals. On the surface, therefore, we could state that our work has already been validated; however, there are deeper issues involved that we discuss in this paper.

In considering how a metacreative system might be validated, there are at least five potential viewpoints that can be considered:

1. The designer: the designer of the system accepts the output as artistically valid;
2. The audience: the work is presented publicly, and the audience accepts the work;
3. The academic experts: the system is described in a technical peer-reviewed paper and accepted for conference or journal publication;
4. The domain experts: the system receives critical attention through the media or non-academic artists via demonstration;
5. Controlled experiments: the system is validated through scientifically accepted empirical methods, using statistical analysis of the results in order to accept or reject the hypothesis made about the system.

In the first instance, any artwork created by a human, and publicly presented, conceivably requires the artist to consider it complete and successful and representative of the artist's aesthetic vision. Similarly, metacreative works have, so far and to our knowledge, reflected the artistic sentiment of their designers. According to this viewpoint, the system evaluation is made directly by the designer. In our case, our metacreative systems have produced works that we find artistically interesting.

The second step reflects an artist's desire to share their work with the public. Whether the audience accepts, appreciates, or enjoys the work is, unfortunately, often difficult to ascertain, as many audiences will politely applaud any work. One could include more quantitative measures, such as audience counts, album sales, or online downloads.

The third case involves peer-review, albeit for a description of the system in technical terms. A different criteria is in place, one dependent less upon the artistic output, and more upon the technical contribution of the system in its novelty and usefulness. Often, the evaluation is also an evaluation of the originality and soundness of the process encoded in the system in regard to the computational creativity literature (Colton, 2008).

Both metacreation software and their output can be discussed in the media. Journalists and critics are different from the regular audience, in that their opinion will be further diffused to the audience: this may influence the audience judgment and the work can gain or lose notoriety as a consequence.

Lastly, empirical quantitative or qualitative validation studies can be undertaken that involve methods long supported by the research community for generating knowledge within the hard and soft sciences. While the computational creativity literature has started investigating these (Pearce and Wiggins, 2001; Pease et al., 2001, Ritchie, 2007, Jordanous, 2011), a great deal remains to be done.

While most previous work regarding the evaluation of musical metacreation (and computationally creative software in general for that matter) have been focusing on dimensions 1, 3 and 5, this paper presents an experimental study realized in the context of the public presentation of artworks in a concert setting (mixing dimensions 2 and 5). Also, there are very few instances of evaluation studies that consider more than one metacreative system at a time; our study is a comparative study of five different systems for computer-generated or computer-assisted composition.

The remainder of this paper discusses our evaluation study and the results we received, but also the questions that were raised. We first describe the different software systems involved, as well as the artistic intent of the compositions produced. We then present the methodology used in gathering audience responses to the compositions, as well as the results garnered from these responses. Finally, we posit our conclusions, as well as potential future work in this area.

Description

The public presentation of the metacreative software systems described in this paper took place as a public concert in December, 2011. Audience included members of the general public, as well some students of the first and third authors. Ten compositions, by two composers, were performed by a professional string quartet, percussionist, and Disklavier (a mechanized piano equipped to interpret MIDI input). The music was produced by five different software systems designed, and coded individually by the two composers. For comparison purposes, two of the pieces were composed without software; in other words, composed completely by human; a third was computer-assisted. The audience was informed beforehand that at least two of the works were human-composed, but were not informed as to which pieces these were; however, the program notes made it rather obvious that *fundatio* and *experiri* were, at most, computer-assisted. See Table 1 for a list of compositions.

The Systems and Compositions

In Equilibrio was generated by a real-time multi-agent system, described in (Eigenfeldt, 2009b). The system is concerned with agent interaction and negotiation towards a integrated melodic, harmonic, and rhythmic framework; its final output are MIDI events. The generated MIDI data was sent to a Yamaha Disklavier; no effort was made to disguise the fact that the performance was by a mechanical musical instrument. Along with the Disklavier and some high-level performance control by the composer, this system was responsible for both the "live" composition and its interpretation.

One of the Above consists of three movements for solo percussion. The music is notated by a system described in (Eigenfeldt and Pasquier, 2012). This system uses multiple evolutionary algorithms, including genetic algorithms, to control how a population of musical motives is presented in time, and how it is combined with other populations of motives. Intended for solo percussionist, the composition is a concentrated investigation in development of rhythmic motives. Each movement of the composition was presented separately, and treated as a unique composition within the evaluation. One additional movement, composed with the same intentions as the other three in this series, is human-composed (for reasons discussed in the Evaluation section).

Dead Slow / Look Left is a notated composition for string quartet and percussion, by a system that employs the harmonic generation algorithm described in (Eigenfeldt and Pasquier, 2010). The composition consists of a continuous overlapping harmonic progression generated using a harmonic analysis of 87 compositions by Pat Metheny, and a third-order Markov model based upon this analysis. In this corpus-based system, durations, dynamics, playing style, range, and harmonic spread were determined using patterns generated by a genetic algorithm. These continuous harmonies were interrupted by contrapuntal sections that interpret tendency masks (Truax 1991), which define such parameters as sequence length, number of instruments, subdivisions, playing style, number of playing styles, dynamics, and the number of gestures in a section.

Other, Previously was generated by a system described generally in (Eigenfeldt, 2009a), while the composition is described more fully in (Eigenfeldt, 2012b). A corpus of MIDI files – in this case 16 measures of the traditional Javanese ensemble composition *Ladrang Wilugeng* – was analysed, and generative rules regarding rhythmic construction was derived from the corpus. These rules were used by a genetic operator to create a population of ever-evolving melodies and rhythms that the system reassembled in a multi-agent environment over a rotating harmonic field. The real-time output was transcribed in a music notation program, and performed by string quartet. The end result is a piece of notated music that reflects many of the tendencies of the original corpus material, without direct quotation. The composer's role was limited to dynamic markings, orchestration, and assembling sections.

Gradual was generated by an extension of the system used to generate *One of the Above*, with an additional module to control pitch aspects integrated into the system. The final output was a notated work for marimba, violin, and Disklavier. While the system achieved the composition on its own, the interpretation was mixed: humans were playing the marimba and violin while the system was in charge of operating the Disklavier.

	Composition	Instrumentation	Experience Level		
			Expert	Novice	Combined
1	*In Equilibrio* [*c*]	Disklavier	3.17 (0.99)	2.71 (1.23)	2.90 (1.14)
2	*One of the Above #1* [*h*]	Solo percussion	4.00 (1.00)	3.36 (1.19)	3.67 (1.13)
3	*Dead Slow /Look Left* [*c*]	String quartet and percussion	4.16 (0.90)	3.08 (1.15)	3.51 (1.16)
4	*One of the Above #2* [*c*]	Solo percussion	3.68 (0.67)	3.16 (1.07)	3.42 (0.93)
5	*fundatio* [*h*]	String quartet	4.29 (0.80)	4.24 (0.83)	4.24 (0.81)
6	*experiri* [*c-a*]	String quartet	4.47 (0.61)	4.36 (0.86)	4.40 (0.76)
7	*One of the Above #3* [*c*]	Solo percussion	3.39 (0.76)	3.12 (1.20)	3.22 (1.04)
8	*Other, Previously* [*c*]	String quartet	4.31 (0.75)	4.50 (0.59)	4.40 (0.66)
9	*One of the Above #4* [*c*]	Solo percussion	3.63 (1.16)	2.71 (1.00)	3.10 (1.16)
10	*Gradual* [*c*]	Violin, marimba, Disklavier	4.05 (0.85)	3.88 (0.95)	3.93 (0.89)

Table 1. Individual composition engagement score means (out of 5). Standard deviations appear in parentheses. [c] = computer- composed. [h] = human-composed. [c-a] = computer-assisted.

fundatio and *experiri* were created by composer and software designer James Maxwell, with the help of his generative composition software that rests on a cognitive model of music learning and production. This software, *ManuScore*, is partially described in (Maxwell et al. 2009, 2011). *ManuScore* is a notation-based, interactive music composition environment. It is not a purely generative system, but rather a system which allows the composer to load a corpus, and proceed with that compositional process while enjoying recommendations from the system of possible continuations as suggested by the model.

fundatio was written using the commercial music notation software, Sibelius, following the compositional process used by the composer for many years, while *experiri* was written using *ManuScore*. Although this latter work remains clearly human-composed, the formal development of the music, and much of the melodic material used, were both directly influenced by the software.

Performances of the compositions can be viewed here:
In Equilibrio: http://youtu.be/x5fIdHbqEhY
Other; Previously: http://youtu.be/gaQfyhOiRio
One of the Above #2: http://youtu.be/gAIjQOiMG54;
One of the Above #3: http://youtu.be/bUYr7T7DKGs;
One of the Above #4: http://youtu.be/cQNQKinbJ-s.
Gradual: http://youtu.be/HZ2_Pr35KyU.
experiri: http://youtu.be/Gr5E7UVUoE8
fundatio: http://youtu.be/rNXt8b-kLMQ

Evaluation Study

The public concert was meant to serve two purposes: firstly, to present the artworks of the metacreative systems to the public, and secondly, to explore the idea of conducting evaluation in concert settings.

The opportunity for serious validation prompted the first composer to write an additional work separate from the metacreative systems, with the same musical goal. The purpose was not to fool the audience in making them guess which piece was not composed by machine, but rather to add human-composed material to the comparative study. While we hope that audiences will, one day, accept machine generated music without bias, Moffat and Kelly (2006) suggest this is not yet occurring. In our case, given three works for solo percussionist, composed in a particularly modernist style, it would be difficult to ascertain whether an audience's appreciation – or lack thereof – was due to the musical style, the restricted timbral palette, the lack of melodic and harmonic material, or any failings of the metacreative system. The human-composed piece allowed the composer to demonstrate the above-mentioned aspects, yet composed by the system designer. If the audience's rating of the human-composed piece was statistically similar to the metacreative works, it would demonstrate that the audience's preferences were based upon style, rather than musical creativity and/or quality.

Methods

Participants were 46 audience members from the general public (rather than only students) who attended a paid concert put on by Simon Fraser University. A program distributed to each audience member explicitly indicated that "machine-composed and machine-assisted musical compositions" would be performed. Each audience member also received an evaluation card on which they were encouraged to provide feedback. Audience members were asked to indicate, on a Likert-scale from 1 to 5, their level of familiarity with contemporary music, followed by ten similar 5 point Likert-scales regarding how "engaging" they found each piece to be. Audience members were also asked to indicate which three pieces they felt were the most directly human-composed. Audience members where also given space to write in their own comments. See Table 1.

Hypotheses

We hypothesized that the machine-generated and computer-assisted works were sufficiently similar in quality and style to the human-composed pieces that audience members would show no preference for the timbrally similar human-composed pieces (null hypothesis). This preference would be indicated by audience members' indication of how "engaging" they found each piece.

Analysis

In order to avoid alpha inflation that arises from multiple comparisons, statistical tests were made using post-hoc Bonferroni corrected alpha levels of .005 (0.5/10). For part of the analysis, the 46 audience members were divided into novice and expert groups depending on the score they indicated for the "familiarity with contemporary music" question. The "novice" group consisted of audience members that gave a score 1, 2, or 3 out of 5 on the familiarity scale (N = 25). The "expert" group consisted of the remaining audience members who gave a 4 or 5 (N = 19). Two audience members failed to provide a familiarity score, so their data was excluded from group comparisons.

Audience did not seem to discriminate between all the percussion pieces. Comparing the average engagement scores for the human-composed solo percussion piece *One of the Above #1* (M =3.59, SD = 1.15) with the average scores for the machine-composed *One of the Above #2* through #4 (M = 3.28, SD = 1.02) was not significant, t(44) = 1.43; p = .16 ns, leaving us unable to suggest that participants were able to discriminate between the human and machine-composed percussion pieces.

Audience did not "recognize" which piece was not computer-made. Assuming participants would find human-composed pieces more engaging, participants' engagement rating of the individual pieces were interpreted as an indication of whether participants could implicitly distinguish human-composed from machine-composed pieces. Tests comparing expert listeners' engagement scores for the human-composed *One of the Above #1* (M = 4.00, SD = 1.00) against the machine-composed alternatives (M = 3.57, SD = 0.88) were not significant (t(18)=1.68; p = 0.11 ns). Similarly, novice listeners' scores for *One of the Above #1* (M = 3.33, SD = 1.20) compared to the alternatives (M = 3.01, SD = 1.08) demonstrated no significant preference for the human-composed piece, t(23)=0.96; p = 0.34 ns.

Comparisons between the expert listener engagement ratings for the two string quartet pieces, the human-composed *fundatio* (M = 4.29, SD = 0.81) and the machine-assisted *experiri* (M = 4.47, SD = .61) were non-significant, t(18) = 1.00; p = .33 ns. Novice ratings for *fundatio* (M = 4.24, SD = 0.83) and *experiri* (M = 4.36, SD = 0.86) were similarly non-significant, t(24) = .72; p = .48 ns. This also failed to show that audience was discerning between the computer-assisted composition made using *ManuScore* and the human-made composition by the same composer.

Together, these results do not support the hypothesis that audience members were able to implicitly pick out which pieces were human-composed.

There was no difference between experts and novice choices. To determine whether audience members' ability to explicitly pick out the human-composed piece could depend on one's familiarity with contemporary music, a chi-square test compared novice and expert listeners' three "most directly human-composed" choices. The results of this test were non-significant, X2 (9, N = 113) = 14.17; p = .51 ns. This result fails to support the hypothesis that expert and novice listeners differ in their ability to explicitly discriminate human-composed pieces from machine-composed pieces.

Discussion

In addition to the above results, several further remarks can be made.

Overall, the evaluation results were pretty successful, showing both a rather high level of engagement from the audience, as well as good range with ranking means varying from 2.7/5 to 4.5/5. The audience did not discern computer composed from human-composed material, which seems to give credit to the five systems presented above. More precisely, this might just mean that the system were successful in portraying the goal, aesthetic and style of the two composers who developed them.

One further general observation that can be made is that while an evaluation in a concert setting allows us to capture the audience reaction to musical output in its "natural" presentation environment, it also introduces many variables that get us out of the usual controlled environment setting. The experimental protocol is also more difficult to follow.

On the other hand, controlled experiments are not the traditional setting in which a musical artwork is presented and this does introduce a number of biases in this type of evaluation. While these are well known, and solutions exist to circumvent them, our goal was to conduct an evaluation study in a live concert setting. We were concerned if conducting an evaluation in a concert setting would risk upsetting the audience's appreciation of the artwork. To our surprise, it did not seem to be the case, and the feedback forms were really welcomed. The whole process triggered a

longer than expected question and answer session at the end of the show. It is to be noted that very few audience members left before the end of the Q&A session.

Conclusions and Future Work

Finally, the whole process shed some light on the difficulty of evaluating computational creativity (and creativity in general). Artificial intelligence addresses the problem of emulating intelligence by having the computer achieve tasks that would require intelligence if achieved by humans. These tasks are usually formalized as well-formed problems. Rational problem solving is then evaluated by comparison to some optimal solution. If the optimal solution is theoretical and not attainable, optimization and approximation techniques can be used to get closer to the optimal, or at least improve the quality of the solution according to some metrics. Computational creativity is faced with the dilemma that, while creative behavior is intelligent behavior, such notions of optimality are not defined. It is often unclear which metrics need to be used to track progresses in the area. As demonstrated by this paper, it is at least an issue for the evaluation of composition systems.

Musical success is subjective in nature. This is why we resort to a comparative study capturing the relative level of success, rather than absolute ones. In the absence of formal metrics, we used human subjects to evaluate musical metacreation. However, creativity is a process (Boden, 2033). When evaluating a musical composition system, one particularly challenging aspect is that the system is capable of generating numerous pieces, with possibly varying levels of success: designing methodologies to measure that variability is an inherent challenge of the area. This is especially true when one has to use human subjects, since getting average relative evaluations of the average system production makes the experimental design particularly challenging.

To our knowledge, this paper is the first one to report on an evaluation experiment of machine-generated material conducted in real-world public situation. Beside the findings exposed above, the research instrument discussed here is a contribution in itself. As the systems presented are musical metacreations, validation and evaluation of such a system's output is itself a relatively novel and challenging research area. Our future work will continue to investigate and try to evaluate the methodologies to do so. Meanwhile, besides the finding exposed above, the paper raises a number of concerns and questions that will likely need further consideration in future work.

Acknowledgements

This research was funded by a grant from the Canada Council for the Arts, and the Natural Sciences and Engineering Research Council of Canada.

References

Boden, M. 2003. The Creative Mind - Myths and Mechanisms (2. ed.). Routledge I-XIII, 1-344

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In Creative Intelligent Systems: Papers from the AAAI Spring Symposium.

Eigenfeldt, A. 2009a. The Evolution of Evolutionary Software: Intelligent Rhythm Generation in Kinetic Engine. Applications of Evolutionary Computing, Berlin.

Eigenfeldt, A. 2009b. Multi-Agency and Realtime Composition: In Equilibrio. eContact 11.4 Toronto Electroacoustic Symposium 2009 http://cec.concordia.ca/econtact/11_4/

Eigenfeldt, A., Pasquier, P. 2010. Realtime Generation of Harmonic Progressions Using Constrained Markov Selection. Proceedings of the First International Conference on Computational Creativity, Lisbon.

Eigenfeldt, A., Pasquier, P. 2012a. Populations of Populations - Composing with Multiple Evolutionary Algorithms, P. Machado, J. Romero, and A. Carballal (Eds.): EvoMUSART 2012, LNCS 7247, 72–83. Springer, Heidelberg.

Eigenfeldt, A. 2012b. Corpus-based Recombinant Composition using a Genetic Algorithm. Soft Computing - A Fusion of Foundations, Methodologies and Applications. Springer Special issue on Evolutionary Music, forthcoming.

Jordanous, A. 2011. Evaluating Evaluation: Assessing Progress in Computational Creativity. Proceedings of the Second International Conference on Computational Creativity, Mexico City.

Maxwell, J., Pasquier, P. and Eigenfeldt, A. 2009. Hierarchical Sequential Memory for Music: A Cognitive Model. Proceedings of the International Society of Music Information Retrieval Conference, Kobe.

Maxwell, J., Pasquier, P. and Eigenfeldt, A. 2011. The Closure-based Cueing Model: Cognitively-Inspired Learning and Generation of Musical Sequences, Proceedings of the 8th Sound and Music Computing Conference, Padova.

Moffat, D., and Kelly, M. 2006. An investigation into people's bias against computational creativity in music composition. Third Joint Workshop on Computational Creativity, Riva del Garda.

Pearce, M. and Wiggins, G. 2001. Towards a framework for the evaluation of machine compositions. In Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in Arts and Science. Brighton: SSAISB. 22–32.

Pease, A., Winterstein, D., and Colton, S. 2001. Evaluating machine creativity. In Workshop on Creative Systems, 4th International Conference on Case Based Reasoning, Vancouver, 56–61.

Ritchie, G. 2007. Some Empirical Criteria for Attributing Creativity to a Computer Program. Minds & Machines. 17, 67–99.

Truax, B. 1991. Capturing musical knowledge in software systems. Interface. 20:3-4, 217–233.

Whitelaw, M. 2004. Metacreation. Art and Artificial Life. Cambridge, MA: MIT Press.

Critical issues in evaluating freely improvising interactive music systems

Adam Linson, Chris Dobbyn and Robin Laney
Faculty of Mathematics, Computing and Technology
Department of Computing
The Open University
Milton Keynes UK
{a.linson, c.h.dobbyn, r.c.laney}@open.ac.uk

Abstract

As freely improvised music continues to be performed, it also continues to be implemented in interactive computer systems. For the scientific study of such systems to be possible, it is important to ensure the fitness for purpose of available evaluation methods. This paper will review several approaches to evaluating interactive computer music systems. It will also examine the uncritically-accepted assumption that quantitative evaluation invariably yields significant data, irrespective of context. Ultimately, it will be argued that, for some interactive computer systems, such as those designed for freely improvised music, qualitative evaluation by experts is the most appropriate evaluation method.

Introduction

Freely improvising computer systems, modelled on an established musical practice that has been called "non-idiomatic" (Bailey 1980/1993), have been around since at least the 1990s (see Rowe 1993; Lewis 1999). There has been a significant amount of academic writing on the topic, including a chapter in *Machine Musicianship* (Rowe 2001), and an entire book, *Hyperimprovisation* (Dean 2003), dedicated to the topic of its subtitle, "*computer-interactive sound improvisation*". As freely improvised music continues to be performed, it also continues to be implemented in interactive computer systems (see, for example, Blackwell and Young 2004; Hsu 2005; Collins 2006). For the scientific study of such systems to be possible, it is important to ensure the fitness for purpose of available evaluation methods.

A significant amount of research is conducted on dominant forms of instrumental and computer music, which has led to a number of evaluation methods and technologies. For example, music with well-defined style-based rules that constrain melodic, harmonic, and/or rhythmic constructs lends itself to generation and analysis techniques based on traditional musical notation. However, less widely studied forms of music such as freely improvised music have different evaluation criteria, and thus pose unique problems to widely adopted approaches to musicological and computational analysis. In particular, for music such as free improvisation, formalisable musical rules and symbolic notation fail to account for the fundamental aspects of the musical practice.

Defining the practice of freely improvised music is not trivial. As MacDonald, et al. (2011) point out, "while there is no generally accepted single definition of improvisation, most accounts highlight the spontaneously generated nature of the musical material and the real-time negotiation of unfolding musical interactions". This characterisation is explicitly extended to cover contemporary improvisation practices including free improvisation. In Clarke's examination of creativity in performance (2005a), he refers to an empirical study on freely improvised music showing that the "interweaving of social and structural factors" serves a central role in such music. (For those unfamiliar with freely improvised music, the artists and recordings mentioned, for example, in Bailey 1980/1993 and Smith and Dean 1997 may provide a useful starting point.)

In considering research into computer music systems that have been developed to perform freely improvised music, it is important to find an appropriate method of evaluation that is well-suited to the context. When computer music systems for free improvisation are assessed according to inappropriate criteria, it can have a potentially stifling effect on the development of new approaches to such systems, as well as potentially devaluing existing effective systems. This paper will review several approaches to evaluating interactive computer music systems. It will also examine the uncritically-accepted assumption that quantitative evaluation invariably yields significant data, irrespective of context. Ultimately, it will be argued that, for some interactive computer systems, such as those designed for freely improvised music, qualitative evaluation by experts is the most appropriate evaluation method.

Evaluation methods in computer music research

Computer music researchers generally acknowledge the need to determine an evaluation method appropriate to their specific research. It is not always apparent, however, to what extent these methods apply to other research in the wider field of computer music. Stowell, et al. (2009) consider a number of quantitative and qualitative approaches to evaluating "live human-computer music-making" (Stowell, et al. 2009), although they do not consider generative systems. Collins (2008), on the other hand, considers approaches to evaluating generative systems, finding promise in approaches that take into account the relationship of software to musical output. These authors find musical improvisation significant enough to merit acknowledgement, although they do not engage with the evaluation issues unique to "player-paradigm" interactive improvising systems, that is, systems with "a musical voice which may be related to, but is still in an audible way distinct from, the performance of a human partner" (Rowe 1996).

Notably, Stowell, et al. (2009) favour studies of expert performers for the evaluation of interactive digital musical instruments that are under performer control, although they do not mention a logical extension of this view, namely, that the same approach can be extended to interactive systems that are not under performer control. Similarly, Pearse and Wiggins (2001) find experts to be capable evaluators of music with enumerable rules (such as period harmonisations), but they do not address the evaluation of music without enumerable rules, such as freely improvised music. Among these researchers, there is a clear recognition that expert human analysis has something to offer, although pragmatic concerns lead to the consideration of alternatives to using human experts, especially computational approaches. But while computational approaches to evaluation can be expected to yield appropriate results in some research contexts, in others, computer-based evaluation techniques may be in principle incapable of discovering evidence that is relevant to the investigation. In support of this claim, Collins (2008) acknowledges that computational analysis, in failing to address emergent features of complex musical output, may have a destructive effect on the (musical) object of study.

A quantitative approach to evaluating improvised music

Pressing (1987), in a comprehensive study of quantitative analysis and improvisation, concludes that while "idiomatic" (Bailey 1980/1993) improvisation such as jazz lends itself well to both macro- and microstructural quantitative analysis, in freely improvised music "the musical meaning is not well described" by the same quantitative analytical approach. To clarify Pressing's terminology, "macroanalysis uses the full panoply of devices from traditional music theory" (primarily those generally found in musicological analysis of composed works), and microanalysis addresses parameters more likely to be found in perception studies of expressivity, such as "interonset and duration distributions", dynamic contours, and "legatoness". Pressing devised a specialised model to account for some of the general structural features of improvised music, which he validated in quantitative empirical studies. In further studies, he found that while his model functioned effectively for analysing improvised jazz music, it could not be effectively extended to freely improvised music without an arbitrary (and thereby subjective) partitioning of "polyphonically overlapping phrase structures". When comparing a jazz improvisation and a free improvisation—both subjectively regarded by Pressing as aesthetically successful—he found that the jazz improvisation contained extensive quantitative evidence of "micro-micro" *and* "micro-macro" correlations (which thus appears to validate his subjective assessment); in the free improvisation, both types of quantitative correlation were "nearly completely absent" (Pressing 1987). His findings suggest that even when quantitative analysis succeeds in apparently similar musical contexts, it is not trivial to extend such analysis to evaluating freely improvised music, whether human- or computer-generated.

Music and qualitative analysis

In some performances, musical features that are apparently insignificant become significant in the course of an analysis. This poses a difficulty for approaches to analysing data that screen for features whose relevance has been determined in advance. A computer-based quantitative analysis cannot overcome this problem, despite other strengths in detecting specific correlations, statistical significance, or other quantitative constructs such as self-similarity. Thus, computational quantitative analysis is limited in what it can discover; it is often confined to providing answers about whether or not a given data set complies to a given rule set.

While computers are a powerful tool to rapidly sift through vast amounts of data, computational analyses are notoriously bad at picking out long-term dependencies or large-scale structures from a body of time-series data, in contrast to human experts. Consider, for example, an analysis by composer, musician, and historian Gunther Schuller (1958) of three Sonny Rollins saxophone solos, all from within a single performance of Rollins' piece, "Blue 7" on the album "Saxophone Colossus" (Prestige LP 7079). Part of Schuller's argument for the merits of Rollins' solos includes the assessment that they are not merely following the fixed harmonic chord progression, nor are they merely variations on a melodic theme, nor do

they merely fulfil both of these (quantitatively measurable) criteria, both of which are typically used to determine that a jazz solo is (formally) allowable. Rather, Schuller points out the musical significance of a number of creative decisions made during the solos. His argument, based on his background expertise in the field, is fundamentally qualitative, and is nonetheless extensively backed up with (in some cases, quantitatively measurable) material evidence (i.e., musicological specificity about particular pitches, phrases, rhythms, etc.). The structural features he identifies in his analysis stand in sharp contrast to those that could be discovered by rule-based quantitative approaches: he identifies semantic information in the particular configuration of musical elements, thus extending his analysis beyond the quantitative measurement of compliance to musical rules. (Another example of this distinction can be found in Clarke's analysis of Jimi Hendrix's "Star Spangled Banner"; Clarke 2005b, Chapter 2.) Other quantitative approaches, such as conducting a survey of listeners' opinions, require large enough sample sizes to find statistical significance, and thus may be applicable when studying the capacities of a given listener population. By determining what is relevant to an analysis, a given analytic approach not only investigates but also characterises the object of study.

Schuller's (1958) expert analysis contains a wide range of assessments that illustrate the strengths of qualitative analysis over quantitative analysis, computational or otherwise. Take, for example, his assertion that a musical phrase introduced by Rollins "at first, seems gratuitous", whereas later in the piece, it "becomes apparent that [the phrase] was not at all gratuitous or a mere chance result, but part of an overall plan". Or, for example, the notion that the final restatement of an initial theme "is drained of all excess notes" and that the "rests [in the original statement of the theme] are filled out by long held notes," serving both to end the piece and "sum up all that came before". His analysis even briefly isolates the Max Roach drum solo, pointing out that two musical ideas, a triplet figure and a snare roll, are built up through permutations and alternations into a complex solo; then, eleven bars after the drum solo has ended, the drummer interestingly and meaningfully re-uses these two elements "in an accompanimental capacity". These examples can be viewed as arguments for the significance of specific musical decisions—what was chosen, or, in some cases, not chosen—among an allowable range of options. For instance, several possible notes may fit a given chord, but there may be a significance to the particular note that is played, such as a long-term dependency that is outside the scope of a computational analysis. Furthermore, a particular note may be chosen over another because of a social connotation, in principle irreducible to a quantitative framework.

In general, assessments of musical significance are relative to listener knowledge and expectation, as well as being strongly affected by listening context (for an extended discussion of this point, see Clarke 2005b). Furthermore, differences in listeners' accounts may extend beyond traditional musicology, and new concepts may be introduced that were not built into the initial evaluation framework. This is not possible when a computer has been limited in advance to a particular analytic framework. Also, in contrast to a quantitative approach, differing assessments of the same material need not contradict each other. In Clarke's Hendrix example, three listeners assess the significance of a particular arpeggiation: Clarke hears a destructive melodic rupture verging on dissolution, another hears the bugle of a military funeral, and yet another hears a pattern of fingerboard traversal. For Clarke's example, an imagined computational analysis would run the risk of shifting the framework of significance to only what can be discovered computationally, potentially excluding *a priori* the three listener assessments. It is difficult to imagine what a computational or other quantitative approach could contribute in this case, beyond support (confirm that it is an arpeggiation; identify the statistical likelihood for the presence and location of the arpeggiation within the melody; confirm its similarity to a given bugle call; investigate melodic possibilities constrained by fingerboard layout). And even if in principle computational analysis could discover *any* item of significance, the necessity of making prior decisions as to what counts as significant is a profound limitation.

Clarke's engagement with musical meaning finds support in the empirical listener perception studies conducted by Deliege, et al. (1997). These studies identify two primary types of perceived musical cues: those that can be confirmed by consulting the musical notation—"'objective' cues (themes, registral usages, etc.)"—and, in contrast, "'subjective' cues, which have psycho-dynamic functions (impressions, for example, of development, or of commencement) which may be experienced differently from one listener to another and are not necessarily identifiable in the score" (Deliege et al. 1997). This account of cues highlights specific, narrowly-defined observations (such as development and commencement), as opposed to the broader semantic framework of Clarke. But both accounts point to the fact that different listeners experience the same musical material in different ways, underscoring the fact that human listeners may be sensitive to information that could otherwise be obscured by more constrained assessments of the same material.

It is not currently possible to computationally model the entirety of human listening possibilities. Thus, when a particular research question is framed to empirically validate a computational model of human listening, the boundaries of listening are constrained, for example, to investigate melodic or harmonic expectations. But for research questions that seek, for example, to uncover the inherent polysemy of a given guitar solo, the diversity of embodied cultural expertise captured by multiple

qualitative accounts is no less scientific, and likely more relevant to the question at hand, than a quantitative study.

The role of experts

Expertise is not necessarily confined to an unworkably small set of specialists. With respect to Clarke's example, the ability to recognise a particular bugle call or guitar fingering can be considered forms of expertise that are shared by many. In practice, these recognitions eluded and thus enhanced his own musicologically astute account of melodic dissolution. Returning to the topic of improvisation, Smith and Dean, in their extensive investigation of improvisation in the arts, suggest that with an improvised work, "the possibility of finite interpretation is not to be expected, or even desirable," and "the ideas of improvisors themselves are very interesting sources for the analysis and understanding of improvisation" (Smith and Dean 1997). The substance of their study is found in the differing perspectives of practising improvisors who are regarded as experts. As Clarke (2005a) states, "the boundaries between the mundane, the creative, and the unacceptably idiosyncratic are constantly shifting, and [...] their position and evaluative significance is a function of judgements made within a shifting cultural and historical context". If we define experts as those with significant experience operating within the given cultural and historical context of a musical practice, it follows that such individuals are better equipped to make effective evaluations about the practice being studied. Especially in light of the aforementioned centrality of the "interweaving of social and structural factors" in freely improvised music, an experienced improvisor is well-suited to serve as an expert qualitative evaluator, capable of attunement to both subtle and complex emergent criteria.

Although some aspects of freely improvised music are amenable to various quantitative criteria (such as those that borrow from compositional analysis, especially melodic and harmonic information), the unique aspects of the music being studied do not necessarily reside in such criteria (see Lehmann and Kopiez 2010). To identify shared features across classical compositions by a single composer, a quantitative analysis would likely suffice, because the melodic and harmonic information comprise a significant degree of what constitutes the compositions. On the other hand, with freely improvised music, Smith and Dean (1997) find that "a multiplicity of semiotic frames can be continually merging and disrupting during a 'free' [...] improvisation," which they find to be an essential characteristic of such music. This represents at least one finding that is more effectively discovered by qualitative human expertise. Furthermore, in their elaborate taxonomy of improvisation, Smith and Dean refer to what they term "stipulated" improvisation, which describes a type of improvisation that derives structure and characteristic style from stipulated aesthetic parameters that are internalised by a community of performers. According to their account, the "stipulated" approach does not fully exploit improvisation because it does not permit the "breaking, remoulding and rebreaking of such 'parameters'", as does freely improvised music, which fundamentally allows for the possibility of "reformulating the parameters on each occasion" (Smith and Dean 1997). Thus, for some complex objects of study, expert qualitative analysis should be recognised as fulfilling an essential role that, at times, can be empirically supported by quantitative means, but never entirely replaced by these means.

Research context and conclusion

Quantitative approaches certainly have independently useful scientific functions (such as examining physical mechanics or features of perception). Yet expert qualitative analysis has the potential to offer a set of results that may, in fact, be more relevant to the particular research being conducted. Unfortunately, qualitative study is often assumed to diminish scientific rigour, despite the well-known criticisms of quantitative studies concerning test bias, determination of statistical significance, and assumptions implicit in classifications and standardised procedures (Hammersley 2009).

Generally speaking, for empirical study, the research question ought to be the determinant of experiment design and evaluation. Among the varieties of computer music research, there are some computer music systems that are not interactive, such as systems designed to output rule-based compositions. In many of these cases, quantitative computational analysis may be the most practical approach to evaluating whether or not a given computer system is successful in achieving its aims, such as rule compliance. When listener surveys are used to evaluate system success, it may be appropriate to use discrimination tests of fixed musical material (for more on discrimination tests, see Ariza 2009). For computer systems that generate widely divergent musical material, studies that focus on the underlying software may offer results more relevant to some research questions (Collins 2008). Alternatively, for studies of interactive computer music systems, the human-computer interaction, rather than the music, may be at the centre of the research. For these studies, the relation between performer intention and system responsiveness is one area of investigation that benefits from both quantitative and qualitative study, such as looking into actual and perceived timing issues (Stowell, et al. 2009). However, when considering interactive computer systems that are not under direct performer control, there is no well-established evaluation method that is widely recognised in the literature.

For some studies of such player-paradigm systems, the focus may be on idiomatic music, for which the evaluation

approaches mentioned for generative composition systems are found to be applicable (Pachet 2002). But for studies of *interaction experience* with player-paradigm systems, it is essential to use expert qualitative analysis to avoid the danger of "measurement that fails to ensure that the assumptions built into measurement procedures correspond to the structure of the phenomena being investigated" (Hammersley 2009). It is a common aim of many studies of computer systems to iteratively improve a system based on assessments of its strengths and weaknesses. In the case of player-paradigm systems, expert qualitative evaluation can be used to identify even broadly defined—or potentially undefinable—weaknesses such as whether or not (and why) a human musical interaction with a system is, for example, "boring". Qualitative expert analysis in this context, though not widely acknowledged, is not entirely disregarded. For example, in Collins' brief account of a "free improvisation simulation" (2006), expert interview data is the primary source of evaluation.

It has been argued here that using qualitative data from experts is one way to approach the problem of evaluating a freely improvising computer music system. This approach is especially relevant for determining whether or not a player-paradigm system itself performs at the level of a human expert. Accounts of interaction experiences such as interview data can be collected, correlated, and analysed, with the aim of applying the data to improve the system. In practice, as part of a longer research program, qualitative data can function in the same manner as quantitative data: after identifying a system's strengths and weaknesses, a second iteration of the system can be built, and a follow-up study can determine what aims have been achieved. In this way, despite the predominance of quantitative evaluation in computer music, qualitative expert analysis can be a viable means of investigating phenomena, and qualitative studies can ultimately serve in making novel contributions to the research field.

References

Ariza, C. 2009. "The Interrogator as Critic: The Turing Test and the Evaluation of Generative Music Systems". *Computer Music Journal* 33:2, 48–70.

Bailey, D. 1980/1993. *Improvisation: its nature and practice in music*. Da Capo Press.

Blackwell, T. and M. Young. 2004. "Swarm granulator". *App. of Evolutionary Computing*: 399-408. Springer.

Clarke, E. 2005a. "Creativity in performance". *Musicae Scientiae* 9:1, 157-182.

Clarke, E. 2005b. *Ways of listening: An ecological approach to the perception of musical meaning*. Oxford University Press.

Collins, N. 2006. "Towards Autonomous Agents for Live Computer Music: Real-time Machine Listening and Interactive Music Systems". PhD Thesis. Centre for Science and Music, University of Cambridge.

Collins, N. 2008. "The analysis of generative music programs". *Organised Sound* 13, No. 3: 237-248.

Dean, R. T. 2003. *Hyperimprovisation: computer-interactive sound improvisation*. A-R Editions.

Deliège, I., M. Mélen, D. Stammers, and I. Cross. 1997. "Musical Schemata in Real-Time Listening to a Piece of Music". *Music Perception* 14(2): 117–60.

Hammersley, M. 2009. "Is social measurement possible, and is it necessary?" In: *The SAGE handbook of measurement*, Sage Publications Ltd.

Hsu, W. 2005. "Using timbre in a computer-based improvisation system". *Proc. of the ICMC*, Barcelona.

Lehmann, A. and R. Kopiez. 2010. "The difficulty of discerning between composed and improvised music". *Musicae Scientiae* 14: 113-129.

Lewis, G. 1999. "Interacting with latter-day musical automata". *Contemporary Music Review* 18:3, 99-112.

MacDonald, R., G. Wilson, and D. Miell. 2011. "Improvisation as a creative process within contemporary music". In: *Musical Imaginations: Multidisciplinary perspectives on creativity, performance and perception*, D. Hargeaves, D. Miell, and R. Macdonald, Eds. Oxford University Press.

Pachet, F. 2002. "The Continuator: Musical Interaction with Style" in *Proc. of the International Computer Music Conference* (ICMA), Gothenberg, Sweden.

Pearce, M. T., and G. A. Wiggins. 2001. "Towards a Framework for the Evaluation of Machine Compositions". In: *Proc. of the AISB'01 Sym. on Artificial Intelligence and Creativity in the Arts and Sciences*. Brighton: 22–32.

Pressing, J. 1987 "The Micro- and Macrostructural Design of Improvised Music". *Music Perception* 5(2): 133-172.

Rowe, R. 1993. *Interactive music systems: machine listening and composing*. MIT Press.

Rowe, R. 1996. "Incrementally Improving Interactive Music Systems." *Contemporary Music Review* 13(2): 47-62.

Rowe, R. 2001. *Machine Musicianship*. MIT Press.

Schuller, G. 1958. "Sonny Rollins and the challenge of thematic improvisation". *Jazz Review* 1, No. 11: 6-9.

Smith, H. and R. T. Dean. 1997. *Improvisation, Hypermedia and the Arts since 1945*. Routledge.

Stowell, D., A. Robertson, N. Bryan-Kinns, and M. D. Plumbley. 2009. "Evaluation of live human-computer music-making: Quantitative and qualitative approaches". *Int. J. of Human-Computer Studies* 67, no. 11: 960-975.

Towards a Mixed Evaluation Approach for Computational Narrative Systems

Jichen Zhu
Drexel University
Philadelphia, PA 19104 USA
jichen.zhu@drexel.edu

Abstract

Evaluation is one of the major open problems in computational creativity research. Existing evaluation methods, either focusing on system performance or on user interaction, do not fully capture the important aspects of these systems as cultural artifacts. In this position paper, we examine existing evaluation methods in the area of computational narrative, and identify several important properties of stories and reading that have so far been overlooked in empirical studies. Our preliminary work recognizes empirical literary studies as a valuable resource to develop a more balanced evaluation approach for computational narrative systems.

Introduction

Evaluation is one of the major open problems in computational creativity research. A set of well-designed evaluation methods not only is instrumental in informing the development of better creative computational systems, but also helps to articulate overarching research directions for the field overall. However, research in creative systems has encountered tremendous difficulties in defining suitable evaluation methods and metrics, both at the level of individual systems and across systems. A recent survey of 75 creative systems shows that, only slightly above half of the related publications give details on evaluation; among those, there is lack of consensus on both the aim of evaluation and the suitable evaluation criteria (Jordanous 2011).

Traditionally, methods for evaluating intelligent computational systems have been mainly developed in two areas: artificial intelligence (AI) and human-computer interaction (HCI). Following the scientific/engineering tradition, evaluation in AI typically relies on quantitative methods to measure the system's performance against a certain benchmark (e.g., system performance, algorithmic complexity, and the expressivity of knowledge representation). A salient example is the measure of "classification accuracy" in machine learning, where new algorithms are evaluated by being compared to standard ones over the same sets of data. Whereas the AI community is primarily concerned with the operation of the system itself, HCI concentrates on the interaction between the user and the system. Borrowing from psychology, human factors, and other related fields, HCI developed a set of quantitative and qualitative *user study* methods to understand the usability of a system along such principles as learnability, flexibility, and robustness (Dix et al. 2003).

Although these existing approaches offer useful insights into creative systems as functional and useful products, they do not fully capture a crucial property of creative systems, that is, they are and they produce cultural artifacts such as stories, music, and paintings. In these areas, there has not be an established tradition of formal evaluation. When we combine artistic expression and system building, evaluation becomes an issue. As Gervás observes in the context of computational narrative, "[b]ecause the issue of what should be valued in a story is unclear, research implementations tend to sidestep it, generally omitting systematic evaluation in favor of the presentation of hand-picked star examples of system output as means of system validation" (2009).

We argue that the difficulty of establishing an evaluation methodology in computational creativity research reflects the cultural clash between the scientific/engineering and the humanities/arts practices. Aligned with Snow's notion of the two cultures (1964), researchers working in the intersection of the two communities have observed the conflict of different and sometimes opposing value systems and axiomatic assumptions (Mateas 2001; Sengers 1998; Manovich 2001; Harrell 2006; Zhu and Harrell 2011). One of the differences is what Simon Penny (2007) calls the "ontological status of the artifact" between the electronic media arts practice and computer science research. For an artwork, the effectiveness of the immediate sensorial effect of the artifact is the primary criterion for success. As a result, most if not all effort is focused on the persuasiveness of the experience, which is built on specificity and complexity. In computer science, the situation is reversed. The artifact functions as a "proof of concept" and hence its presentation can be overlooked; the real work is inherently abstract and theoretical. These differences, Penny argues, illustrate that the insistence upon "alphanumeric abstraction," logical rationality, and desire for generalizability in science is fundamentally at odds with the affective power of artwork. In the context of evaluation, this conflict takes the form of the clash between the productivity- and value-based methodologies adopted by both AI and HCI communities, and the general resistance to empirical studies in the arts.

In this position paper, we present our initial work of developing a more balanced evaluation approach that takes into

account *both system and cultural* aspects of creative systems, focusing on computational narrative systems and their output. Our work is not intended to replace the function of literary criticism and close reading with empirical studies and statistical analysis. Simplistic attempts to reproduce art as a scientific experiment without an in-depth understanding of the former's tradition and value systems are shortsighted (as discussed in Ian Horswill's panel presentation at the Fourth Workshop on Intelligent Narrative Technologies, Palo Alto, 2011) and counter-productive to the long-term goal for computational creativity research. In the meantime, we also believe that evaluation is a critical process to inform the development of creative systems and to deepen the understanding of computational creativity. Therefore, more research and discussions about evaluation are needed.

In the rest of paper, we examine existing evaluation methods in the area of computational narrative, and identify several important properties of stories and reading that have so far been overlooked in existing evaluation methods. Our preliminary work suggests that empirical literary studies can be a valuable resource to develop a more balanced evaluation approach for computational narrative systems.

Existing Work on Narrative Evaluation

Broadly speaking, discussions of evaluating creative systems have taken place at two levels. At the level of computational creativity in general, researchers have attempted to come up with domain-independent evaluation criteria to measure a system's level of creativity, both in terms of its process and output. For example, Colton (2008) and Jordanous (2011) proposed standardized frameworks to empirically evaluate system creativity. The importance of these approaches is that, in addition to evaluating specific systems, they also allow potential cross-domain comparison between systems. At the level of specific creative domains, evaluations are conducted to validate a specific creative system and its output in that domain. For instance, the recent work by Vermeulen et al. in the IRIS project (2011) proposed a list of standardized, systematic assessment criteria for interactive storytelling systems using concepts that "play a key role in users' responses to interactive storytelling systems."

This section provides an overview of existing evaluation methods in the area of computational narrative. Our main focus is on the evaluation of story generation systems and their output, but some of our observations can also be applied to (non-generative) interactive digital storytelling systems. Recent examples of evaluating the latter type can be found in (Thue et al. 2011; Schoenau-Fog 2011). Although we do not specifically deal with high-level constructs such as 'novelty' and 'value,' we believe that more comprehensive evaluation criteria at the domain-specific level can indirectly contribute to the recognition and formulation of these high-level creativity constructs at the first level. Based on our survey of major text-based story generation systems, existing evaluation methods can be categorized into three broad approaches.

System Output Samples

As Gervás pointed out above, providing sample generated stories is one of the most common approaches for validating the system as well as the stories it generates. This approach started from the first story generation system *Tale-Spin* (Meehan 1981), where sample stories (translated from the logical propositions generated by the system into natural language by the system author) are provided to demonstrate the system's capabilities. In addition to successful examples, Meehan also picked different types of "failure" stories to illustrate the algorithmic limitation of the system for future improvement. Similarly, many later computational narrative systems such as *BRUTUS* (Bringsjord and Ferrucci 2000), and *ASPERA* (Gervás 2000) use selected system output for validation. Besides the lack of established specific evaluation metrics, the reason for the wide appeal of this approach is that it aligns with the tradition in literary and art practice where the final artifact should stand on its own without formal evaluation.

However, simply showing the "successful" and/or "interesting" output without explicitly stating the system author's selection criteria can be potentially problematic. Some recent work in this approach has attempted to make this selection process more transparent. For example, in the evaluation of the *GRIOT* system, Harrell (2006) evaluates the generated poems based on the quality and novelty of the metaphors they invoke. When the system generates "my world was so small and heavy," the author evaluates it by the metaphor it evokes — "Life is a Burden." Similarly, the *Riu* system (Ontañón and Zhu 2011) automatically assesses the generated stories by measuring the semantic distances of the analogies in the stories based on the WordNet knowledge base.

Evaluating the System's Process

The second approach is to evaluate the system primarily based on its underlying algorithmic process. Among the three evaluation approaches, this one is most aligned with traditional AI evaluation methods. Cognitive systems often use this approach to show that the system's underlying processes are cognitively sound. For instance, the evaluation of the *Universe* system (Lebowitz 1985) included fragments of the system's reasoning trace, along with the corresponding story output. It is intended to illustrate the system's capability to expand its plot-fragments library by generalizing from given example stories. Although the sample output and the process are relatively simple compared to those of the previous approach, Lebowitz intends to show, especially through the system processes, that the learning process is a necessary condition to creativity.

In a more complex example, the *Minstrel* system (Turner 1993), presented as a model for the creative process and storytelling, is evaluated in two ways. First, Turner evaluates the system by comparing it to related work in psychology, creativity, and storytelling. *Minstrel*'s process is contrasted to existing AI models of creativity both in the similar domain of narrative (e.g., *Tale-Spin* and *Universe*) and in different ones (e.g., AM (Lenat 1976)). Second, *Minstrel* is empirically studied in terms of its plausibility and quality as a test

bed for evaluating different hypotheses of creativity. Specifically, plausibility is evaluated based on 1) the quantity of possible output stories, by testing the system in different domains, and 2) the quality of output stories through a series of user studies (details in the next section). In the evaluation of the "test bed" criteria, Turner studies why some TRAMS (i.e., problem-solving strategies) were added, removed, etc. to prove that one can experiment with different models of creativity. For instance, to test its model of "boredom" as how many repeated elements are there in the stories, *Minstrel* was asked to generate stories about the same topic four times. The differences and similarities between these stories are analyzed to evaluate how boring these stories are.

User Studies

Evaluating the system's process alone, however, does not provide insights into the quality of the output. For systems that are more geared towards seeing narrative as a goal in its own right, user studies provide a way to assess the output story without counting solely on the author's own intuition. As a result, user studies has been increasingly adopted both as a standalone evaluation method and as a complement to other approaches.

For example, the *MEXICA* system (Pérez y Pérez and Sharples 2001) is evaluated through an Internet survey. The users rated seven stories by answering a set of 5-point Likert scale questions over five factors (i.e., coherence, narrative structure, content, suspense, and overall experience). Among these seven stories, four were generated by *MEXICA* using different system configurations (with or without certain modules). Two stories were generated by other computational narrative systems (i.e., *GESTER* and *MINSTREL*). The last story was written by a human author using "computer-story language." The scores each stories received is used to determine *MEXICA*'s level of "computerised creativity" (c-creativity) in reference to human writers and other similar systems.

In a more complex example, in addition to the methods mentioned above, the stories generated by *Minstrel* are evaluated through a series of independent user studies. In the first user study, users were given the generated stories, without being told that they were generated by a computer. Then they were asked to answer questions regarding their impressions of the author and the stories. In the second study, a different group of users repeated the above test, except the generated stories were rewritten by a human writer for better presentation with improved grammar and more polished prose. In the third study, the users were presented an unrelated story written by a 12-year-old and asked to answer the same set of questions.

User studies of narrative systems do no always adopt some form of the Turing Test. In the *Fabulist* system (Riedl 2004), the system author conducted two quantitative evaluations without using human writers as a benchmark. The first study evaluates plot coherence, measured based on the assumption that unimportant sentences decrease plot coherence. A group of users independently rates the importance of each sentence in the generated story and hence the coherence of the plot. Second, character believability is evaluated by asking users to rate the difference in characters' motivation in stories generated by two configurations of the system.

What is Missing

Computational narrative is still in its early stage, both in terms of the depth and breath of the narrative content. It is especially true when we compare these generated stories with what we typically conceive as literary text produced by human authors. In this regard, the different methods described in the previous section are arguably adequate for the current state of these systems. As argued above, however, evaluation methods play an important role not only in assessing existing systems, but also in informing what kind of future systems should be built. In this regard, waiting for the narrative systems to mature before starting to develop suitable evaluation criteria is detrimental to the research community.

As computational narrative research moves forward, a set of more comprehensive evaluation methods can help to reduce the gap between computer generated stories and traditional literature. *Our position is that* many important lessons from literary criticism and communication theory are by and large overlooked in computational narrative. We argue that they can be instrumental to developing evaluation methods that not only focus on the algorithmic and usability aspects of narrative systems, but also the expressiveness of the generated stories as cultural artifacts.

Below is our preliminary work in identifying some crucial elements that are missing in many existing evaluation methods. It is not intended to be seen as a comprehensive list, but rather as an initial step towards incorporating *fundamental* knowledge and concerns from related fields in the arts and the humanities.

Different Modes of Reading

Reading is a complex activity. Depending on the setting, purpose of the reading, and background of the reader, different aspects of the text are highlighted. Vipond and Hunter (1984) distinguished among point-driven, story-driven, and information-driven orientations for reading. Shown by recent studies in Reader Response theory (Miall and Kuiken 1994), ordinary readers typically adopt the story-driven approach, that is, to read for plot. They contemplate what characters are doing, experience the stylistic qualities of the writing, and reflect on the feelings that the story has evoked. This mode is adopted while we read for pleasure.

By contrast, the point-driven orientation is the foundation for literary criticism. Experts perform informed close reading — a complex act of interpretation at the linguistic, semantic, structural, and cultural levels — in order to understand the "point" of plot, setting, dialogue, etc. Point-driven reading assumes that the text is a purposeful act of communication between the author and the reader, and the "points" in the story have to be constructed through the reader's careful examination of the text.

Finally, in the information-driven orientation, a reader is more concerned about extracting specific knowledge from the text. We adopt this orientation while, for example,

following a recipe or checking facts in an encyclopedia. Information-driven reading places a strong emphasis on the coherence and informativeness of the text. This orientation is less common in computational narrative.

Different reading orientations place different emphasis on evaluation methods. As story-driven reading is primarily concerned with creating the "lived-through experience" for the reader, compatible evaluation needs to focus on the immersiveness of the story world. In computational narrative, most existing evaluation criteria presume the story-driven reading orientation and center on interestingness, presence, and engagement of the stories (e.g. plot coherence and character believability). Additionally, this orientation requires the participants of the evaluation to be close to an "average reader." A point-driven-based evaluation requires participants, usually experts, to perform more in-depth reading of the text beyond the surface plot. The effectiveness of different literary techniques, such as thematic structures, linguistic patterns, and points of view in the story can be evaluated in ways similar to traditional literary criticism.

To the best of our knowledge, there have not been attempts of point-driven-based evaluation in the context of computational narratives. There are many complex reasons for this. Some may argue that computational narrative, at its current stage, is too simple for this level of close reading. However, electronic literature (e-lit) work demonstrated that less algorithmically complex systems can still produce rich meanings. Establishing these evaluation criteria helps to develop a wider range of computational narrative.

Authorial Intention

Contradictory to the tradition of literary criticism, the evaluation of computational narrative systems has by and large ignored the intention of the authors. If we subscribe to the assumption that storytelling is a form of communication between the author and the reader, authorial intention should play a role in evaluating how effective these stories are. For instance, a user's report of unpleasantness may be positive or even desirable, if the system author intends to use her stories to challenge the reader's belief system, in ways similar to Duchamp's Urinal. A more balanced evaluation needs to differentiate this scenario from unpleasantness caused either by poorly written story or by unintuitive user interface. Similarly, intentional ambiguity in the story can be a powerful device, leaving something undetermined in order to open up multiple possible meanings. In the history of literature, intentionally ambiguous works such as Henry James's 1898 novel *The Turn of the Screw* have triggered many distinctive interpretations and vigorous debates about them.

Mixed Methods

A large percentage of the evaluations we surveyed gravitate towards quantitative methods with qualitative methods as a supplement, if at all. Through surveys and experiments, numerical data is collected, then analyzed statistically to provide an average user response. Although these methods have the clear advantage of being relatively easy to collect and analyze, they filter out the specificity and contextualization that is crucial to cultural artifacts.

Several research projects have attempted to address this issue. Mehta et al. (2007) devised an empirical study for the *Façade* system, which was intended by its authors to evoke rich exchange of meanings. Mehta et al. acknowledge that the standard quantitative criteria in the conversational system research community (e.g., task success rate, turn correction ratio, concept accuracy and elapsed time) are not adequate because they assume a task-based philosophy, where conversational interaction is framed as a simple exchange of clear, well-defined meanings. As a result, they made a deliberate choice to use more in-depth but less statistically significant ethnographic methods to study a small group of users' perceptions and interpretations of their conversations with non-player characters. Using video recording and retrospective interviews, their study found that participants created elaborate back-stories to make sense of character reactions in order to fill in the gaps of AI failures, an insight difficult to capture with pure quantitative methods.

The limitation of quantitative methods is echoed in Höök, Sengers and Andersson's user study of their digital art project (Höök, Sengers, and Andersson 2003). They observed, "[g]rossly speaking, the major conflict between artistic and HCI perspectives on user interaction is that art is inherently subjective, while HCI evaluation, with a science and engineering inheritance, has traditionally strived to be objective. While HCI evaluation is often approached as an impersonal and rigorous test of the effects of a device, artists tend to think of their system as a medium through which they can express their ideas to the user and provoke them to think and behave in new ways." As a response, their interpretive methods (open-ended interviews) focuses on giving the artists a grounded feeling for how the interactive system was interpreted and their message was communicated. Despite the sentiment against user studies in the interactive arts community, some artists involved in the project acknowledged that laboratory evaluations can help artists to uncover problems in interaction design.

Because of these limitations, we believe that a mixed methods approaches may be more suitable for evaluating computational narrative outputs. In addition to the closed-ended questions and surveys, qualitative methods such as phenomenology, grounded theory, ethnography, case studies can better capture the plurality of meanings interpreted by different readers and the complexity of such readings.

In literary studies, a group of researchers have started developing methods to empirically study readers' responses to literature. Due to the field's predisposition to point-driven interpretation, these methods offer a good example of balancing expert interpretation and ordinary readers' responses to and experience of the stories under evaluation. For example, Miall (2006) identified four kinds of empirical literary studies. First, studies that manipulate a literary text to isolate a particular effect. Second, studies that use an intact text in which the researchers hypothesize that intrinsic features of the text influence the reader. Instead of manipulating a text, each text itself provided a naturally varying level of foregrounding from high to low. A third kind of study involves comparison of two or more texts. Four, readers are asked to think aloud about a text during or after reading it. All of

these can be further explored and potentially incorporated into the evaluation of computational narrative systems.

Conclusion

In this position paper, we discussed the challenge of designing evaluation methods for creative systems due to their dual status. Focusing in the area of computational narrative, we surveyed existing evaluation approaches in story generation systems and identified crucial aspects of computational narrative, as a potential form of cultural artifacts, that have been so far downplayed. Penny warned us of the danger of the "unquestioned axiomatic acceptance of the concept of generality as being a virtue in computational practice especially when that axiomatic assumption is unquestioningly applied in realms where it may not be relevant" (Penny 2007). We suggest that work in empirical literary study research can offer valuable insights of developing more interdisciplinary and more balanced evaluation methods.

References

Bringsjord, S., and Ferrucci, D. A. 2000. *Artificial Intelligence and Literary Creativity: Inside the Mind of BRUTUS, a Storytelling Machine*. Hillsdale, NJ: Lawrence Erlbaum.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *Proceedings of the AAAI 2008 Spring Symposium in Creative Intelligent Systems*. AAAI Press.

Dix, A.; Finlay, J.; Abowd, G.; and Beale, R. 2003. *Human-Computer Interaction*. Edinburgh Gate, England: Prentice Hall.

Gervás, P. 2000. An expert system for the composition of formal spanish poetry. *Journal of Knowledge-Based Systems* 14:200–1.

Gervás, P. 2009. Computational approaches to storytelling and creativity. *AI Magazine* 30(3):49–62.

Harrell, D. F. 2006. Walking blues changes undersea: Imaginative narrative in interactive poetry generation with the griot system. In Liu, H., and Mihalcea, R., eds., *AAAI 2006 Workshop in Computational Aesthetics: Artificial Intelligence Approaches to Happiness and Beauty*, 61–69. Boston, MA: AAAI Press.

Höök, K.; Sengers, P.; and Andersson, G. 2003. Sense and sensibility: evaluation and interactive art. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 241–248.

Jordanous, A. 2011. Evaluating evaluation: Assessing progress in computational creativity research. In *Proceedings of the Second International Conference on Computational Creativity (ICCC-11)*, 102–107.

Lebowitz, M. 1985. Story-telling as planning and learning. *Poetics* 14(6):483–502.

Lenat, D. B. 1976. *Am: an artificial intelligence approach to discovery in mathematics as heuristic search*. Ph.d., Stanford University.

Manovich, L. 2001. Post-media aesthetics, available at http://www.manovich.net/docs/post_media_aesthetics1.doc.

Mateas, M. 2001. Expressive ai: A hybrid art and science practice. *Leonardo* 34(2):147–153.

Meehan, J. 1981. Tale-spin. In Riesbeck, C. K., ed., *Inside Computer Understanding: Five Programs Plus Miniatures*. New Haven, CT: Lawrence Erlbaum Associates.

Mehta, M.; Dow, S.; Mateas, M.; and MacIntyre, B. 2007. Evaluating a conversation-centered interactive drama. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, 8:1–8:8.

Miall, D. S., and Kuiken, D. 1994. Foregrounding, defamiliarization, and affect response to literary stories. *Poetics* 22:389–407.

Miall, D. S. 2006. *Literary Reading: Empirical and Theoretical Studies*. NewYork: Peter Lang.

Ontañón, S., and Zhu, J. 2011. On the role of domain knowledge in analogy-based story generation. In *Proceedings of the Twenty-Second International Joint Conferences on Artificial Intelligence (IJCAI-2011)*, 1717–1722.

Penny, S. 2007. Experience and abstraction: the arts and the logic of machines. In *Proceedings of PerthDAC 2007: 7th Digital Arts and Culture Conference*.

Pérez y Pérez, R., and Sharples, M. 2001. Mexica: A computer model of a cognitive account of creative writing. *Journal of Experimental & Theoretical Artificial Intelligence* 13(2):119–139.

Riedl, M. 2004. *Narrative Generation: Balancing Plot and Character*. Ph.D. Dissertation, North Carolina State University.

Schoenau-Fog, H. 2011. Hooked! evaluating engagement as continuation desire in interactive narratives. In *Proceedings of the Fourth International Conference on Interactive Digital Storytelling (ICIDS 2011)*, 219–230.

Sengers, P. 1998. *Anti-Boxology: Agent Design in Cultural Context*. Ph.D. Dissertation, Carnegie Mellon Universit.

Snow, C. P. 1964. *The Two Cultures*. New York: Menton Books.

Thue, D.; Bulitko, V.; Spetch, M.; and Romanuik, T. 2011. A computational model of perceived agency in video games. In *Proceedings of the Seventh Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*, 91–96.

Turner, S. R. 1993. *Minstrel: a computer model of creativity and storytelling*. Ph.D. Dissertation, University of California at Los Angeles, Los Angeles, CA, USA.

Vermeulen, I.; Roth, C.; Vorderer, P.; and Klimmt, C. 2011. Measuring user responses to interactive stories: Towards a standardized assessment tool. In *Proceedings of the Fourth International Conference on Interactive Digital Storytelling (ICIDS 2011)*, 38–43.

Vipond, D., and Hunt, R. A. 1984. Point-driven understanding: Pragmatic and cognitive dimensions of literary reading. *Poetics* 13:261–277.

Zhu, J., and Harrell, D. F. 2011. *Navigating the Two Cultures: a Critical Approach to AI-based Literary Practice*. Singapore: World Scientific. 222–246.

A Creative Improvisational Companion Based on Idiomatic Harmonic Bricks[1]

Robert M. Keller
Harvey Mudd College
Claremont, CA, USA
keller@cs.hmc.edu

August Toman-Yih
Harvey Mudd College
Claremont, CA, USA
August_Toman-Yih@hmc.edu

Alexandra Schofield
Harvey Mudd College
Claremont, CA, USA
aschofield@hmc.edu

Zachary Merritt
University of Central Florida
Orlando, FL, USA
zbmerritt@gmail.com

Abstract

We describe an improvisational companion based on the concept of harmonic bricks, as articulated by Cork and others. Our companion is software that can play background for, and trade melodies with, a human soloist. While exhibiting creativity itself, its greater purpose is to improve creativity of its user. Bricks, originally intended for memorization of chord progressions, are used here as a structuring device for improvised melodies within a tune, as a basis for interaction, and as a means of learning new grammars for the purpose of generating melodies by the companion. A user interface for a partially implemented system is presented.

Introduction

Jazz musicians often make use of *play-along* audio tracks to practice their improvisations. Such tracks feature a recorded *rhythm section* (e.g. drums, bass, and piano, organ, or guitar), with the solo part omitted. A well-known example is the Aebersold (1967) series, comprised of over 130 volumes and still growing.

One performance aspect helpful in practice is that of *trading*, wherein different soloists alternate playing over consecutive four or eight-measure segments of a tune. Our interest here is a computational companion in which the roles of the rhythm section and all but one improviser are played by the computer program. Such a companion does not require the assembly of other musicians and is essentially tireless. It can also be used as a basis for improving its user's understanding of the underlying theory and tune structure.

Because playing vs. listening alternate in small manageable chunks, trading can be a valuable learning device for the jazz musician. This paper proposes that trading can be structured using the concept of a *brick*, a shorthand term for an idiomatic harmonic phrase. While there are thousands of tunes that comprise the jazz literature, fewer than one hundred bricks suffice to describe most of the tunes. Thus significant intellectual economy is achieved by working on melodic lines for bricks vs. entire tunes.

We describe a preliminary implementation, with an outline for how bricks can also be used for machine learning, thereby improving the quality of the improvisational companion with time.

[1] The authors thank the NSF (CNS REU #0753306) and Harvey Mudd College for their generous support.

Background and Related Work

Jazz solo improvisation most commonly consists of a soloist spontaneously creating and playing a melody over a chordal and rhythmic background provided by the rhythm section. The harmony typically consists of a chord progression from a standard tune. In the vernacular of the jazz musician, the progression is referred to as "the changes", i.e. the transitions from one chord to another. Negotiating the changes while playing is one of the aspects of jazz that provides both pleasure and challenge for the soloist, and ideally pleasure for the listener as well.

Certain idiomatic chord sub-sequences, such as cadences, have long been used by improvisers, but Cork (1988, 2008) was one of the main proponents of providing informal *labels* for a significant variety of these sequences, which he called "LEGO bricks", after the well-known toy building block system. We will simply call them *bricks* here. Later Clark (2007) and Elliott (2009) analyzed a much larger set of standard songs and extended the set of bricks suggested by Cork. Elliott's work and analysis representation, which he called "road maps", further extended and refined the set of analyses. Our focus is on the computational aspects of using bricks in an improvisational companion.

The work of the aforementioned authors emphasized chordal aspects of bricks. The present paper redirects the focus toward melody, with the intent that bricks are also a useful organizational concept in learning to improvise. Aebersold (1967) took a similar approach, mentioning a few common progressions. Berg (1992) provided theoretical underpinnings, but did not use the brick terminology. He focused primarily on cadences, to which he applied the term "turnarounds", targeting chords diatonic to major or minor keys.

A cognitive discussion of creativity in jazz improvisation was provided by Johnson-Laird (2002), who observed "The cognitive problem for jazz musicians is to create a novel melody that fits the harmonic sequence and the metrical and rhythmic structure of the theme." He also cited Cork (1988) as recognizing the possibility of modulation between arbitrary keys in standard tunes, which Cork called "joins".

Most notable for computer performance, Biles (1994) used a genetic algorithm to generate jazz licks. Walker (1997) and Thom (2000) each researched and prototyped

their concept of an improvisational companion. Although neither system is currently accessible, both are compatible with our specific suggestion, which focuses on a particular basis for learning.

Bricks

This section provides a few examples of bricks that occur in standard tunes, such as found in the jazz music literature. The illustrations are taken from the graphical user interface of our improvisational companion. Each brick consists of three rows. The top row is the *inferred* key of the brick, the second row is the name of the *inferred* brick, and the third row is the input chord sequence in the brick. The inference algorithm is described in a separate paper (under review). Definitions of bricks are specified as text, in a grammar file accessible by the user.

The first example is the most common type of jazz cadence in a major key, classically described as a ii-V^7-I cadence. The brick name is *straight cadence*, to distinguish it from other types of cadence. We use m7 to note a minor seventh chord.

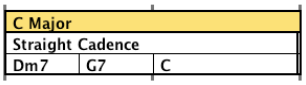


Figure 1: Straight cadence

The second example extends the first by adding two more chords to the front. Termed a *long cadence*, one of its functions is to prolong the tension leading up to the final resolution.

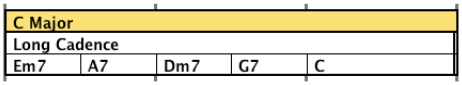


Figure 2: Long cadence

The third example extends the long cadence even further by adding two more chords. This is called a *starlight cadence*, in reference to the tune *Stella by Starlight*, by Victor Young, which contains a typical instance of this cadence.

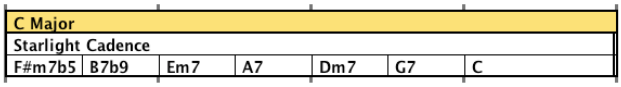


Figure 3: Starlight cadence

There are other bricks in addition to cadences. For example, a *turnaround* in our terminology is a brick that take the progression from a chord of a given function, (the tonic chord by default) toward a chord of another function (also the tonic by default). Figure 4 illustrates the *POT* (*Plain Old Turnaround*) brick (Aebersold, 1979), while figure 5 illustrates the *Ladybird Turnaround,* after the Tadd Dameron tune *Ladybird*, which entails making *tritone-substitutions* (*cf.* Berg, 1992) for the non-tonic chords in the POT.

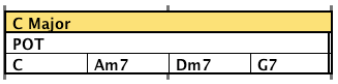


Figure 4: POT (Plain Old Turnaround)

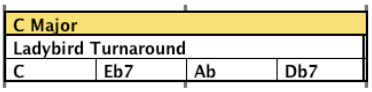


Figure 5: Ladybird Turnaround

Turnarounds do not always target the tonic. For example, the *dropback* (Cork, 2008) targets the ii chord. There are many variations of dropback, all ending with the secondary dominant for the ii chord (V^7 of ii = VI^7). Two of them are shown in Figures 6 and 7. *TTFA* stands for "turnaround to further away", with *further away* (from the tonic) being the phrase Cork used for the pre-dominant ii chord. *TINGLe* (Elliott, 2009) stands for *There is No Greater Love*, a tune by Isham Jones, that begins with this brick.

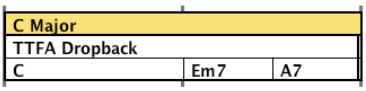


Figure 6: TTFA Dropback

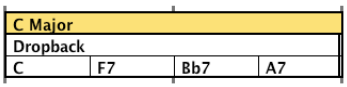


Figure 7: TINGLe Dropback

Figure 8 illustrates a roadmap produced by the companion, representing the analysis of a complete tune, in this case *Confirmation* by Charlie Parker, into bricks. The input that produced the roadmap consists of a text file with the chord symbols, bar markers, and section markers. In this case there are four sections, one per line. Most of the brick types in this tune have been defined above. Ones that haven't are *major on*, *sad approach*, and *straight launcher*. The word "on" simply means the tonic chord in the key of the moment. In this tune, FM7, which stands for F major seventh, is the *on* chord. An *approach* is the part of a cadence not including its resolution, and a *launcher* is an approach that resolves in the start of a new section.

One can also notice in Figure 8 the presence of *joins* Cork (2008), which the software shows as small tag boxes below some bricks. Joins represent transitions between bricks. For example there are six *sidewinder* joins in this tune. The sidewinder join is often used to signal a transition from a major key to its relative minor, for example F major to D minor at the start of the tune. Although important to understanding the tune as a whole, and practice should take into account all twelve join possibilities (one for each chromatic interval) over time, joins do not play a major role in the current exposition. Not every transition has an identifiable join, revealing a gap in Cork's method.

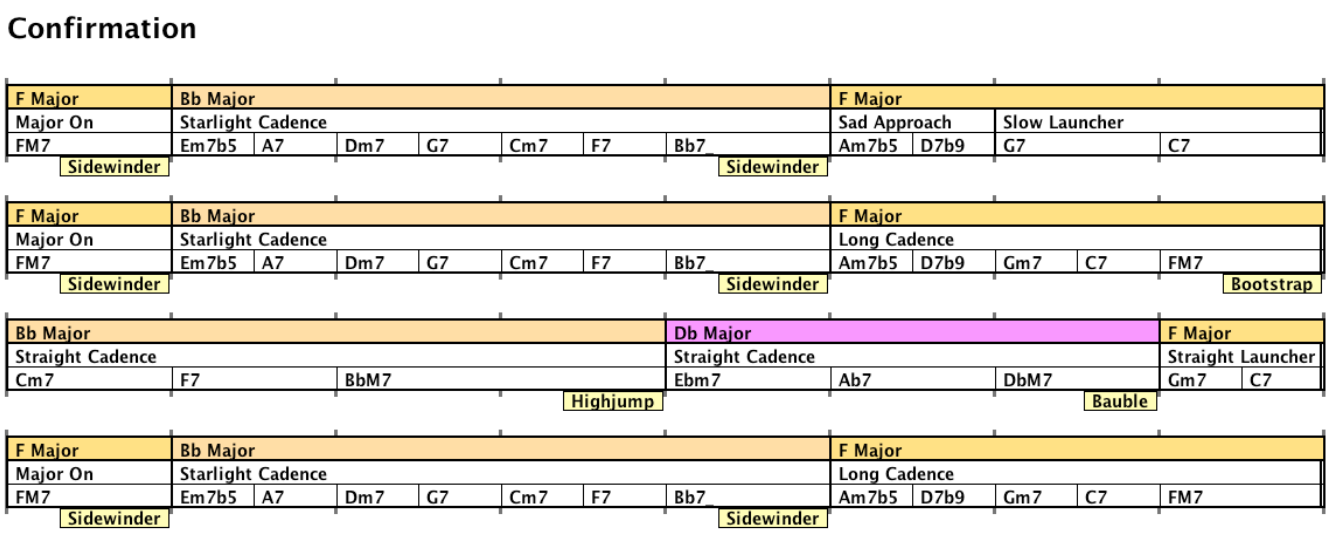


Figure 8: Screen capture of the algorithmically produced roadmap of a complete 32-measure tune, *Confirmation*

Automatic Brick Analysis

The feasibility of our proposal is enhanced by the development and implementation of an algorithm for analyzing the chord progression of a tune into bricks. The implementation underlies our user interface, starting with a lead sheet as input and resulting in a roadmap as output. Because the algorithm is described in another paper submitted for publication, we will take it for granted here.

One of the challenges of such an algorithm is that a given chord sequence can be interpreted as more than one brick sequence. The algorithm uses section sub-divisions of the tune to help reduce the ambiguity, and produces a final unique parse based on a cost assignment, representing user-specifiable precedence levels for various brick types.

It can also be noted in Figure 8 that some of the starlight cadences end in Bb7_. The underscore tells the program that this Bb7 chord functions as a *tension tonic* (Cork, 2008), rather than as a dominant, its function by default.

Interaction Modes

Once the brick roadmap for a tune is made available, an improviser can make use of the bricks for practice. Our user interface allows the repeated play ("looping") of bricks. There are various *modes* of looping:

- *Simple looping* mode plays only the automatically generated background. The user can play over it as long as desired.
- *Trading* mode has the companion play melody every other iteration. The intention is that the user will play when the companion is not playing the melody. The melodies can be generated by two possibilities:
  - Generated on the fly by a grammar.
  - Selection from a pre-composed database.
- *Recording* mode can be used with either of the above modes. Whatever the user plays is recorded.
- *Learning* mode augments recording by the program learning a grammar from the melodies played by the user.

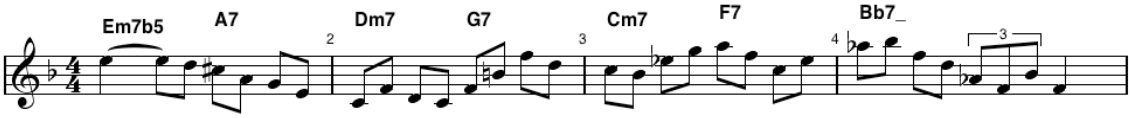


Figure 9: Screen capture of a melody played by the improvisation companion over a starlight brick

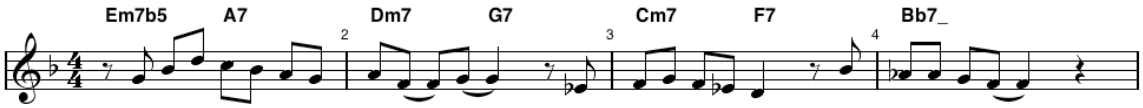


Figure 10: Edited screen capture of a response played as by a user over the starlight brick

The tradeoff between grammar and database creation of melodies is that grammars can be much more creative, generating a wider variety of material and do not require the laborious process of predetermining the database. However, use of database allows one to avoid sub-standard melodies.

At this writing, recording and learning have not been completely implemented, although a sufficient technology base exists to establish their feasibility. Recording from a MIDI instrument, such as a keyboard or EWI (Electronic Wind Instrument), is relatively straightforward and available. Recording from audio will require the addition of a module for audio to MIDI transcription. Software tools such as *Smart Music* (2012) and *Intelliscore* (2012) establish the feasibility of recording and learning from audio.

Figure 9 shows a screen capture of a companion-generated melody over the starlight cadence bricks in Figure 8, while Figure 10 shows a possible user response, played on a MIDI instrument input at 100 beats per minute. Figure 10 was edited for readability to account for *swing* feel in the user's playing. Automation of the reversal of swing feel for visualization purposes is a solvable problem still to be implemented in our system.

Our user interface provides additional features to enhance its applicability:

- The various play modes are not limited to just single bricks. Any contiguous combination of bricks can be played, allowing gradual range expansion.
- A variety of different styles, such as swing, bossa, rock, etc., can be generated as background for the bricks. Tempos can be varied to suit the player.
- Large bricks can often be broken down hierarchically into sub-bricks, as bricks in general are defined using a grammar-like notation in a user-modifiable dictionary. This feature can facilitate incremental learning by the user.
- The user can drag bricks together to create the harmony of a totally new composition, then save the result as a lead sheet.

Figure 11 shows a screen capture of the full roadmap interface as it currently stands, including the brick dictionary, from which the user can select bricks.

Using Bricks to Improve Grammar Learning

Gillick, et al. (2010) demonstrate a method for automating learning of grammars for generating melodies over a sequence of chords, by using a set of transcriptions of solos as input. The set of solos can be as small as a single solo, or a part of a solo. This learning method, as used in Impro-Visor (2012), involves scanning the solo using a fixed-length moving window one or two measures long. Each segment scanned in the moving window is converted, based on the underlying chords, into an *abstract melody* wherein the notes are not actual pitches, but rather categories (chord tones, color tones, approach tones, etc.). The melodic contours are represented by a series of *slopes*, groups of notes that uniformly rise or fall, with parametric bounds on the number of semitones between notes.

Once the segments have been extracted by the moving window method described above, they are clustered by similarity. Then a small set of *representatives* is chosen for each cluster, and the representatives are *chained* probabilistically using a Markov chain, which is conveniently representable as productions in the overall probabilistic context-free grammar. The transition probabilities are derived empirically from occurrence frequencies in the transcribed solos.

The grammar is used to generate melodies by instantiating the representative abstract melodies into actual melodies. Due to the manner of abstracting and chaining, the results exhibit stylistic characteristics of the original solos without being rote copies of them. Probabilistic selection of grammar productions is the root source of apparent creativity emerging from the algorithm.

Our new contribution is to use bricks as the windows, rather than having a fixed-size window as in the case of Gillick, et al. An advantage of using bricks is that they are already harmonically coherent units. As one can expect that a solo performed by a professional player will have melodic segments that conform to the harmonic units (Johnson-Laird, 2002), such a correspondence represents the transcribed soloist's understanding of the harmonic flow of the tune.

Another advantage of using bricks as windows is that the Markov chaining used for sequencing small fixed-length segments can be eliminated. Chaining still might find uses in connecting brick-based melodies themselves; however, we expect that this macro-level chaining will not be extraordinarily useful in jazz, where large-scale coherence tends to be the exception.

Learning from the User

A set of recorded melodies produced by a user can be a viable basis for grammar learning. Such learning can either take place off-line or, if the processing demands are not too great, while the companion's melody is being played. Learning from the user's own melodies as a basis for a grammar can provide positive or negative reinforcement, as the resulting grammar is fundamentally an embodiment of the kinds of melodies being played by the user.

Assessment

As our companion is proposed as a learning vehicle, it is worthwhile to consider what kinds of mechanisms might be possible to provide feedback for the user. Coloration of user-played notes to indicate chord tones, color tones, approach tones, and other, as employed by Impro-Visor, provides one means of judging how well the tones being played by the user conform to the underlying harmony. This contrasts with note coloration used by, e.g. Smart

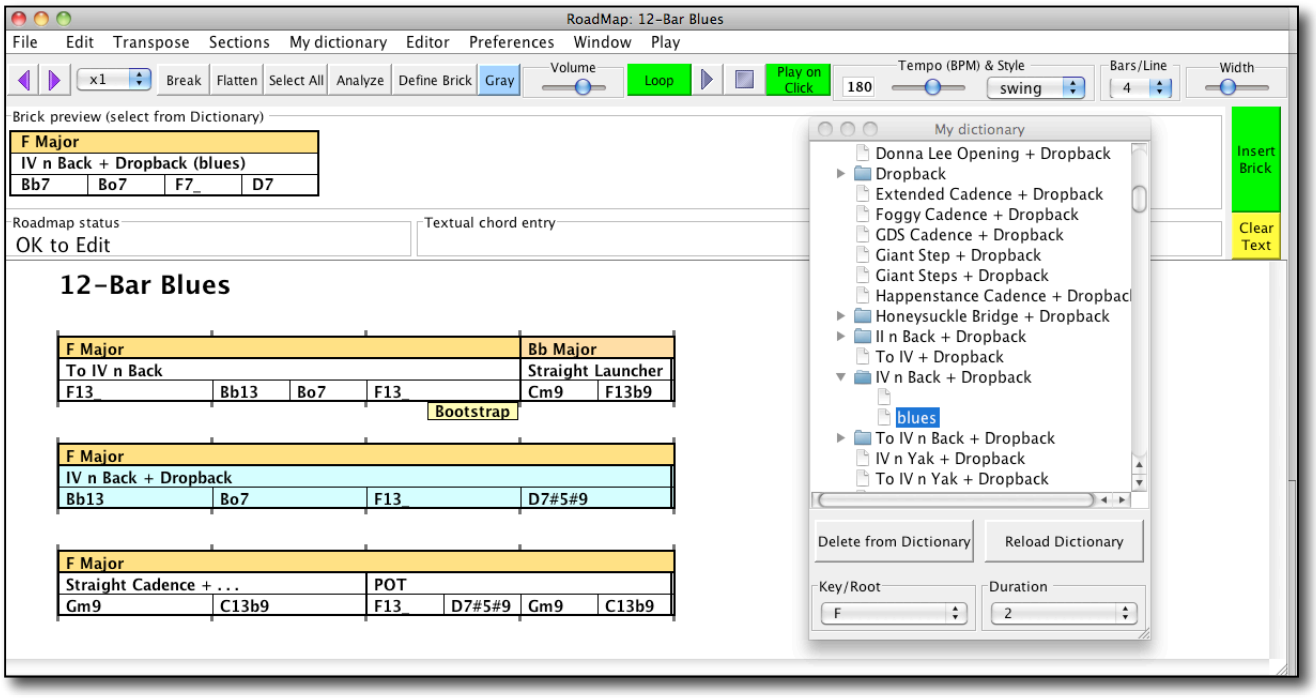


Figure 11: The roadmap user interface, showing the brick dictionary inset on the right

Music (2012), which informs the user whether a note is the one specified in the score. With improvisation, there is no single correct note, so Impro-Visor's note categorization is more appropriate in this context. Another capability that could be added concerns timing. Smart Music can inform the user whether a note is played early or late. But what is desired for jazz is to inform the user whether his or her timing *swings* or not, an aspect difficult to capture, and one that remains as a topic of future research.

Conclusion

We have proposed an approach toward a jazz improvisation companion based on the idea of harmonic bricks. The latter were suggested by Cork (1988, 2008) as a means of remembering jazz chord progressions. Our suggestion is to also use bricks as a basis for improvising melodies. Toward that end, we have prototyped a software improvisation companion based on this idea. Although the learning aspects of the tool is work in progress, the implementation has progressed far enough that we are confident that the approach will be useful in an educational setting to help enhance the creativity of its users.

References

Aebersold, J. 1967. http://www.jazzbooks.com/

Aebersold, J. 1979. *Turnarounds, Cycles, and II/V7's*, Jamey Aebersold publisher.

Berg, S. 1992. *Jazz Improvisation*: *The Goal Note Method*. Kendor Music, Inc.

Biles, J. 1994. Genjam: A genetic algorithm for generating jazz solos. In Proceedings of the 1997 ICMC, 1994

Clark, P. 2007. A book of LEGO. http://www.cs.hmc.edu/~keller/jazz/BookOfLegoPhilClark.pdf

Cork, C. 1988. *Harmony by LEGO Bricks*: *A New Approach to the Use of Harmony in Jazz Improvisation*. Leicester: Tadley Ewing Publications.
Cork, C. 2008. *The New Guide to Harmony with LEGO Bricks*. London: Tadley Ewing Publications. http://www.tadleyewing.co.uk.

Elliott, J. 2009. *Insights in Jazz*: *An Inside View of Jazz Standard Chord Progressions*. http://www.dropback.co.uk/

Gillick, J.; Tang, K.; Keller, R. 2010. Machine learning of jazz grammars. *Computer Music Journal*, September 2010.

Impro-Visor. 2012. http://www.impro-visor.com/

Intelliscore. 2012. http://www.intelliscore.net/

Johnson-Laird, P. 2002. How jazz musicians improvise. *Music Perception*, Spring 2002, Vol. 19, No. 3, 415–442.

Smart Music. 2012. http://www.smartmusic.com/

Thom, B. 2000. BoB: An interactive improvisational music companion. *Proceedings, Fourth International Conference on Autonomous Agents*. Barcelona, Spain. ACM.

Walker, W. 1997. A computer participant in musical improvisation. *CHI '97 Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM.

Automatic Composition from Non-musical Inspiration Sources

Robert Smith, Aaron Dennis and Dan Ventura
Computer Science Department
Brigham Young University
2robsmith@gmail.com, adennis@byu.edu, ventura@cs.byu.edu

Abstract

In this paper, we describe a system which creates novel musical compositions inspired by non-musical audio signals. The system processes input audio signals using onset detection and pitch estimation algorithms. Additional musical voices are added to the resulting melody by models of note relationships that are built using machine learning trained with different pieces of music. The system creates interesting compositions, suggesting merit for the idea of computational "inspiration".

Introduction

Musical composition is often inspired by other musical pieces. Sometimes, the new music closely resembles the inspiring piece, perhaps being an intentional interpretation or continuation of its themes or ideas. Other times the connection between the pieces is not identifiable (or even conscious). And, such sources of inspiration are, of course, not limited to only the musical realm. A composer can be inspired by the sight of a bird, the smell of industrial pollution, the taste of honey, the touch of rain or the sound of a running stream. Since this is the case, an interesting question for the field of computational creativity is whether a similar mechanism can be effected in computational systems. If so, new, interesting mechanisms for the development of (musical) structure become viable.

Many attempts have been made at computational composition. These attempts use mathematical models, knowledge based systems, grammars, evolutionary methods and hybrid systems to learn music theory, specifically whatever music theory is encoded in the training pieces applied to the algorithms (Papadopoulos and Wiggins 1999). Some of these techniques have been shown to be capable of producing music that is arguably inspired by different music genres or artists (Cope 1992). Some computational composers focus on producing melodies (Conklin and Witten 1995), but most focus on producing harmonies to accompany a given melody (Chuan and Chew 2007)(Allan and Williams 2005). Ames (Ames 1989) and others have described training Markov models on existing artists or styles and generating similarly sounding melody lines. No system that we have found models the idea of artistic inspiration from non-musical sources.

We present a computational system which implements a simple approach to musical inspiration and limit our focus to (non-musical) audio inspirational sources. Our system can autonomously produce a melody and harmonies from non-musical audio inputs with the resulting compositions being novel, often interesting and exhibiting some level of acceptable aesthetic.

Methodology

Our approach to automatic composition from non-musical inspirational sources is composed of four steps: (1) audio input and melody generation, (2) learning voice models, (3) harmony generation and (4) post-processing.

Audio Input and Melody Generation

Inspirational audio input was selected from various sources. Our samples included baby noises, bird chirpings, road noises, frog croakings, an excerpt from Franklin Delano Roosevelt's "A Date Which Will Live in Infamy" speech, and an excerpt from Barack Obama's 2004 DNC speech.

The melody generator takes an audio file (.wav format) as input and produces a melody. The input signal typically contains many frequencies playing simultaneously and continuously, and the generator's job is to produce a sequence of non-concurrent notes and rests that mimics the original audio signal. To do so, it uses an off-the-shelf, free audio utility called Aubio to detect the onset of "notes" in the audio file (as well as to estimate their duration) and to extract the dominant pitch at each of these onset times. Aubio is intended for analyzing recordings of musical pieces in which actual notes are played by instruments; however, in our system it is used to analyze any kind of audio signal, which means Aubio extracts "notes" from speeches or recordings of dogs barking or anything else. A thresholding step discards generated notes that are too soft, too high, or too low. The result is a collection of notes, extracted from the raw audio, composing a melody.

Learning Voice Models

To produce harmonization for the generated melody, we employ a series of voice models, M_i, learned from a collection of MIDI files representing different musical genres and artists. Each such model is trained with a different set of training examples, constructed as follows. First, because

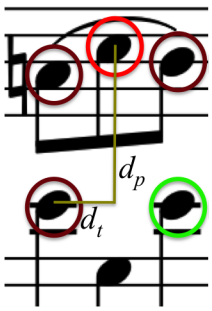


Figure 1: *Finding neighbor notes.* The top center note (circled in red) is the current melody note. In this case, $k = 3$, and, assuming $w_p = w_t$, the k closest neighbors are the two notes surrounding the melody note on the top staff and the first note on the bottom staff (circled in dark red). d_t refers to the distance in time between the melody note and neighbor, and d_p refers to the change in pitch. The $(k + 1)$th note is the rightmost note on the bottom staff (circled in green).

there is no restriction on the time signature of the input or output pieces, note durations are converted from number of beats to seconds.

Second, to identify the melody line of the training piece (and later to identify the melody line of the output piece), we use a simple heuristic assumption that the highest pitched note at any given time is the melody note.

Third, for each melody note, we find the $k + 1$ nearest neighbor notes using the distance function (see Figure 1):

$$d(n_1, n_2) = \sqrt{w_t d_t(n_1, n_2)^2 + w_p d_p(n_1, n_2)^2}$$

where n_1 and n_2 are notes, and weights w_t and w_p allow flexibility in how chordal or contrapuntal the training data will be. d_t and d_p compute absolute difference in onset time and pitch, respectively, so

$$d_t(n_1, n_2) = |\text{onset}(n_1) - \text{onset}(n_2)|$$

and

$$d_p(n_1, n_2) = |\text{pitch}(n_1) - \text{pitch}(n_2)|$$

Training instances are constructed from a musical piece's melody notes and its $k + 1$ closest notes. The training inputs are the melody note and its k nearest neighbors, while the $(k + 1)$th closest note is used as the training output (see Figure 1). The melody note is encoded as a 2-tuple consisting of the note's pitch and duration. The neighbor notes and the output note are encoded using a 3-tuple consisting of the time (d_t) and pitch (d_p) differences between the neighbor note and the melody note and its duration (see Figure 2). When building the training set for voice model M_i (with i indexed from 0), $k = i + 2$. So, after training, voice model M_i computes a function, $M_i : \mathbb{R}^{3i+8} \rightarrow \mathbb{R}^3$.

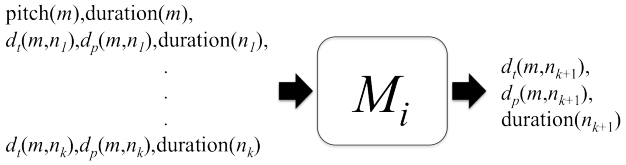


Figure 2: *Training the voice models.* For each melody note m of each training piece, a training instance is created from the melody note and the $k + 1$ closest neighboring notes $(n_1, ...n_{k+1})$. The k closest neighbors are used, along with m as input, and, as training output, the $(k + 1)$th closest neighbor is used. The melody note is represented as a pitch and a duration. Each of the other notes is represented as a 3-tuple consisting of d_t, d_p, and duration, where d_t and d_p refer respectively to the differences in start time and pitch between the neighbor note and the melody note.

Harmony Generation

The harmony generator is applied iteratively to add notes to the composition. Each pass adds an additional voice to the composition as follows. For the iteration 0, $k = 2$ and voice model M_0 is used with the melody as input. Each note, in turn, is used as the melody note, and it and it's two nearest neighbors are used as input to the model, which produces an output note to add to the harmonizing voice. This does not imply that each harmony note is produced to occur at the same time as its associated melody note. For each melody note the model produces as output values for d_t, d_p, and duration; the harmony note will only start at the same time as the associated melody note if $d_t = 0$.

When all melody notes have been used as input, the additional harmonic voice is then combined with the original melody line and the first iteration is complete. For iteration 1, $k = 3$ and voice model M_1 is used with the new two-voice composition as input, and the process is repeated, with the following caveat. We use the "melody" notes of the current piece (that is, the highest pitched notes) instead of the original melody notes (along with their k neighbors) as input to the model. This allows the melody notes to change from iteration to iteration, since the system can output notes that are higher than the (current) melody. The end result is another harmonic voice that is combined with the two-voice composition to produce a three-part musical composition (see Figure 3).

This process is repeated for v iterations, so that the final composition contains $v + 1$ voices in total. Empirically, we found that $v = 3$ resulted in the most pleasing outputs. With $v < 3$ there was not enough variation to distinguish the output from the original melody. For higher values of v, the less musical and more cluttered the output became.

Post-processing

After the output piece has been composed, the composition is post-processed in two ways which we call *snap-to-time* and *snap-to-pitch* (and to which we refer collectively as *snap-to-grid*).

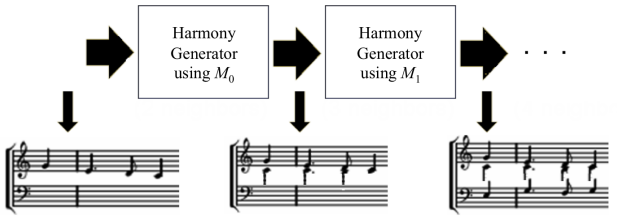


Figure 3: *Adding voices.* The harmony generator is applied iteratively over the melody line and generated harmony lines, using successively complex voice models. These iterations add successive voices to a composition.

Algorithm 1 Snap-To-Time. This algorithm adjusts note start times in the final composition to compensate for lack of uniform timing across input and training pieces. First, δ_{min}, the minimum difference in start time between any two notes in the melody, is calculated. Each note is then shifted so that its start time is an integer multiple of δ_{min} from the start time of the composition's initial note.

$\delta_{min} \leftarrow \infty$
for all notes n_1 **do**
  for all notes n_2 **do**
    $\delta \leftarrow |\text{onset}(n_1) - \text{onset}(n_2)|$
    if $\delta < \delta_{min}$ **then**
      $\delta_{min} \leftarrow \delta$
    end if
  end for
end for
for all notes n **do**
  $\Delta \leftarrow \lfloor \text{onset}(n)/\delta_{min} + .5 \rfloor * \delta_{min} - \text{onset}(n)$
  onset(n)←onset(n)+Δ
end for

Due to the beat-independent durations of the generated notes, the note onsets in the composition can occur at any time during the piece, which can result in unpleasant note timings. To correct this, we implement a snap-to-time feature.

To do so, we first analyze the melody line to determine the shortest time, δ_{min}, between any two (melody) note onset times. Then each composition note onset is shifted so that it is an integer multiple of δ_{min} from the onset of the first note in the composition (see Algorithm 1). In other words, each note is snapped to an imaginary time grid whose unit measure is δ_{min}, with the result being music with a more regular and rhythmic quality.

Because each voice is generated independently, there is no explicitly enforced (chordal) relationship between notes which occur at the same time. The voice models may provide some of this indirectly; however, this implicit relationship is not always strong enough to guarantee pleasing harmonies—there exists the possibility of discordant notes. To remedy this, we implement the snap-to-pitch algorithm.

If two notes occur at the same time, the difference in their pitches is computed. The pitches are then adjusted until the pitch interval between the notes is acceptable (here, for simplicity, acceptable means one of {*major third*, *perfect fourth*, *perfect fifth*, *major sixth*}). See Algorithm 2.

Algorithm 2 Snap-To-Pitch. The notes n_1 and n_2 start at the same time. If the interval between them is not one of {*major third*, *perfect fourth*, *perfect fifth*, *major sixth*}, snap-to-pitch modifies the pitch of one of n_2 so that it is.

$\delta \leftarrow \text{pitch}(n_1) - \text{pitch}(n_2)$
if $\delta > 0$ **then**
  if $\delta < 4$ **then**
    $\delta = 4$
  else
    while $\delta \notin \{4, 5, 7, 9\}$ **do**
      $\delta \leftarrow \delta - 1$
    end while
  end if
else if $\delta < 0$ **then**
  if $\delta > -3$ **then**
    $\delta = -3$
  else
    while $|\delta| \notin \{3, 5, 7, 8\}$ **do**
      $\delta \leftarrow \delta + 1$
    end while
  end if
end if
$\text{pitch}(n_2) \leftarrow \text{pitch}(n_1) - \delta$

As a summary, Algorithm 3 gives a high-level overview of the entire compositional process.

Results

Musical results are better heard than read. We invite the reader to browse some of the system's compositions at http://removedforblindcopy.

In some cases the melody generator produces melody out-

Algorithm 3 Algorithmic Overview Of System. A melody is generated by detecting pitch, onset, and duration of "notes" in an inspirational audio sample. Additional voices are added by creating increasingly complex voice models and iteratively applying them to the composition. The entire composition is then post-processed so that it incorporates a global time signature of sorts and to improve its tonal quality.

$composition \leftarrow \text{extractMelody}(inspirationAudio)$
for $i = 0$ to v **do**
  $k = i + 2$
  $trainset \leftarrow \emptyset$
  for all training pieces t **do**
    $trainset \leftarrow trainset \cup \text{extractInstances}(t, k))$
  end for
  $\text{trainModel}(M_i, trainset)$
  $composition \leftarrow \text{addVoice}(M_i, composition)$
end for
$composition \leftarrow \text{snapToTime}(composition)$
$composition \leftarrow \text{snapToPitch}(composition)$

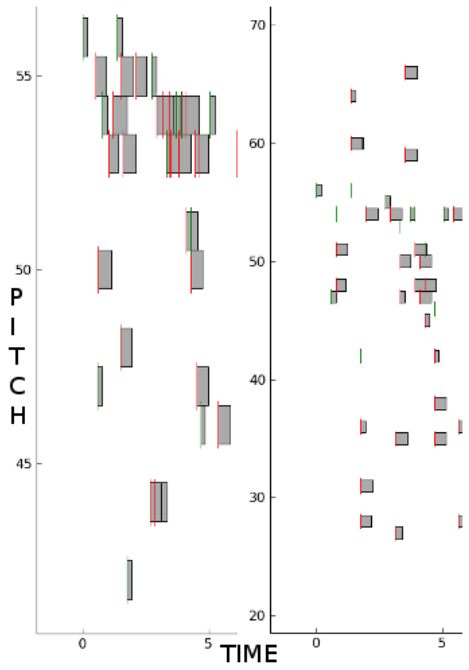


Figure 4: *Snap-to-grid*. The first graph shows the layout of an output composition based on CarSounds without snap-to-grid post-processing. The second graph shows another CarSounds output with snap-to-grid. Note the change in the pitch scale that reflects the increase in pitch range which is a result of adjusting concurrent notes to an aesthetically pleasing interval.

puts which are readily identifiable with their inspirational source audio files. Examples include compositions inspired by a speech by President Obama and by a bird's song. In both cases, the resulting melody line synchronises nicely with the original audio when both are played simultaneously. In contrast, other compositions sound very different from their inspirational source. Examples include a recording of a frog's repetitive croaking and a monotonous recording of road noise in a moving car. In the case of the road noises one would expect an output melody that is monotonous, mirroring the humanly-perceived characteristics of the input audio file. However, the melody generator composes a low-pitched, interesting, and varied melody line when given the road noise audio file, making it hard to identify how the melody relates to its source.

In all outputs there is a general lack of traditional rhythm and pitch patterns. This is, of course, not surprising given that our audio sources for inspiration are not required to be in any particular musical key or to follow traditional key changes, nor do they have any notion of a time signature. Additionally, we do not restrict our training sets in either of these traditional ways. As a consequence, it is likely that in any given training set there will be instances which are in different keys and/or time signatures than the melody. In light of these conditions, it is to be expected that the output would not be traditional music.

Training	w_t	w_p	Percent Chords
TwoDance	1	1	83
TwoDance	1	3	44
TwoDance	3	1	80
TwoBlues	1	1	67
TwoBlues	1	3	47
TwoBlues	3	1	71

Table 1: This table shows the effect of the weights w_t and w_p. The input was the FatFrog audio file and voice models were trained using either two songs from the Dance genre or two songs from the Blues genre. Generally, as w_p increases (with respect to w_t), the number of chords produced in the output composition decreases.

The snap-to-grid feature is helpful. We have posted audio examples on the web comparing outputs with and without snap-to-grid. An example graph of each is given for visual comparison in Figure 4. Snap-to-time doesn't significantly change the landscape of the pieces, but it proves to be essential in synchronizing notes which were composed as chords but are not easily recognized as such because of the high precision of start times. Snap-to-pitch has a dramatic effect on the pitch of certain notes but is limited to those notes which occur at the same time.

We explored several values for w_t and w_p (see Table 1), and, as expected, when $w_p > w_t$ there are less chordal note events than single notes compared to when $w_p < w_t$. Interestingly, the baseline $w_t = w_p = 1$ for the case of voice models trained with two Dance songs is slightly more chordal even than $w_t = 3, w_p = 1$.

We could not detect any significant difference in effect when using different genres or artists for training the voice models. No distinguishable qualities of dance music were discernible in the outputs composed using models trained only on dance music. No distinguishable qualities of Styx songs were discernible in the outputs composed using models trained only on songs by Styx. In short, each variable on training input successfully introduced novel variations in the output compositions in an untraceable way. Choice of training pieces did not produce a predictable pattern for aesthetic quality. The fact that our (admittedly simple) voice models failed to capture the distinct qualities of certain artists or genres suggests that our methods for encoding the musical qualities of training pieces are less effective at capturing such information than they are at capturing interesting note combinations and timings (see Figure 5).

As described, the standard system uses the $k + 1$ closest neighboring notes of each melody note for training the voice models, and this works. However, as a variation on this approach, randomly sampling $k + 1$ notes from the $4k$ closest notes adds some extra variation in the composition and can lead to more aesthetically pleasing outputs.

Snap-to-grid proved to be very useful for contributing to the aesthetic quality of the compositions. Compositions without snap-to-grid have more atonal and discordant chords which play at undesirable intervals. Using snap-to-grid allows a compromise between the uniqueness of the compo-

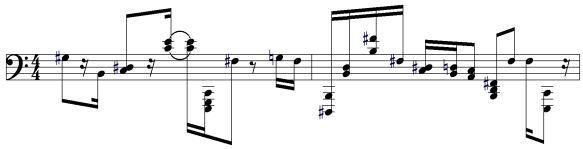


Figure 5: *Composition sample*. These two measures are taken from one of the compositions produced by our system. The system produces interesting rhythms with varying chordal texture.

sitional style and regular timing intervals and chordal structure.

Future Work

At this point, our system is quite simple and many of the techniques it employs are somewhat naïve musically. Some of this naïveté is for convenience at this early stage of system development, and some of it is design decisions that allow for greater variety in system output. The snap-to-grid processing is a post-hoc attempt to impose some level of musical "correctness" on the system's output. Given the unconstrained nature of the inspirational input, it is an interesting question to ask how one might naturally incorporate useful aspects of music theory directly in the melody generation process while still allowing significant effect from the source. Also, it is natural to suggest incorporating more traditional and mature harmonization schemes for the generated melodies. Finally, to this point, only the melody has been (directly) affected by the inspiring piece; it would be interesting to develop methods for using the inspirational source to directly influence other musical characteristics such as harmonization, style, texture, etc. However, all of these necessary improvements are relatively minor compared to the real open issues.

The first of these is the development of an evaluation method for judging aesthetic and other qualities of the compositions. To this point, our measure of "interestingness" has been only our own subjective judgment. The development of more principled, objective metrics would be useful as a filtering mechanism, and, at a more fundamental level, as feedback for directing the system to modify its behavior so that it produces better (novel, interesting, and surprising) compositions. In addition, such results may also be vetted in various kinds of human subject studies.

The second of these is the development of a mechanism for autonomously choosing which inspirational sources the system will use as input. This requires the development of some type of "metric" for inspiration. Or, perhaps another way to think about this problem is to ask the question, "what makes a sequence of sounds interesting (or pleasing, or arousing, or calming, or ...)?" Is this quantifiable or at least qualifiable in some way? Some potential starting points for this type of investigation might include work on identifying emotional content in music (Li and Ogihara 2003; Han et al. 2009) as well as work on spectral composition methods (Esling and Agon 2010).

This, in turn, introduces further considerations, such as in which quality or qualities the system might be interested and how those interests might change over time. An additional consideration is that of a second level of inspiration – rather than the system being inspired by the aural qualities of the input alone (as it is at present), is it possible to construct a system that can be inspired by metaphors those aural qualities suggest? And is it then possible for the system to communicate the metaphor to some degree in its output?

References

Allan, M., and Williams, C. K. 2005. Harmonising chorales by probabilistic inference. In *Advances in Neural Information Processing Systems 17*, 25–32.

Ames, C. 1989. The Markov process as a compositional model: A survey and tutorial. *Leonardo* 22(2):175–187.

Chuan, C. H., and Chew, E. 2007. A hybrid system for automatic generation of style-specific accompaniment. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*.

Conklin, D., and Witten, I. H. 1995. Multiple viewpoint systems for music prediction. *Journal of New Music Research* 24:51–73.

Cope, D. 1992. Computer modeling of musical intelligence in EMI. *Computer Music Journal* 16(2):69–83.

Esling, P., and Agon, C. 2010. Composition of sound mixtures with spectral maquettes. In *Proceedings of the International Computer Music Conference*, 550–553.

Han, B.; Rho, S.; Dannenberg, R. B.; and Hwang, E. 2009. SMERS: Music emotion recognition using support vector regression. In *Proceedings of the 10th International Conference on Music Information Retrieval*, 651–656.

Li, T., and Ogihara, M. 2003. Detecting emotion in music. In *Proceedings of the 4th International Conference on Music Information Retrieval*, 239–240.

Papadopoulos, G., and Wiggins, G. 1999. AI methods for algorithmic composition: A survey, a critical view and future prospects. In *Proceedings of the AISB Symposium on Musical Creativity*, 110–117.

Creativity in Configuring Affective Agents for Interactive Storytelling

Stefan Rank[1], Steve Hoffmann[2], Hans-Georg Struck[3], Ulrike Spierling[2], Paolo Petta[1]

[1] Austrian Research Institute for Artificial Intelligence (OFAI), Austria

stefan.rank/paolo.petta @ ofai.at

[2] Hochschule RheinMain, University of Applied Sciences, DCSM, Germany

Steve.Hoffmann/Ulrike.Spierling @ hs-rm.de

[3] Independent Screenwriter

georgstruck @ foni.net

Abstract

Affective agent architectures can be used as control components in Interactive Storytelling systems for artificial autonomous characters. Creative authoring for such systems then involves configuration of these agents that translate part of the creative process to the system's runtime, necessarily constrained by the capabilities of the specific implementation. Using a framework for presenting configuration options based on literature review; a questionnaire evaluation of authors' preferences for character creation; and a case study of an author's conceptualisation of the creative process, we categorise available and potential methods for configuring affective agents in existing systems regarding creative exploration. Finally, we present work-in-progress on exemplifying the different options in the ActAffAct system.

Introduction

Interactive Digital Storytelling (IDS) is concerned with the creation of a new media art form that allows for real-time interaction with a developing narrative. In the terminology of (Boden and Edmonds 2009), the aim is a form of CI-art or VR-art, i.e., computer-generated and responsive to audience interaction, possibly in the form of virtual reality. Creating new methods for adaptivity, generativity, and interactivity is seen as the prime method for advancing beyond traditional linear media, and while there are recent examples of technical approaches that are close to video-based media, e.g. video recombination (Porteous et al. 2010), and of conceptual approaches to story generation, e.g. based on analogy-mapping (Ontañón and Zhu 2011), a large part of the research has focused on enabling interactive storyworlds inhabited by synthetic characters (Rank 2005; Si, Marsella, and Pynadath 2005; Louchart and Aylett 2007). The assembly of autonomous conversational actors endowed with some degree of autonomy poses significant integration challenges (Gratch et al. 2002), including the development of authoring methodologies that support the creative process inside the boundaries of an IDS system.

We take the point of view of authors with IDS experience to find ways beyond the current disparity between needs of authors and capabilities and interfaces of existing systems. In order to examine the support for creative authoring in these systems, we look at methods for configuring one crucial element for translating parts of the creative process to runtime: affective agents. After introducing affective agent architectures and a framework for presenting their configuration options to authors that draws on literature review; the evaluation of authors' preferences for character creation using a questionnaire; and a case study of an author's conceptualization of the creative process, we examine available and potential methods for configuring affective characters in existing systems. Finally, we report on work-in-progress translating these options to the ActAffAct system.

Affective Characters

Affective agents are a specialization of intelligent agents for domains in which emotional and related phenomena are important. A key dimension for agent control architectures of synthetic characters is affective competence: believable portrayal of emotional reactions and the capability of selecting appropriate expressiveness; variability within the consistent boundaries perceived as personality (Ortony 2003); and the recognition of subjective relevance in the agent's (social) environment (Marsella, Gratch, and Petta 2010). The origins of such architectures often lie with scenarios of use (Rank and Petta 2006) that target other application areas, cognitive modeling or modeling of psychological theories. Their configuration is not necessarily suited directly for the authoring of synthetic characters.

In the context of IDS, synthetic characters translate parts of the creative decisions of authoring to the runtime system. Affective competence helps to ensure the emotional aspect of character portrayal as well as of the causal connections in a story, down to the fine-grained level of audience interaction. Here, we focus on techniques that model characters and their behaviour *explicitly* in order to achieve levels of motivational and behavioral autonomy that facilitate the generativity and interactive flexibility that IDS strives for. On the spectrum from 'strong autonomy' to 'strong story' (Mateas and Stern 2000; Swartjes 2010), this places the approach towards the autonomy-end, more compatible with the idea of emergent narrative that nevertheless requires purposeful authoring (Louchart et al. 2008), in addition to affective and situated competencies, to be successful. Even in a strongly story-based interactive system, autonomously competent agents are valuable if they can be configured to act 'in character' during episodes that are not directly controlled by a story-based framework or if the system explicitly

represents emotional links between characters as part of the authoring process (Pérez y Pérez 2007). For the context of this work, the top-level of an IDS system, drama management (Roberts and Isbell 2008), is not considered: we intentionally focus on single characters and their autonomous behaviour: *character goals* rather than *author goals* (Riedl 2009).

The nature of interactive systems entails that judgment of creativity cannot focus on novelty and value of static artifacts as for systems that clearly separate a generative part, see also (Gervás 2009): The use of affective agents carries aspects that are clearly separable for static artifacts, such as chronology, causality, and the distinction between fabula and discourse, i.e. what is told and how it is told, into the runtime of a system. Rather, we try to map the conceptual spaces that are established by affect models as well as the ways for exploring and potentially transforming them. This approach is similar to work in the area of game design (Smith and Mateas 2011). In (Spierling and Hoffmann 2010), the authors state that creative authorship is far from obsolete in the context of IDS. The creative output consists to a substantial degree of the specific configuration of the control system. Abstractions at the creative conceptual level are seen to be distinct from the more formal abstractions (Crawford 2004) required for implementation. Rather, this relationship between authorial conceptual abstractions and implementation-specific abstractions of a more technical kind and the special kind(s) of dedicated support required of an IDS system for this transformation needs to be *investigated* for each case.

Configuration Options for Affective Agents

We use a framework for the configuration of affective characters as seen from the author's perspective (Rank et al. 2012), based on feedback received from a questionnaire study performed during an authoring workshop[1], as well as a case study of one author's practice, strongly rooted in drama theory. The questionnaire used free-text feedback and a set of Likert-scale questions intended to gauge the relative importance of different strategies for creating characters for a story-world. The free-text feedback reflected a wide range of approaches to character creation, including placing the focus on events, conflicts, or personality-specific feelings that happen to the character; the reliance on known characters or on personal experience as a starting point; and picking the background and underlying goals of characters as central element. At the same time, the reported results point at the complementarity of different approaches to character creation while the evaluation of preferences for different approaches showed no significant differences at that level of investigation (Rank et al. 2012).

As a second source for authors' viewpoints on the problem of configuring synthetic characters in an effective way, we draw on a case study of the authoring experience of one of the co-authors (Struck 2005) with both traditional linear narratives (i.e., script writing) and interactive storytelling systems. The central tenet of this author's experience is the conceptualization of narrative as a *sequence of emotional effects*. Underlying this approach are character-centered drama models based on work such as by Frank Daniel (see (Howard and Mabley 1995)) and the notion of a *conceptual space* (Struck 2005). Furthermore, the practice corresponds to the cyclic process of *engagement* and *reflection* that has been proposed as a model of creativity in writing (Sharples 1999).

As a quick way to concretize a character's role in a narrative, it is proposed to answer the question: What does the character want the most? Motivations and aversions of a character are then considered subordinate to this main desire. The notions of aspirations, vocations, and goals are crucial to derive a character's *fears*. As an example, consider a character that wants to see the world: Potential candidates for suitable high-impact fears would then be fear of flying or fear of crossing open water. A narrative is then seen to pitch a character's goals against obstacles involving the need to risk high stakes. Anything lacking in connection to a character's hopes and fears is omitted. Conversely, everything that is shown relates to this backstory of the character. Note that such constraints quickly extend beyond individual characters to comprise their social stances and interrelationships (Spierling and Hoffmann 2010). In addition, this selection principle contributes to perceived believability by allowing for inference and prediction of motivations and intentions through observations, e.g. (Riedl and Young 2010), p.220.

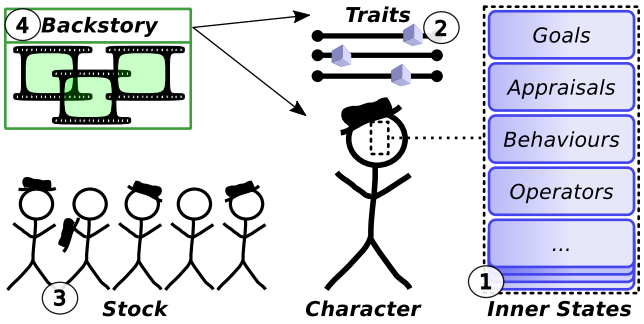


Figure 1: The levels of presenting configuration of characters in IDS.

Based on this investigation of the authors' viewpoint, and considering available systems, four different *non-distinct* levels of presenting configuration are distinguished, as illustrated in Figure 1:

1. Direct changes to initial *inner states* of agents as motivating factors for the character's behaviour.
2. Parameter settings that influence the inner working of an agent in correspondence to a theoretically persistent characteristic of the agent as a whole: *traits*.
3. Complete *stock characters* with a particular personality that can be used as a basis for customization.
4. Configuration based on the selection of *backstory experiences* that influence background beliefs and emotional parameters.

[1]The IRIS authoring summer school http://iris.interactive-storytelling.de/Summerschool

Levels 1 and 2 define a conceptual space that can be explored exhaustively in theory, though not in practice for most systems. Levels 3 and 4, while relying on the same methods of character control, present configuration as a transformational process starting from exemplars[2]. As mentioned, these levels are not distinct and can be seen as complementary, progressing towards a higher level of abstraction.

Configuration in Existing Systems

In the following, we review current systems and their support for exploring configuration in terms of inner states and traits, as well as the potential for supporting stock characters and backstory experiences. An important biasing factor for the selection of example systems was the open or confidential availability of source code.

As mentioned above, the inner states of existing systems are often strongly tied to the origin of the agent architecture and not geared towards authoring. Many affective agent architectures can be seen as extensions of *belief-desire-intention (BDI) agent models* based on ideas about resource-bounded practical reasoning (Bratman, Israel, and Pollack 1988). These 'BDI+E' architectures rely on: *beliefs* that represent what an agent holds to be true, *desires* (or goals) that an agent tries to fulfill, and a representation of what the agent is capable of doing, often in terms of a *plan library*. The third name-giving element, *intention*, refers to capabilities activated in pursuit of a current goal. The main influence on agent behaviour that all these architectures share and that are directly amenable to exploration are the relative importance of different desires, i.e. their *utility*, and the set of capabilities available to a specific agent.

Examples of architectures that add an operationalization of *affective* appraisal are FAtiMA (Dias and Paiva 2005) or ActAffAct (Rank 2005). The addition of affective appraisal results in further parameters for exploring configurations: the relative importance of standards, i.e. the *evaluation of different types of behaviour*; the initial relative importance of the actors and objects in the storyworld; and the *creation thresholds* and *decay rates* for different types of emotions. Further, such architectures were extended to consider *mood as a meta-level effect*, i.e. as an aggregate of previous emotions. An important additional parameter in this respect is the number of emotions considered. One extension of FAtiMA (Doce et al. 2010) applies the so-called OCEAN or five-factor model of personality. An individual's personality is expressed as values of five *personality traits* that can be used to explore the possible resulting behaviours: openness, conscientiousness, extraversion, agreeableness, neuroticism. The values for the five OCEAN factors influence the appraisal process in terms of thresholds and decay rates for emotion instances, but also coping and planning as well as expressivity in an animation system.

Both the BDI framework and personality theory frame the configuration of agents in terms that are still close to the authorial conceptualization presented above. For other mechanisms, the matching of explorable settings is less direct. In implementations based on PSI theory, such as MicroPSI (Bach 2003) or ORIENT (Lim et al. 2012), another extension of FAtiMA, emotions are described as sets of modulators that influence processing directly: *arousal*, i.e. the propensity for action, *resolution level*, i.e. the accuracy of internal processing, and *selection threshold*, i.e. resistance to change the current intentions. Exploring the effect of different situations on the values for these modulators opens a space for an agent's personality. On a more general level, PSI theory introduces the individual settings of so-called *motivators* (affiliation, integrity, energy, certainty, and competence), homeostatic variables with an influence on behaviour based on the deviation from set-points.

Further affective architectures are built on top of cognitive architectures that in themselves provide a wide range of possibilities for configuring individual differences in terms of *inner states*. An example is Soar and the emotion models based on it, such as EMA (short for emotion and adaptation) (Marsella and Gratch 2009) or PEACTIDM (Marinier and Laird 2006). Soar itself provides a general processing cycle that can be used in different styles, which in turn results in a wide range of configuration options. Corresponding to the BDI approach, the *utility of goals* and the *availability of operators* form the core of any configuration. Similarly, in Thespian (Si, Marsella, and Pynadath 2005; Si, Marsella, and Pynadath 2009), goals, policies and beliefs about self and others are the determining factors of single agent behaviour. In addition, in support of authorial control, characters can be configured by specifying multiple story-paths that are used to deduce their goals. Thereby, this approach employs a strong-story element to parameterize a system that is originally autonomy-driven. In EMA, the overall affective assessment is based on a causal interpretation of the current state of the world. On a conceptual level, the *granularity of this representation* forms an important part of configuring the personality of an agent. One focus of affective architectures in general are coping activities defined as the inverse operation of appraisal. Coping thus involves the identification and influencing of the believed causes for the currently significant state. Different coping strategies can be available to a single agent and the selection of these strategies represents a new level of potential configuration and a suitable candidate for a high-level exploration of the conceptual space of affective agents. Further, *availability and relative priority of different coping strategies* can be linked to backstory experiences.

In the VirtualStoryteller framework (Swartjes, Kruizinga, and Theune 2008), *late commitment* is used for the autonomous characters to determine the values of internal parameters and the state of the storyworld at runtime rather than beforehand at authoring time. To inform these delayed decisions, an assessment of the benefit of available options for story development is computed, thus reducing options of direct authoring of character *states* in favour of more global *traits* of the character in the story context.

Planning and scheduling techniques add further implementation-specific parameters for configuring individual differences that are comparatively far removed

[2] See (Rook and Knippenberg 2011) on the influence of quality of exemplars on creativity and imitation depending on the regulatory focus of authors.

from the author's perspective. As an example, planning algorithms can use *quality measures*, e.g., time and cost, and *resource constraints* to decide between alternative paths of action. Configuration of these mechanisms involves relative weighting of different quality measures.

Table 1: Configuration options for inner states and traits.

Inner States	Beliefs, granularity of internal representation; Availability of capabilities; Utility of goals; Standards, evaluation of types of behaviour; Thresholds and decay rates for emotion types; Availability and priority of coping strategies
Traits	History considered for meta-level mood; Qualities/constraints for planning/scheduling; Openness, conscientiousness, extraversion, agreeableness, neuroticism; Arousal, resolution level, selection threshold; Importance of affiliation, integrity, energy, certainty, competence

Table 1 summarizes relevant options for configuration of affective behaviour, distinguishing options related to inner states from traits of the agent as a whole. Practical and intentional examples for stock characters and configuration based on backstory experiences are rare. However, most architectures were designed for a specific purpose and therefore a set of characters is available, at least in principle. To the best of our knowledge, explicit use of backstory experiences to influence individual character behaviour has not been implemented in any system so far.

Extending Configuration of ActAffAct

ActAffAct (Rank 2005) is a proof-of-concept system that relies solely on the affective competences of individual characters and their configuration in terms of *beliefs* and *desires* to generate interesting but very simple plots within an interactive storyworld setting comprising a hero; a villain; a victim; and a mentor, as well as simple props such as a sword; a rope; or a bouquet of flowers. As a system with BDI-background and a practical reasoning system at its core, the modification of *inner states* is the direct way of exploring character designs. The use of a mood-system allows for the modification of *traits*: influence factors of different emotion types on mood and the decay rate of the mood state, that influence the character as a whole. Our work-in-progress considers the support for stock characters and backstory experiences. Due to the use of "cliché" story characters, archetypes, as agents in the original system, a set of stock characters can be derived directly. Most interesting though is the realisation of backstory experiences: The implementation of the appraisal process includes the selection of coping styles and the relative weighting of different types of emotions. These two elements lend themselves to be configured by a selection from automatically generated episodes: every episode shows reactions of a character in the possible combinations of encounters with other characters and objects in the storyworld. Evaluation of the resulting authoring possibilities is planned, including a questionnaire study.

Conclusion

In this paper, we used a conceptualization of character creation from the author's viewpoint to review different levels of configuration that current affective agent architectures provide. A mapping of notions of the author's creative process to configuration options is not straightforward. Rather, due to the roots of many affective agent architectures in areas other than IDS, parameters offered for exploration are often far removed from the author's perspective on character creation. On the other hand, parameters that stem from theoretical and practical considerations in agent architectures potentially provide new sources of creative inspiration for authors, both in terms of details of modeling and in terms of additional factors of influence that implementation-specific configurations expose for exploration. Overall, the conceptualization helps to frame the support for creativity in authoring IDS systems, and points to future extensions of approaches for character configuration. Finally, based on the review of configuration options in related systems, we presented ongoing work in extending the ActAffAct system regarding new ways for configuring autonomous characters.

Acknowledgements

This work is partially supported by the European Commission under grant agreement IRIS (FP7-ICT-231824). The Austrian Research Inst. for AI is supported by the Austrian Federal Ministry for Transport, Innovation, and Technology.

References

[Bach 2003] Bach, J. 2003. The MicroPsi Agent Architecture. In *Proc.5.Int.Conf.on Cognitive Modeling ICCM-5, Bamberg DE, Apr.10-12 2003*, 15–20. Universitäts-Verlag, Bamberg DE.

[Boden and Edmonds 2009] Boden, M. A., and Edmonds, E. A. 2009. What is generative art? *Digital Creativity* 20:1–2.

[Bratman, Israel, and Pollack 1988] Bratman, M. E.; Israel, D. J.; and Pollack, M. E. 1988. Plans and Resource Bounded Practical Reasoning. *Computational Intelligence Journal* 4(4):349–355.

[Crawford 2004] Crawford, C. 2004. *Chris Crawford On Interactive Storytelling*. New Riders Publishing, Indianapolis.

[Dias and Paiva 2005] Dias, J., and Paiva, A. 2005. Feeling and Reasoning: a Computational Model. In *Progress in AI, EPIA 2005*, 127–140. Springer LNCS 3808.

[Doce et al. 2010] Doce, T.; Dias, J.; Prada, R.; and Paiva, A. 2010. Creating Individual Agents Through Personality Traits. In *Intelligent Virtual Agents, 10th Int.Conf. IVA 2010, Philadelphia PA USA, Sep.20-22 2010. Proc.*, 257–264. Springer.

[Gervás 2009] Gervás, P. 2009. Computational Approaches to Storytelling and Creativity. *AI Magazine* 30(3):49–62.

[Gratch et al. 2002] Gratch, J.; Rickel, J.; André, E.; Cassell, J.; Petajan, E.; and Badler, N. 2002. Creating Interactive Virtual Humans: Some Assembly Required. *IEEE Intelligent Systems* 17(4):54–63.

[Howard and Mabley 1995] Howard, D., and Mabley, E. 1995. *The Tools of Screenwriting - A Writer's Guide to the Craft and Elements of a Screenplay*. St. Martin's Griffin.

[Lim et al. 2012] Lim, M. Y.; Dias, J.; Aylett, R.; and Paiva, A. 2012. Creating adaptive affective autonomous NPCs. *Autonomous Agents and Multi-Agent Systems* 24(2):287–311.

[Louchart and Aylett 2007] Louchart, S., and Aylett, R. 2007. From Synthetic Characters to Virtual Actors. In *Proc.3.Artificial Intelligence and Interactive Digital Entertainment Conf.*, 88–90. AAAI Press.

[Louchart et al. 2008] Louchart, S.; Swartjes, I.; Kriegel, M.; and Aylett, R. 2008. Purposeful Authoring for Emergent Narrative. In *Proc.1.Joint Int.Conf.on Interactive Digital Storytelling (ICIDS 2008)*, 273–284. Springer LNCS 5334.

[Marinier and Laird 2006] Marinier, III, R. P., and Laird, J. E. 2006. A Cognitive Architecture Theory of Comprehension and Appraisal. In *Proc.18.Meeting on Cybernetics and Systems Research, Apr.18-21 2006, Univ. of Vienna*. Austrian Soc.for Cybernetic Studies. 589–594.

[Marsella and Gratch 2009] Marsella, S. C., and Gratch, J. 2009. EMA: A process model of appraisal dynamics. *Cognitive Systems Research* 10(1):70–90.

[Marsella, Gratch, and Petta 2010] Marsella, S.; Gratch, J.; and Petta, P. 2010. Computational Models of Emotion. In *A Blueprint for Affective Computing - A sourcebook and manual*. Oxford Univ. Press. 21–46.

[Mateas and Stern 2000] Mateas, M., and Stern, A. 2000. Towards Integrating Plot and Character for Interactive Drama. In *Working Notes of the Social Intelligent Agents: The Human in the Loop Symposium. AAAI Fall Symposium Series. Menlo Park CA*. AAAI Press. 113–118.

[Ontañón and Zhu 2011] Ontañón, S., and Zhu, J. 2011. The SAM Algorithm for Analogy-Based Story Generation. In *Proc.7.AAAI Conf.on AI and Interactive Digital Entertainment, AIIDE 2011, Oct.10-14 2011*, 67–72. AAAI Press.

[Ortony 2003] Ortony, A. 2003. On Making Believable Emotional Agents Believable. In *Emotions In Humans And Artifacts*. MIT Press. 189–212.

[Porteous et al. 2010] Porteous, J.; Benini, S.; Canini, L.; Charles, F.; Cavazza, M.; and Leonardi, R. 2010. Interactive Storytelling Via Video Content Recombination. In *Proc.18.Int.Conf.on Multimedia 2010, Firenze Italy, Oct.25-29 2010*, 1715–1718. ACM.

[Pérez y Pérez 2007] Pérez y Pérez, R. 2007. Employing Emotions to Drive Plot Generation in a Computer Based Storyteller. *Cognitive Systems Research* 8(2):89–109.

[Rank and Petta 2006] Rank, S., and Petta, P. 2006. Comparability Is Key to Assess Affective Architectures. In *Proc.18.Meeting on Cybernetics and Systems Research, Apr.18-21 2006, Univ. of Vienna*. Austrian Soc.for Cybernetic Studies. 643–648.

[Rank et al. 2012] Rank, S.; Hoffmann, S.; Struck, H. G.; Spierling, U.; Mayr, S.; and Petta, P. 2012. Authoring vs. Configuring Affective Agents for Interactive Storytelling. *Applied AI, to appear*.

[Rank 2005] Rank, S. 2005. Towards Reusable Roleplayers Using an Appraisal-Based Architecture. *Applied AI* 19(3-4):313–340.

[Riedl and Young 2010] Riedl, M. O., and Young, R. M. 2010. Narrative Planning: Balancing Plot and Character. *J.of AI Research* 39:217–268.

[Riedl 2009] Riedl, M. O. 2009. Incorporating Authorial Intent Into Generative Narrative Systems. In *Intelligent Narrative Technologies II, Papers from the 2009 AAAI Spring Symposium*, 91–94. AAAI Press.

[Roberts and Isbell 2008] Roberts, D. L., and Isbell, C. L. 2008. A Survey and Qualitative Analysis of Recent Advances in Drama Management. *Int.Trans.On Systems Science And Applications* 4(2):61–75.

[Rook and Knippenberg 2011] Rook, L., and Knippenberg, D. V. 2011. Creativity and Imitation: Effects of Regulatory Focus and Creative Exemplar Quality. *Creativity Research Journal* 23(4):346–356.

[Sharples 1999] Sharples, M. 1999. *How We Write: Writing as Creative Design*. Routledge London.

[Si, Marsella, and Pynadath 2005] Si, M.; Marsella, S. C.; and Pynadath, D. V. 2005. Thespian: Using Multi-Agent Fitting to Craft Interactive Drama. In *Proc.4.Int.Joint Conf.on Autonomous Agents and Multi-Agent Systems, Utrecht NL, July 25-29 2005*, 21–28. ACM Press.

[Si, Marsella, and Pynadath 2009] Si, M.; Marsella, S.; and Pynadath, D. 2009. Directorial Control in a Decision-Theoretic Framework for Interactive Narrative. In *Interactive Storytelling, 2.Joint Int.Conf.on Interactive Digital Storytelling, ICIDS 2009, Guimarães Portugal, Dec.9-11 2009. Proc.*, 221–233. Springer LNCS 5915.

[Smith and Mateas 2011] Smith, A. M., and Mateas, M. 2011. Knowledge-Level Creativity in Game Design. In *Proc.2.Int.Conf.on Computational Creativity, 27-29 Apr.2011*, 16–21. Univ. Autònoma Metropolitana MX.

[Spierling and Hoffmann 2010] Spierling, U., and Hoffmann, S. 2010. Exploring Narrative Interpretation and Adaptation for Interactive Story Creation. In *Interactive storytelling: 3.Joint Conf.on Interactive Digital Storytelling, ICIDS 2010, Edinburgh UK, Nov.1-3 2010. Proc.*, 50–61. Springer LNCS 6432.

[Struck 2005] Struck, H.-G. 2005. Telling Stories Knowing Nothing: Tackling the Lack of Common Sense Knowledge in Story Generation Systems. In *Virtual Storytelling - Using VR Technologies for Storytelling, 3.Int.Conf.(ICVS 2005), Strasbourg FR, Nov.30-Dec.2 2005. Proc.*, 189–198.

[Swartjes, Kruizinga, and Theune 2008] Swartjes, I.; Kruizinga, E.; and Theune, M. 2008. Let's pretend I had a sword: late commitment in emergent narrative. In *Proc.1.Joint Int.Conf.on Interactive Digital Storytelling (ICIDS 2008)*, 264–267. Springer LNCS 5334.

[Swartjes 2010] Swartjes, I. 2010. *Whose story is it anyway? How improv informs agency and authorship of emergent narrative*. Ph.D. Dissertation, Univ. of Twente NL.

A Meme-Based Architecture for Modeling Creativity

Shinji Ogawa
Nagoya, Aichi, Japan
perfectworld@nyc.odn.ne.jp

Bipin Indurkhya and **Aleksander Byrski**
AGH University of Science and Technology, Cracow, Poland
{bipin, olekb}@agh.edu.pl

Abstract

This research is a collaborative work between a visual artist, a computer scientist, and a cognitive scientist, and focuses on the creative process involved in connecting two pictures by painting another picture in the middle. This technique was involved in four *Infinite Landscape* workshops conducted at Art Museums in Japan and Europe over the last five years. Based on the artist's verbal recollection of the ideas that occurred to him as he drew each of the connecting pictures, we identify the micro-processes underlying these ideas, and propose a meme-based, evolutionary-inspired architecture for modeling them.

Introduction

Research in recent years has revealed that though creativity may involve an *aha* moment with a gestalt shift or a sudden perceptual or conceptual reorganization, it is typically preceded and followed by several micro-processes that play an equally important role as the *aha* moment itself (Dunbar 1997; Sawyer 2006). These micro-processes can occur within a cognitive agent itself, or in different agents within a group or society. Our goal in this research is to study and model these micro-processes.

Infinite Landscape Workshops

This research is a collaborative effort between a visual artist [henceforth *the Artist*], a computer scientist and a cognitive scientist. Over the last five years, the Artist conducted four workshops at art museums in Japan and in Europe with the common theme *connecting different spaces*. In each workshop, there were 15-19 participants, all children (8-14 years) except in one workshop there were six adults. All the workshops followed the following modus operandi.

In the first step, the children were shown about 20 photographs of scenery from around the world, and then they were asked to draw imaginary landscapes using the building, people, animals etc. in these pictures as they liked. In the second step, the Artist brought the children's imaginary landscapes to his studio, and then he drew one picture to be inserted between every two children's pictures, so that all three pictures form a seamless scene. One such trio of pictures is shown in Fig. 1: scenes 9 and 10 were drawn by participants, and the Artist drew S9 to connect the two.

Figure 1: First strip

In the third and final step, all the pictures were connected in a ring without a beginning and an end, and the completed ring was suspended from the ceiling of the museum where the workshop was held. The ring was placed with the paintings on the inner side, so that the viewer is surrounded by the work while viewing it.

Overview of the Project and Methodology

Specifically, our goal in this project is to model the micro-processes involved in creating the connecting picture. Our methodology is as follows. In the first step, the Artist has recorded various ideas that occurred to him as he drew each of the connecting pictures. In the second step, we analyze these steps to identify and classify underlying processes. In the third step, we outline a model for implementing these processes. Finally, we would like to do experiments with the implemented system and evaluate the results.

In the current paper, we report our observations from analyzing the data from the workshop conducted at the Meguro Museum of Art, Tokyo (Japan) on 2 August 2007. The Meguro workshop was different from the other three workshops in that the participants were given only pencil and paper; there was no color, so the focus was on forms, shapes and space. Also, this workshop included six adults among nineteen participants; the remaining 13 were children (8-14) years. Based on our observations, we identify various micro-processes and how they interacted with each other to create the macro-level connecting pictures. Finally, we propose a meme-based, multi-agent architecture for modeling the underlying cognitive process, and discuss future research directions.

Observations on the 'Connecting' Process

We analyzed data from ten connecting pictures that the Artist drew for this workshop. Here we present the Artist's self-reflection on the genesis of ideas that led to the creation of connecting pictures. We include here seven of the more interesting cases. (The original comments were in Japanese. Translation and slight editing is by one of the other authors of this paper.)

We start with the Artist's observations on connecting 9 and 10 (Fig. 1): "These two had completely different atmosphere from each other. Sketch 9, drawn by an adult participant, is a scene set at dusk; a person looking at the artist is drawn wearing a sad expression. Sketch 10 has a bright atmosphere with flowers, fountains, buildings on a hill, and a horse. Moreover, each picture had an important character in the bottom left. The idea for connecting these sketches came to me while looking at the wonderful horse in 10. I thought of putting a parent horse running nearby. Because the background color of 9 and the body color of the horse in 10 was the same, I transformed the background of 9 into the parent horse in S9, which became a nested image structure. Then I extended the baby horse and the hill with the buildings."

On connecting 11 and 12 (Fig. 2): "There was the ground and the sky in the left one-third of 11, but the sea covered the remaining part on the right. In 12, a vast meadow was drawn with rich pictorial details. Here my attention was drawn to the connection between the color of the giant bridge in 11 and the color of the sky in 12. In S11 I drew the enlarged bridge of 11 and connected it with the picture on 12, which resulted in a nested image structure."

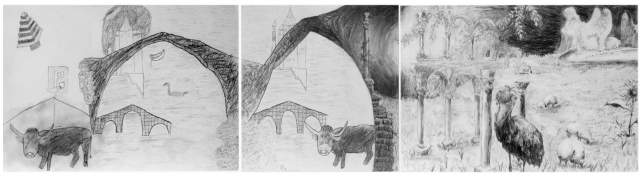


Figure 2: Second strip

On connecting 12 and 13 (Fig. 3): "I felt these two could not be connected with the techniques I had used so far. Then I noticed the wall on the top-right corner of 12 and the curved ledge surrounding the fountain in 13. Using these two curves, I drew a large Mobius strip in S12. As this Mobius strip divided S12 into four sections, in each section I extended the adjacent scenery. It felt like pouring in the scenery. Accordingly, I was able to connect them without blending, and this became the first work with this technique."

On connecting 7 and 8 (Fig. 4): "Because 8 was a richly detailed realistic presentation, to contrast it with the presentation in 7, I decided to stress dimensionality in the connection. The realistic rocks and the bridge in 8 were rendered in 3-d and were connected with the bridge in 7 that was extended in 2-d. To make this connection smoother and give an accent to the picture, I drew 3 Russian onion domes from 7 into S7."

Figure 3: Third strip

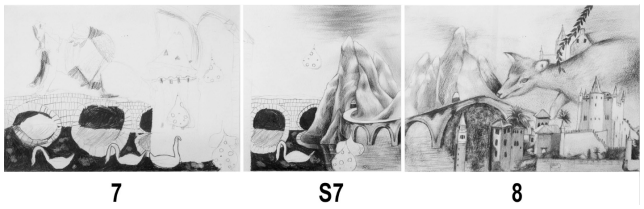


Figure 4: Fourth strip

On connecting 5 and 6 (Fig. 5): "Sketches 5 and 6 could be naturally connected. However, I had decided to refuse ordinary, conventional way of connecting things. I got the hint from the composition of 6. Oddly, on the right of 6, everything is drawn tilted towards the bottom left along a vector, but in the middle part, another horizontal vector appears. As a result, the horizon is split into two: one horizontal and another pointing to bottom left. I further emphasized this split of horizon, and drew a horizon pointing to the sky where the cow is, and another horizon that is sinking down where the buildings are."

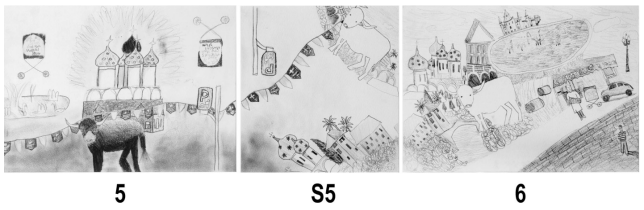


Figure 5: Fifth strip

On connecting 4 and 5 (Fig. 6): "Connect 5 on the right of 4. I was very interested in the row of flags that was hanging in 5 from left to right. On the right edge of 4, there is an upside-down building. What a challenge! I took that challenge and extended the gate of that fort-like building, and turned the top-right part of it into water surface. I extended that dark water surface to the right, making it narrower, and connected it with the contour of the lake in 5. On top of it, I placed the swans and plants from 5. I left the top-right part of the picture white in order create a contrast effect with the black space that is extended to the left. In the bottom right, I extended the flags."

On connecting 3 and 4 (Fig. 7): "I had a strong impression that the participants were expressing their own images instead of sketching by sampling from the photographs of the scenery I had shown. An extreme case of this is 4.

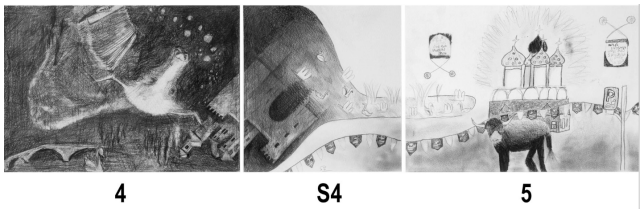


Figure 6: Sixth strip

At first the sketch was filled-in completely black, and then brightened by the eraser. It had no earth and sky, but an ambiguous space from a dark fantasy I decided to connect this dark picture with 3, which had a child-like pictorial space. However, it would be impossible to connect the two in an ordinary way. Here, I decided to ignore all the meaning in these pictures, but to focus on the pattern of light and dark instead. I said to myself, 'it is just a blotch'. The only connecting point in both pictures was the street in 3 and the bridge on the bottom left of 4. I could connect this street and the bridge. Luckily, bottom left of 4 looked like the sea, and bottom right of 3 also looked like a body of water. In S3, I extended the road in 3 in S-shaped curve and connected it with the bridge in 4. Continuing, I also extended the sea The problem was what to do on top of this. On the left part of S3, the only possibility was to extend the street-side houses on 3, so I did that in the same touch. Then I gradually changed the color of houses from gray to black, while introducing spatial distortion, and changing them from solid to liquid. I floated a swan in the dark pond that the buildings were turned into."

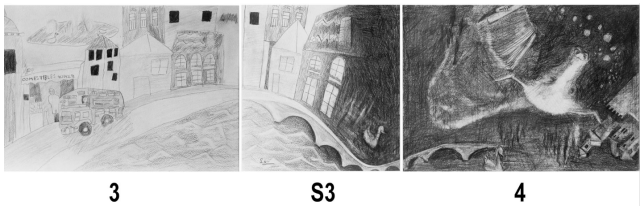


Figure 7: Seventh strip

Identifying Micro-Processes in 'Connecting'

Carefully going through all these comments, as well as examining the trio of pictures ourselves, we came with the following list of micro-processes that played a role in the genesis of connecting two pictures:

Copy elements This was by far the most common operation. Elements were copied from both the left and right pictures and incorporated in the connecting pictures just like that. One can see examples of this in almost every instance of connection. Among the examples presented above, one can see that swan is copied from 4 to S3, flags, plants and swan from 5 to S4, onion domes and swan from 7 to S7, small bridge and bull from 11 to S11, and so on.

Copy elements and transform This is similar to the above except that the element gets transformed while copying. For example, the rocky peak and the bridge are rendered in 3-d while they are copied from 8 into S7, and parking sign is turned around as it is copied from 5 into S5.

Copy elements and swap attributes Here elements that are being copied interact during copying and swap attributes. One example is provided in 6-S6-7 (Fig. 8), where two people are copied from 6 into S6, but their poses and the object one of them is holding are swapped.

Extend elements An element is continued in the adjacent picture; for example, the sea from 3 into S3, the masonry from 6 into S6, and the meadow from 12 into S12.

Same form (shape, shade,...) → search for meaning This is illustrated by 9-S9-10 (Fig. 1), where the same shading for the horse's body in 10 and the background in 9 led to the idea that the background in 9 can be morphed in the mother horse in S9. This process can also be evidenced between S11 and 12 (Fig. 2).

Similar form and semantic association → morph forms This is evidenced in 3-S3-4 (Fig. 7), where a semantic association between the road and the bridge, and similar forms (notice that they are similar but not the same) led to the idea that they can be joined by morphing one into the other.

Form-based continuation This is different from the extend element above in that the continuation is based on the shape and shade only, and does not involve meaning. This is seen in S3 and 4 (Fig. 7).

Form-contrast → concept-contrast This is illustrated by 7-S7-8 (Fig. 4). The contrast between a richly detailed sketch (8) and a plain sketch (7) suggested a 3-d vs. 2-d contrast.

Form-similarity → unifying concept In 12-S12-13 (Fig. 3) form-similarity between the wall on the top right of 12 and the ledge around the fountain on the bottom left of 13 suggested the idea of a Möbius strip.

Emphasize concept In 5-S5-6 (Fig. 5), different planes (horizons) in 6 were incorporated in S5 and emphasized. This is similar to *copy element* and *transform* except that the element is a concept rather then a concrete object.

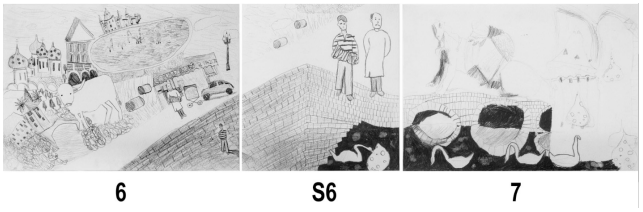


Figure 8: Eight strip

Meme: A Representation for Ideas

In order to represent all these micro-processes, we propose to use the formalism of *meme,* which was popular-

ized by Richard Dawkins in his celebrated *The Selfish Gene* (Dawkins 1989). Memes are cultural counterpart of genes, and represent ideas that can be generated, be passed on, get transformed, be combined with each other, and die out. As we observed many of the similar operations and interactions among the micro-processes in connecting two pictures, we chose meme as a unit of representation for modeling.

In our particular domain, a meme can be an element like a swan, a horse, or a building. It is a particular element, so it carries specific attributes. In other words, the horse meme that plays a key role in 9-S9-10 (Fig. 1) is not the general concept of a horse, but carries concrete attributes like the shade and the shape of the horse that was drawn in 10. There can also be conceptual memes, for example, 'horizon tilted to top-right', '3-d rendering', or 'dark shade'. Such memes represent specific operations or attributes that can be imparted to an element or a scene.

It is possible to have generalized memes and to organize them in a hierarchy. So, for example, there can be a 'horse' meme of which the horse meme of 10 (Fig. 1) would be an instance; or there can be a 'tilted horizon' meme, which would be a parent of the 'horizon tilted to the top-right' meme. But for the time being we are not considering such general memes.

Following actions can be carried out on individual memes:

Copy or replicate In this case the element is copied as it is, or the concept is applied as it is. So a swan is copied with all its attributes intact, or the horizon can be tilted toward the top-right corner of the pictures for a part of the scene that is selected.

Copy with transformation In this case, the element is copied but one or more of its attributes are changed along the same dimension. For example, its size can be made bigger or smaller, its color or shade can be changed, its orientation can be changed, and so on. For a concept meme, some of its parameters are changed during application; for example 'horizon tilted to top-right' can change to 'horizon tilted to top-left'.

Two memes can also interact with each other and we specify the following four modes of interaction:

Swap attributes Two memes can swap attributes of each other. We saw an example of this in Fig. 8 where the pose and the "object-held' attributes of two people were exchanged.

Overwrite attribute In this case the attribute of one meme overwrites the attribute of the other meme. So, for example, the size or the color of one meme can be rendered according to the other meme. This is illustrated by an instance in the Osaka workshop, where the silhouettes of cliffs were made to conform to the silhouettes of buildings.

Unify This allows two memes to bond together and act as one meme. Any common attributes of the two become the attributes of the unified meme, and in addition some extra attributes may be created based on the spatial or other relationships between the two. This is similar to the grouping operation in many graphic editors.

Create a new meme This allows creation of a new meme with attributes inherited from each of the parent memes.

There are a number of other features that we are not considering at the moment. For example, it may be possible for a meme to activate another meme. We saw an example of it in our observation above when the Mobius strip idea was suggested by the similarity in form between the wall and the fountain ledge 3. However, in order to model this mechanism, we need to have some kind of global associative knowledge network.

A Memetic Architecture

We are implementing a meme-based system to model the process of creating the intermediate picture. In particular, our system incorporates the following features: 1) modeling of visual attention to identify prominent elements or areas in the neighboring pictures; 2) specifying memes for spatial relationships among the picture elements; 3) specifying memes for general techniques like extension and continuation; and 4) various heuristics for choosing among competing memes.

For a lack of space, and also as our system is currently being implemented, we limit ourselves to only pointing out that we are exploring two approaches to generating the connecting image for a given pair of images:

- Evolutionary algorithm: the two images should be digitalized, and potential solution generated stochastically from them with the use of crossover and mutation (Michalewicz 1998). Formulation of the fitness function should take into consideration the similarity of the potential solution to both of images. We also plan to incorporate aesthetic criteria in the fitness function (Norton, Heath, and Ventura 2010).
- Agent-based approach: some complex approaches utilizing multi-agent notions (Byrski and Kisiel-Dorohinicki 2005)) bring interesting additions to the process, as autonomous individuals, as agents are, may utilize other means to evaluate the resulting images, and may choose different crossover and mutation operators in an intelligent way to apply them to the current solution.

Both approaches may leverage concepts well-known from the memetic computation—local search (Moscato and Cotta 2010)—thus applying a number of mutation operators (instead of only one) before final evaluation.

Relation with Previous Research

Needless to say, the ideas and the architecture presented here are based on a number of existing and past research efforts to model different aspects of creativity. The origin of the parallel, competition-cooperation architecture can be traced back to Selfridge (Selfridge 1959). Subsequently, Lesser et al. (Lesser, Fennell, and Reddy 1975) formalized it as *blackboard* architecture and used it for speech recognition; and in our earlier research (Indurkhya 1997) we used a similar approach to model creativity in legal reasoning. Hofstadter

and his colleagues (Hofstadter 1995) proposed a *parallel terraced scan* architecture for modeling creativity in analogical thinking and our approach outlined above is heavily influenced by their work. One key point of difference is that a meme is more like an agent that carries its own data with it, unlike a knowledge source in the blackboard architecture or a codelet in Hofstadter's architecture.

The system proposed above also draws from the *meme media* architecture of Tanaka, Fujima and Kuwahara (Tanaka and Kuwahara 2008). They have developed the C3W wrapper framework that allows the user to open a web application page, clip some input and output portions as pads, and link them with pads clipped from other web applications.

A number of approaches have been developed for applying evolutionary algorithms to generate visual art (Sims 1991; Lewis 2007; Machado, Romero, and Manaris 2007), but their goal is to generate aesthetically pleasing visual objects. In the long run, it may be possible to use some of these techniques by incorporating constraints from the neighboring picture objects to generate novel but related picture objects for the connecting picture.

As for systems that generate constrained visual objects or scenes, there has been some research on automatic collage generation (Krzeczkowska 2009) and on completing a partially drawn picture in the intended style (Colton 2008), and some of the techniques developed therein can be exploited in our system as well.

Conclusions and Future Research

We analyzed data from the Artist's verbal recollection of his thoughts as he drew the middle pictures to connect pairs of pictures seamlessly. From this analysis, we identified a number of micro-processes that led to the big picture idea. We described a memetic approach to formalize these micro-processes, and outlined an evolutionary-inspired approach to support the process of generating the connecting picture.

We are also interested to study the cognitive processes of the viewers as they look at the trio of pictures. It has been noted in the past that surface-level perceptual similarities influence how viewers connect pairs of images and relate them conceptually (Indurkhya et al. 2008). It would be interesting to see how this process is affected when there is an intervening picture in the middle. We plan to conduct behavioral and eye-tracking experiments to measure the viewers' response and incorporate those observations in our model.

References

Byrski, A., and Kisiel-Dorohinicki, M. 2005. Immunological selection mechanism in agent-based evolutionary computation. In Klopotek, M. A.; Wierzchon, S. T.; and Trojanowski, K., eds., *Intelligent Information Processing and Web Mining : proceedings of the international IIS: IIPWM '05 conference : Gdansk, Poland*, Advances in Soft Computing, 411–415. Springer Verlag.

Colton, S. 2008. Experiments in constraint-based automated scene generation. In *Proc. of the 5th International Joint Workshop on Computational Creativity*.

Dawkins, R. 1989. *The Selfish Gene (2nd ed.)*. Oxford University Press.

Dunbar, K. 1997. How scientists think: On-line creativity and conceptual change in science. In Ward, T.; Smith, S.; and Vaid, J., eds., *Creative thought: An investigation of conceptual structures and processes*. American Psychological Association.

Hofstadter, D. 1995. *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought*. New York: Basic Books.

Indurkhya, B.; Kattalay, K.; Ojha, A.; and Tandon, P. 2008. Experiments with a creativity-support system based on perceptual similarity. In Fujita, H., and Zualkernan, I., eds., *New Trends in Software Methodologies, Tools and Techniques*. IOS Press: Amsterdam. 316–327.

Indurkhya, B. 1997. On modeling creativity in legal reasoning. In *Proceedings of the Sixth International Conference on AI and Law*, 180–189.

Krzeczkowska, A. 2009. Automated collage generation from text. Master's thesis, Imperial College, London.

Lesser, V.; Fennell, R.D. nad Erman, L.; and Reddy, D. 1975. Organization of the hearsay-ii speech understanding system. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 23(1):11–24.

Lewis, M. 2007. Evolutionary visual art and design. In Romero, and Machado, P., eds., *The Art of Artificial Evolution: A Handbook on Evolutionary Art and Music*, 3–37. Berlin: Springer-Verlag.

Machado, P.; Romero, J.; and Manaris, B. 2007. Experiments in computational aesthetics: An iterative approach to stylistic change in evolutionary art. In Romero, and Machado, P., eds., *The Art of Artificial Evolution: A Handbook on Evolutionary Art and Music*, 381–415. Berlin: Springer-Verlag.

Michalewicz, Z. 1998. *Genetic Algorithms + Data Structures = Evolution Programs*. Springer.

Moscato, P., and Cotta, C. 2010. A modern introduction to memetic algorithms. In Gendrau, M., and Potvin, J.-Y., eds., *Handbook of Metaheuristics*, volume 146 of *International Series in Operations Research and Management Science*. Springer, 2 edition. 141–183.

Norton, D.; Heath, D.; and Ventura, D. 2010. Establishing appreciation in a creative system. In *Proceedings of the International Conference on Computational Creativity: ICCC-X,*, 26–35.

Sawyer, K. 2006. *Explaining Creativity*. Oxford University Press.

Selfridge, O. 1959. Pandemonium: A paradigm for learning. In Blake, D., and Uttley, A., eds., *Proc. of the Symposium on Mechanisation of Thought Processes*, 511–529. H.M. Stationery Office: London.

Sims, K. 1991. Artificial evolution for computer graphics. *ACM Computer Graphics* 25:319–328.

Tanaka, Y., F. J., and Kuwahara, M. 2008. Meme media and knowledge federation. In Dengel, A., and et al., eds., *Proc. of KI2008: LNAI 5243*, 2–21. Springer Verlag.

Corpus-Based Generation of Content and Form in Poetry

Jukka M. Toivanen, Hannu Toivonen, Alessandro Valitutti and Oskar Gross
Department Of Computer Science and
Helsinki Institute for Information Technology, HIIT
University of Helsinki, Finland

Abstract

We employ a corpus-based approach to generate content and form in poetry. The main idea is to use two different corpora, on one hand, to provide semantic content for new poems, and on the other hand, to generate a specific grammatical and poetic structure. The approach uses text mining methods, morphological analysis, and morphological synthesis to produce poetry in Finnish. We present some promising results obtained via the combination of these methods and preliminary evaluation results of poetry generated by the system.

Introduction

Computational poetry is a challenging research area of computer science, at the cross section of computational linguistics and artificial intelligence. Since poetry is one of the most expressive ways to use verbal language, computational generation of texts recognizable as good poems is difficult. Unlike other types of texts, both content and form contribute to the expressivity and the aesthetical value of a poem. The extent to which the two aspects are interrelated in poetry is a matter of debate (Kell 1965).

In this paper we address the issues of generating content and form using corpus-based approaches. We present a poetry generator in which the processing of content and form is performed through access to two separate corpora, with minimal manual specification of linguistic or semantic knowledge.

In order to automatically obtain world knowledge necessary for building the content, we use text mining on a *background corpus*. We construct a word association network based on word co-occurrences in the corpus and then use this network to control the topic and semantic coherence of poetry when we generate it.

Many issues with the form, especially the grammar, we solve by using a *grammar corpus*. Instead of using an explicit, generative specification of the grammar, we take random instances of actual use of language from the grammar corpus and copy their grammatical structure to the generated poetry. We do this by substituting most words in the example text by ones that are related to the given topic in the word association network.

Our current focus is on testing these corpus-based principles and their capability to produce novel poetry of good quality on a given topic. At this stage of research, we have not yet considered rhyme, rhythm or other phonetic features of the form. These will be added in the future, as will more elaborate mechanisms of controlling the content.

As a result of the corpus-based design, the input to the current poetry generator consists of the background and the grammar corpora, and the topic of the poem. In the intended use case, the topic is directly controlled by the user, but we allow the grammar corpus to influence the content, too. Control over form is indirectly over the choice of the two corpora. The only directly language-dependent component in the system is an off-the-shelf module for morphological analysis and synthesis.

The current version of our poetry generation system works in Finnish. Its rich morphology adds another characteristic to the current implementation. However, we believe that the flexible corpora-based design will be useful in transferring the ideas to other languages, as well as in developing applications that can adapt to new styles and contents. A possible application could be a news service in the web, with a poem of the day automatically generated from recent news and possibly triggering, in the mind of the reader, new views to the events of the world.

After briefly reviewing related work in the next section, we will describe the corpus-based approach in more detail. Then, we will give some examples of generated poetry, with rough English translations. We have carried out an empirical evaluation of the generated poetry with twenty subjects, with encouraging results. We will describe this evaluation and its results, and will then conclude by discussing the proposed approach and the planned future work.

Related Work

The high complexity of creative language usage poses substantial challenges for poetry generation. Nevertheless, several interesting research systems have been developed for the task (Manurung, Ritchie, and Thompson 2000; Gervás 2001; Manurung 2003; Diaz-Agudo, Gervás, and González-Calero 2002; Wong and Chun 2008; Netzer et al. 2009). These systems vary a lot in their approaches, and many different computational and statistical methods are often combined in order to handle the linguistic complexity and creativity aspects. State of the art in lexical substitution but not in poetical context is presented, for instance, by Guerini et

al. (2011). We next review some representative poetry generation systems.

ASPERA (Gervás 2001) employs a case-based reasoning approach. It generates poetry out of a given input text via a composition of poetic fragments that are retrieved from a case-base of existing poems. In the system case-base each poetry fragment is annotated with a prose string that expresses the meaning of the fragment in question. This prose string is then used as the retrieval key for each fragment. Finally, the system combines the fragments by using additional metrical rules. In contrast, our "case-base" is a plain text corpus without annotations. Additionally, our method can benefit from the interaction of two distinct corpora for content and form.

The work of Manurung et al. (2000) draws on rich linguistic knowledge (semantics, grammar) to generate a metrically constrained poetry out of a given topic via a grammar-driven formulation. This approach requires strong formalisms for syntax, semantics, and phonetics, and there is a strong unity between the content and form. Thus, this system is quite different from our approach. The GRIOT system on its part (Harrell 2005) is able to produce narrative poetry about a given theme. It models the theory of conceptual blending (Fauconnier and Turner 2002) from which an algorithm based on algebraic semantics was implemented. In particular, the approach employs "semantics based interaction". This system allows the user to affect the computational narrative and produce new meanings.

The above mentioned systems have rather complex structures involving many different interacting components. Simpler approaches have also been used to generate poetry. In particular, Markov chains (n-grams) have been widely used as the basis of poetry generation systems as they provide a clear and simple way to model some syntactic and semantic characteristics of language (Langkilde and Knight 1998). However, the characteristics are local in nature, and therefore standard use of Markov chains tends to result in poor sentence and poem structures. Furthemore, form and content are learned from a single corpus and cannot be easily separated.

Methods

We next present our approach to poetry generation. In the basic scenario, a topic is given by the user, and the proposed method then aims to give as output a novel and non-trivial poem in grammatically good form, and with coherent content related to the given topic.

A design principle of the method is that explicit specifications are kept at minimum, and existing corpora are used to reduce human effort in modeling the grammar and semantics. Further, we try to keep language-dependency of the methods small.

The poetry generator is based on the following principles.

1. *Content:* The topics and semantic coherence of generated poetry are controlled by using a simple word association network. The network is automatically constructed from a so-called background corpus, a large body of text used as a source of common-sense knowledge. More specifically, the semantic relatedness of word pairs is extracted from their co-occurrence frequency in the corpus. In the experiments of this paper, the background corpus is Finnish Wikipedia.

2. *Form (grammatical):* The grammar, including the syntax and morphology of the generated poetry, is obtained in an instance-based manner from a given grammar corpus. Instead of explicitly representing a generative grammar of the output language, we copy a concrete instance from an existing sentence or poem but replace the contents. In our experiments, the corpus consists mainly of old Finnish poetry.

3. *Form (phonetic):* Rhythm, rhyme, and other phonetic features can, in principle, be controlled when substituting words in the original text by new ones. This part has not been implemented yet but will be considered in future work.

The current poetry generation procedure can now be outlined as follows:

- A topic is given (or randomly chosen) for the new poem. The topic is specified by a single word.
- Other words associated with the topic are extracted from the background graph (see below).
- A piece of text of the desired length is selected randomly from the grammar corpus.
- Words in the text are analyzed morphologically for their part of speech, singular/plural, case, verb tense, clitics etc.
- Words in the text are substituted independently, one by one, by words associated with the topic. The substitutes are transformed to similar morphological forms with the original words. The original word is left intact, however, if there are no words associated with the topic that can be transformed to the correct morphological form.
- After all words have been considered, the novelty of the poem is measured by the percentage of replaced words. If the poem is sufficiently novel it is output. Otherwise the process can be re-tried with a different piece of text.

For the experiments of this paper, we require that at least one half of the words were replaced. This seems sufficient to make readers perceive the new topic as the semantic core of the poem.

We next describe in some more detail the background graph construction process as well as the morphological tools used.

Background Graph

A background graph is a network of common-sense associations between words. These associations are extracted from a corpus of documents, motivated by the observation that (frequent) co-occurrence of words tends to imply some semantic relatedness between them (Miller 1995).

The background graph is constructed from the given background corpus using the log-likelihood ratio test (LLR). The log-likelihood ratio, as applied here for measuring associations between words, is based on a multinomial model of word co-occurrences (see, e.g., Dunning (1993) for more information).

The multinomial model for a given pair $\{x, y\}$ of words has four parameters $p_{11}, p_{12}, p_{21}, p_{22}$ corresponding to the probability of their co-occurrence as in the contingency table below.

	x	$\neg x$	Σ
y	p_{11}	p_{12}	$p(y;C)$
$\neg y$	p_{21}	p_{22}	$1-p(y;C)$
Σ	$p(x;C)$	$1-p(x;C)$	1

Here, $p(x;C)$ and $p(y;C)$ are the marginal probabilities of word x or word y occurring in a sentence in corpus C, respectively.

The test is based on the likelihoods of two such multinomial models, a null model and an alternative model. For both models, the parameters are obtained from relative frequencies in corpus C. The difference is that the null model assumes independence of words x and y (i.e., by assigning $p_{11} = p(x;C)p(y;C)$ etc.), whereas the alternative model is the maximum likelihood model which assigns all four parameters from their observed frequencies (i.e., in general $p_{11} \neq p(x;C)p(y;C)$).

The log-likelihood ratio test is then defined as

$$LLR(x,y) = -2\sum_{i=1}^{2}\sum_{j=1}^{2} k_{ij}\log(p_{ij}^{null}/p_{ij}), \qquad (1)$$

where k_{ij} is the respective number of occurrences. It can be seen as a measure of how much the observed joint distribution of words x and y differs from their distribution under the null hypothesis of independence, i.e., how strong the association between them is. More complex models, such as LSA, pLSA or LDA could be used just as well.

Finally, edges in the background graph are constructed to connect any two words x, y that are associated with $LLR(x,y)$ greater than an empirically chosen threshold. To find words that are likely semantically related to the given topic, first-level neighbours (i.e., words association with the topic word) are extracted from the background graph. If this set is not large enough (ten words or more in the experiments of this paper), we add randomly selected second-level neighrbours (i.e., words associated to any of the first-level neighbors).

In the future, we plan to use edge weights to control the selection of substitutes, and possibly to perform more complex graph algorithms on the background graph to identify and choose content words.

Morphological Analysis and Synthesis

Morphological analysis is essential and non-trivial for morphologically rich languages such as Finnish or Hungarian. In these languages, much of the language's syntactic and semantic information is carried by morphemes joined to the root words. For instance, the Finnish word "juoksentelisinkohan" (I wonder if I would run around) is formed out of the root word "juosta" (run). Hence, morphological analysis provides valuable information of the syntax and to some degree of the semantics. In our current system, morphological analysis and synthesis are carried out using Omorfi[1], a morphological analyzer and generator of Finnish language based on finite state automata methodology (Lindén, Silfverberg, and Pirinen 2009).

[1]URL: http://gna.org/projects/omorfi

With the help of Omorfi we can thus generate substitutes that have similar morphological forms with the original words. For instance, assume that the topic of the poetry is "ageing" and we want so substitute "juoksentelisinkohan" by a word based on "muistaa" (remember). Omorfi can now generate "muistelisinkohan" (I wonder if I would think back) as a morphologically matching word.

Examples

We next give some example poems generated by the current system with the original example texts used to provide structure for these poems. We also give their rough English translations, even though we suspect that poetical aesthetics somewhat change in translation. The substituted words are indicated by *italics*.

The first example poem is generated around the topic "(children's) play". We first give the Finnish poem with the template used to construct it (on the right) and then the English translation of both the generated and original poems.

Kuinka hän leikki *silloin*	kuinka hän leikki kerran
uskaliaassa, uskaliaassa //	suuressa vihreässä //
kuiskeessa	puistossa
vaaleiden puiden alla.	ihanien puiden alla.
Hän oli *kuullut* huvikseen,	Hän oli katsellut huvikseen,
kuinka hänen *kuiskeensa*	kuinka hänen hymynsä
kanteli helkkeinä tuuloseen.	putosi kukkina maahan,

Original by Uuno Kailas: Satu meistä kaikista, 1925

How she played *then*	how she played once
in a *daring, daring whispering*	in a big green park
under the *pale* trees.	under the lovely trees.
She had *heard* for fun	She had watched for fun
how her *whispering*	how her smile
drifted as jingle to the wind.	falled down as flowers,

The next poem is generated with "hand" as the topic. The template used is shown below the generated poem and thereafter the translations, respectively.

Vaaleassa kourassa
sopusuhtaisessa kourassa ovat *nuput* niin kalpeita
kuvassasi lepää *lapsikulta* jumala.

Alakuloisessa metsässä
Hämärässä metsässä ovat kukat niin kalpeita
Varjossa lepää sairas jumala

Original by Edith Södergran: Metsän hämärä, 1929

In a *pale fist*
in a *well-balanced fist*, the *buds* are so pale
in *your image* lies a *dear child* god.

In a gloomy forest
In a dim forest flowers are so pale
In the shadow lies a sick god

The final example poem has "snow" as its topic.

Elot sai karkelojen teitä,	Aallot kulki tuulten teitä,
lumi ajan *kotia*,	aurinko ajan latua,
hiljaa *soi kodit autiot*,	hiljaa hiihti päivät pitkät,
hiljaa *sai armaat karkelot* -	hiljaa hiipi pitkät yöt -
laiho sai lumien riemut.	päivä kutoi kuiden työt,

Original by Eino Leino: Alkusointu, 1896

Lives got the frolic ways,
snow the home of time,
softly *chimed abandoned homes*,
softly *got frolics beloved* -
ripening crop got the snows' joys.

Waves fared the wind's ways,
sun the track of time,
slowly skied for long days,
slowly crept for long nights -
day wove the deeds of moons

Subjectively judging, the generated poems show quite a wide range of grammatical structures, and they are grammatically well formed. The cohesion of the contents can also be regarded as fairly high. However, the quality of generated poetry varies a lot. Results from an objective evaluation are presented in the next section.

Evaluation

Evaluation of creative language use is difficult. Previous suggestions for judging the quality of automatically generated poetry include passing the Turing test or acceptance for publishing in some established venue. Because the intended audience of poetry consists of people, the most pragmatic way of evaluating computer poetry is by an empirical validation by human subjects. In many computer poetry studies both human written and computationally produced poetry have been evaluated for qualities like aesthetic appreciation and grammaticality.

In this study we evaluated poetry using a panel of twenty randomly selected subjects (typically university students). Each subject independently evaluated a set of 22 poems, of which one half were human-written poems from the grammar corpus and the other half computer-generated ones with at least half of the words replaced. The poems were presented in a random order, and the subjects were not explicitly informed that some of the poems were computer-generated.

Each subject evaluated each text (poem) separately. The first question to answer was if the subject considered the piece of text to be a poem or not, with a binary yes/no answer. Then each text was evaluated qualitatively along six dimensions: (1) How typical is the text as a poem? (2) How understandable is it? (3) How good is the language? (4) Does the text evoke mental images? (5) Does the text evoke emotions? (6) How much does the subject like the text? These dimensions were evaluated on the scale from one (very poor) to five (very good). (The interesting question of how the amount of substituted words affects the subjective experience of topic, novelty and quality is left for future research.)

Evaluation results averaged over the subjects and poems are shown in Figures 1 and 2. Human-written poems were considered to be poems in 90.4% of the time and computer-generated poems 81.5% of the time (Figure 1). Intervals containing 66.7% of the poems show that there was more variation in the human-written poetry than in the computer generated poetry. Overall, these are promising results, even though statistically the difference between human-written and computer generated poetry is significant (p-value with Wilcoxon rank-sum test is 0.02).

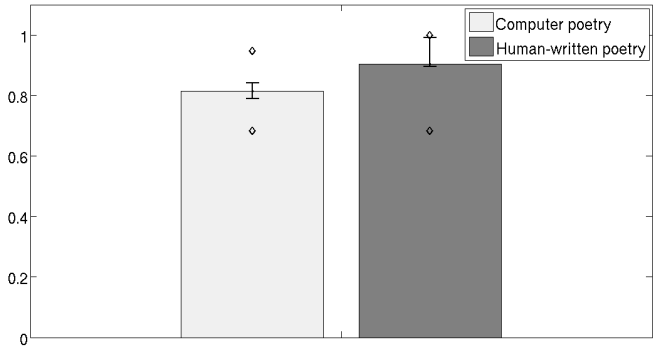


Figure 1: Relative amounts of texts (computer-generated and human-written poetry) subjectively considered to be poems, averaged over all subjects. The whiskers indicate an interval of 66.7% of poems around the median. Points indicate the best and worst poems in the both groups.

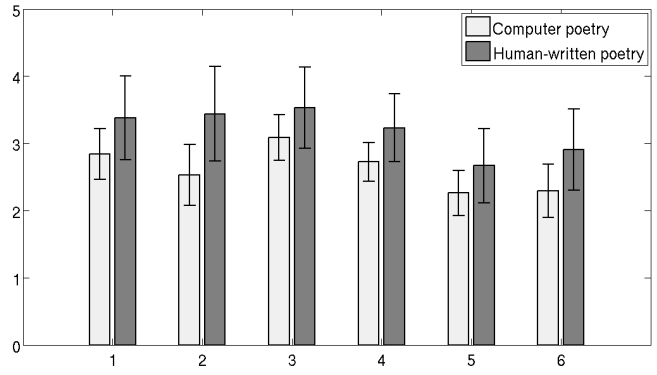


Figure 2: Subjective evaluation of computer-generated and human-written poetry along six dimensions: (1) typicality as a poem, (2) understandability, (3) quality of language, (4) mental images, (5) emotions, and (6) liking (see text for details). Results are averaged over all subjects and poems. The whiskers indicate one standard deviation above and below the mean.

The evaluated qualities have a similar pattern (Figure 2): The average difference between human-written and computer-generated poetry is not large, and in many cases there is a lot of overlap in the ranges of scores, indicating that a good amount of (best) computer-generated poems were as good as (worst) human-written ones. Statistically, however, the differences are highly significant (all p-values below 0.001). The biggest drop in quality was in understandability (dimension 2). However, somewhat controversially, the language remained relatively good (dimension 3). An interesting observation is that some of the generated poems were rated to be quite untypical but their language quality and pleasantness were judged to be relatively high.

Discussion

We have proposed a flexible poetry generation system which is potentially able to produce poetry out of a wide variety of different topics and in different styles. The flexibility is achieved by automating the processes of acquiring and applying world knowledge and grammatical knowledge. We use two separate corpora: background corpus for mining lexical associations, and grammar corpus for providing grammatical and structural patterns for the basis of new poetry. We have implemented the system for Finnish, a morphologically rich language. We carried out a preliminary evaluation on the produced poetry, with promising results.

It may be questioned whether the current approach exhibits creative behaviour, and whether the system is able to produce poetry that is interesting and novel with respect to the text that is used as the basis of new poetry. First, the generated poems are usually very different from the original texts (our subjective view, to be evaluated objectively in the future). Second, some of the generated texts were rated to be quite untypical, even though recognized as poems. The pleasantness and language quality of these poems were still judged to be relatively high. According to these observations we think that at least some of the system's output can be considered to be creative. Thus, the system could be argued to automatically piggyback on linguistic conventions and previously written poetry to produce novel and reasonably high quality poems.

Our aim is to develop methods that can be applied to other languages with minimal effort. In our current system, morphological analysis and synthesis are clearly the most strongly language-specific components. They are fairly well isolated and could, in principle, be replaced by similar components for some other language. However, it may prove to be problematic to apply the presented approach to more isolating languages (i.e., with a low morpheme-per-word ratio), such as English. In agglutinative languages (with higher morpheme-per-word ratio), such as Finnish, a wide variety of grammatical relations are realized by the use of affixation and the word order is usually quite free. We currently consider implementing the system for other languages, in order to identify and test principles that could carry over to some other languages.

So far, we have not considered controlling rhythm, rhyme, alliteration or other phonetic aspects. We plan to use constraint programming methods in the lexical substitution step for this purpose. At the same time, we doubt this will be always sufficient in practice since the space of suitable substitutes can be severely constrained by grammar and semantics. Another interesting technical idea is to use n-gram language models for computational assessment of the coherence of produced poetry.

We consider the approach described in this paper to be a plausible building block of more skillful poetry generation systems. The next steps we plan to take, in addition to considering phonetic aspects, includes trying to control the emotions that the poetry exhibits or evokes. We are also interested in producing computer applications of adaptive or instant poetry.

Acknowledgements: This work has been supported by the Algorithmic Data Analysis (Algodan) Centre of Excellence of the Academy of Finland.

References

Diaz-Agudo, B.; Gervás, P.; and González-Calero, P. A. 2002. Poetry generation in COLIBRI. In *ECCBR 2002, Advances in Case Based Reasoning*, 73–102.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics* 19(1):61–74.

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

Gervás, P. 2001. An expert system for the composition of formal spanish poetry. *Journal of Knowledge-Based Systems* 14(3–4):181–188.

Guerini, M.; Strapparava, C.; and Stock, O. 2011. Slanting existing text with valentino. In Pu, P.; Pazzani, M. J.; André, E.; and Riecken, D., eds., *Proceedings of the 2011 International Conference on Intelligent User Interfaces*, 439–440. ACM.

Harrell, D. F. 2005. Shades of computational evocation and meaning: The GRIOT system and improvisational poetry generation. In *In Proceedings, Sixth Digital Arts and Culture Conference*, 133–143.

Kell, R. 1965. Content and form in poetry. *British Journal of Aesthetics* 5(4):382–385.

Langkilde, I., and Knight, K. 1998. The practical value of n-grams in generation. In *Proceedings of the International Natural Language Generation Workshop*, 248–255.

Lindén, K.; Silfverberg, M.; and Pirinen, T. 2009. HFST tools for morphology - an efficient open-source package for construction of morphological analyzers. In *Proceedings of the Workshop on Systems and Frameworks for Computational Morphology*, 28–47.

Manurung, H. M.; Ritchie, G.; and Thompson, H. 2000. Towards a computational model of poetry generation. In *Proceedings of AISB Symposium on Creative and Cultural Aspects and Applications of AI and Cognitive Science*, 79–86.

Manurung, H. 2003. *An evolutionary algorithm approach to poetry generation*. Ph.D. Dissertation, University of Edinburgh, Edinburgh, United Kingdom.

Miller, G. A. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

Netzer, Y.; Gabay, D.; Goldberg, Y.; and Elhadad, M. 2009. Gaiku : Generating haiku with word associations norms. In *Proceedings of NAACL Workshop on Computational Approaches to Linguistic Creativity*, 32–39.

Wong, M. T., and Chun, A. H. W. 2008. Automatic haiku generation using VSM. In *Proceedings of ACACOS*, 318–323.

Crossing the Theshold Paradox: Modelling Creative Cognition in the Global Workspace

Geraint A. Wiggins
Centre for Digital Music
School of Electronic Engineering and Computer Science
Queen Mary University of London
Mile End Road, London E1 4NS, UK
geraint.wiggins@eecs.qmul.ac.uk

Abstract

I present a hypothetical global model of everyday creative cognition located within Baars' Global Workspace Theory, based on theories of predictive cognition and specific work on statistical modelling of music perception. The key idea is a proposal for regulating access to the Global Workspace, overcoming what Baars calls the Threshold Paradox. This idea is motivated as a general mechanism for managing the world, and an argument is given as to its evolutionary value. I then show how this general mechanism produces effects which are indistinguishable from spontaneous creative inspiration, best illustrated by Wallas' (1926) "Aha!" moment. I argue that W. A. Mozart's introspective account of compositional experience closely matches the proposed process, and refer to a computational system which will form the basis of an implementation of the ideas, for musical composition.

Introduction

Computational Creativity is mired, practically speaking, in the problem of evaluation. Artefacts created by computer cannot be judged by the computer's aesthetic, for that is obscure, and evaluating them in terms human aesthetics has been shown to be unreliable due to negative preconceptions (Moffat and Kelly, 2006). One solution to this might be to compensate for that bias statistically, given the necessary models. Another is to avoid the issue of artefact evaluation altogether, and focus on process, and on building systems that apply it. Colton (2009) catchily entitles this point "Paradigms Lost", making the point that AI sometimes overtheorises, and paints itself into a corner by the application of problem solving methods to a domain in the abstract, instead of getting on and building something that concretely explores it: the subtext may be that this tendency arises from rigour envy. Colton raises a point that benefits from emphasis: he finishes the section with "the production of beautiful, interesting and valuable artefacts", and this occludes the key point in the final sentence: "the need to embrace *entire* intelligent tasks" (my italics).

The vexed question is "how?" Modelling an *entire, novel* creative process, evaluation, reflection, and all, in the abstract leads us back to the initial problem: the only way to judge it from outside is in terms of its outputs (Ritchie, 2001).

The rest of the paper is structured as follows. First, I explain the theoretical methods used, and make and important distinction between what I shall call *inspiration* and *creative reasoning*; the current proposal addresses only the first of these. Next, I describe some of the extended background to the thinking presented here, and in terms of the surrounding and supporting cognitive theory, including an apparent inherent paradox identified by its creator. Then I present the evolutionary argument for the theoretical stance taken here, and derive the (simple) principles on which my proposal is based from it. Next, turning to implementation, I summarise earlier modelling work, explain its connection with the current proposal, and describe what is necessary to extend it into the model proposed here.

The technical contributions of the paper are a variant notion of AI Agent, based on prediction from sense data, rather than on sensing, and a mechanism for deployment of that agent in a particular kind of reasoning system. The key philosophical contribution is the fact that, once this mechanism is deployed, the kind of creativity that is addressed here, inspiration, is explained *within* the basic reasoning, and needs no further explanation.

Methodology

To overcome the methodological problem introduced above, my approach is to attempt to replicate an existing creative process. The only existing creative process ready available for inspection is that of humans; these have the built-in advantage, mostly, of being able to explain what (they thought) they did, and elegant paradigms exist to empirically deconstruct that majority of aspects of human behaviour of which introspective reports are unreliable. I therefore aim to apply cognitive modelling theory and technology to human creative process, and then to evaluate the success of the enterprise with respect not only to the outputs of the computational systems produced, but to compare the various aspects of their operation with human creators. While this approach solves only part of the general problem of computational creativity, it is an area where refutable hypotheses can be made, and so demonstrable progress in a research programme (Lakatos, 1970) may take place.

For this attempt to succeed in a scientific sense, before one even considers the artefacts that the replicant creative system may produce, the theory and its associated computational

system must conform to at least the following constraints, to be said to model *human creative cognition*.

1. **Falsifiability** The system must not behave in ways which are arguably or demonstrably different from human creators while it is operating. Since we cannot, currently, know how human creators create, this is the strongest falsifiability constraint that can be applied.
2. **Evolutionary context** There must be an account of the evolutionary advantage conferred by the mechanisms proposed, a corresponding order of development, and an analysis of their appearance in successive species over evolutionary time. This account cannot be verifiable, but the lack of one leaves the biological development of the proposed solution unavailable to scientific scrutiny.
3. **Learning capability** The system must be capable of learning its creative domain. Learning should be appropriate to the domain: for example, in music, perceptual aspects should be *implicit*—that is, teaching or supervision should not be required; however, in some domains, such as mathematics, minimal supervision is evidently unavoidable, because of the need to know the meaning of symbols, to give semantics to what is being learned[1].
4. **Production capability** The system must be able to produce artefacts that are demonstrably within its creative domain, whether or not they are of quality comparable with a human creator's output. While the judgement of whether an artefact is or is not a particular kind of thing is subjective, it is not as difficult as the subjectivity of quality. For the purposes of experiment, restricted domains with clear tests must be set up, using appropriate theory from the corresponding human-creative domain.
5. **Reflection** The system must be capable of reflecting on its behaviour, modifying it, and explaining it—where necessary via indirect indicators such as those used for understanding the behaviour of humans.

In this paper, I present a hypothetical, but partly implemented, computational model of a particular kind of human creativity, and suggest that it conforms to criteria 3–4, and partly to criterion 2, though further research is required to provide more evidence against criterion 1. Criterion 5, Reflection, is conferred by location of the model within Baars' (1988) Global Workspace Theory, whose focus is consciousness; so it falls beyond the scope of the present proposal.

Background

Creativity: Inspiration and Reasoning

Wiggins (2012) introduces a distinction between two kinds of creativity: on one hand, *inspiration* and, on the other, *creative reasoning*. Respectively, these terms are intended to distinguish what appears spontaneously in consciousness—the "Aha!" moment that Wallas (1926) suggests follows "incubation"—from what is produced by the deliberate application of creative method. The spectrum between the two allows us to make distinctions between conscious creation in the deliberate planning of a formalist composer, the semi-spontaneous but cooperative and partly planned creation of the jazz improviser in a trio, and entirely spontaneous singing in the shower. Note that a non-polar position on this spectrum necessarily entails a *combination* of explicit technique and implicit imagination: there is not a smooth transition in kind between the two, but rather a *mixture* containing some of each in varying proportion.

Having made this point, I reserve creative reasoning for future work, not least because it entails that we address consciousness, which is difficult, but also because Baars' theory already provides a framework in which it may be considered, *given* a mechanism for inspiration. This is not to dismiss the deliberate end of the scale, nor to suggest that it does not exist, but merely to focus the current work on a separable aspect of the complex.

Global Workspace Theory

Bernard Baars (1988) introduces a theory of conscious cognition called the Global Workspace Theory. There is not space to describe this wide-ranging and elegant theory here, so I summarise the relevant important points. The theory posits a framework within which consciousness can take place, based around a multi-agent architecture (Minsky, 1985) communicating via something like an AI blackboard system (Corkill, 1991), but with particular constraints, which I outline below. The approach taken is to avoid Chalmers' "hard" question of "what is conscious?" (Chalmers, 1996) and instead ask "what is it conscious of, and how?" This is especially appropriate in cases such as the current paper, where consciousness is not the central issue, but presentation of information to it is.

Baars casts the non-conscious mind as a large collection of expert generators (not unlike the multiple experts in Minsky's *Society of Mind*, 1985), performing tasks by applying algorithms to data in massive parallel, *compete for access* to a Global Workspace via which (and only via which) information may be exchanged; crucially, information must cross a notional threshold of "importance" before it is allowed access. The Global Workspace is always visible to all generators, and contains the information of which the organism is conscious at any given time. However, it is capable of containing only one "thing" at a time, though the scope of what that "thing" might be is variable. The Global Workspace is highly contextualised, and meaning contained therein is context sensitive and structured; contexts can contain goals, desires, etc., of the kind familiar from broader AI. Aside from further discussion of the "threshold" idea, below, this is all that is needed to understand the purpose of the competition mechanism proposed here. Baars mentions the possibility of creativity within this framework in passing, implicitly equating entry of a generator's output into consciousness with the "Aha!" moment (Wallas, 1926). However, he does not develop this idea further beyond noting that a process of refinement may be implemented as cycling of information into the Workspace and out again; that process

[1] To ask the system to learn the semantics of the symbols to which it is exposed from context is not, in principle, unreasonable, as there is every evidence that humans do so. However, to require the system to do so when the scientific research focus is creativity seems unnecessarily difficult.

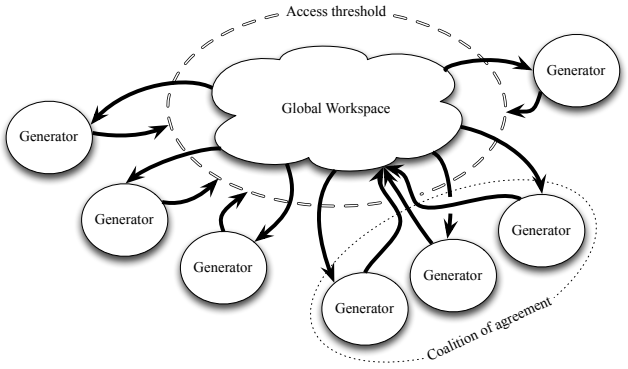


Figure 1: Illustration of Baars' Threshold Paradox. Generators generate, but need a means of recruiting support for their outputs. Individuals cannot break in; they must recruit coalitions, as shown. The only way to do so is via the Global Workspace, but before they can do so, they need the support they are trying to recruit, and therein lies the paradox.

may be equivalent to my creative reasoning. To the best of my knowledge, however, creativity in the Global Workspace has not been addressed elsewhere in the related literature.

In the later developments of the theory, Baars proposes that information integration may take place in stages, via something that one might (but he does not) call local workspaces, that integrate information step by step in a sequence, rather than all in one go as it arrives in the Global Workspace. This information integration approach has been extended by Tononi and Edelman (1998), who propose information-theoretic measures of information integration as a measure of consciousness of an information-processing mechanism. Baars has embraced the information-theoretic stance, too, and the three authors have jointly proposed to begin implementing a conscious machine (Edelman, Gally, and Baars, 2011) based on their ideas. The current work may contribute to this endeavour, though probably at a level more abstract from neurophysiology than these authors intend.

The Threshold Paradox

Baars (1988, pp. 98–99) addresses what he acknowledges is a problem for his theory. He proposes a threshold for input access to the Global Workspace, crossing of which is thought of in terms of recruiting sufficient generators to produce information that is somehow coordinated, or synchronised between them: it must be metaphorically "loud" enough to be "audible" in the Workspace. However, in terms of the Global Workspace alone, there is no means of doing this: generators can only be coordinated (whatever that means) via the Global Workspace, and so the generators are faced with the beginning artist's dilemma: you have to be famous to show your work, but you have to show your work to become famous. This form of the Workspace is illustrated in Figure 1. Baars presents two possible solutions to the paradox, which is the motivation of the current paper, but both are presented somewhat half-heartedly, leaving a gap in the theory. Here, I present a possible solution, in terms of the evolutionary argument required by my criterion 2, above.

Perception, Anticipation, and Evolution

Reaction vs. Anticipation

I now present a mechanism for managing the competition between generators in Baars' system. This mechanism may be implemented either directly or indirectly (that is, by means of some other effect)—the difference is immaterial at the current theoretical level. The key distinctions are a) between the information content and entropy (defined below) of various stimuli; and b) between organisms that react and organisms that anticipate. The design of this mechanism is motivated by evolutionary thinking: that is, by consideration of the evolutionary advantage conferred by the resulting behaviours, in humans and other animals. Thus, the evolutionary argument presented here is part of the design, not merely an example.

Russell and Norvig (1995), in their well-known AI text book, define an AI agent (of which an AI *creative* agent is presumably an instance) as a program or robot with a behaviour cycle that consists of perceiving the world and then acting on the perceptions. It seems not unreasonable to present this as a model of lower organisms, such as insects, which seem to do nothing more than react to environmental conditions, coping poorly when their evolved reactive program is interrupted. However, to model higher cognitive development, one can propose a more predictive system, in which an organism is predicting continually, from a learned model of previous sensory data, what is likely to come next, and *comparing* this with current sensory input. Doing so gives a simple but effective mechanism for spotting what is unusual, what, therefore, constitutes a potential new opportunity or threat, and what deserves cognitive resource, or *attention*. In the simplest case, the anticipatory agent can in principle avoid a threat *before* it becomes apparent, while the reactive one has to experience the threat in order to respond.

The consequence of sequence: managing uncertainty with expectation

The most important feature of an autonomous agent is not, as sometimes supposed in AI, that it is able to identify or categorise a situation from available data. What gives it the edge is that it can, in some sense, *imagine* what is to come next, and react, or perhaps *pre*act, in advance. Of course, the word "imagine" is loaded, and suggests the involvement of consciousness and even volition; I use it here deliberately to draw attention to the point that consciousness need not be implicated in this process, which can be described in completely mechanistic terms, of *prediction* alone.

In order to predict usefully in a changing world, it is necessary for an organism to learn. It must be able to learn not just categorisations (to understand what something is), but also associations (to associate co-occurrence of events with reward or threat), and, crucially here, sequence.

However, a simple statistical learning mechanism is not subtle enough (Huron, 2006). Since evolutionary success entails that an organism breeds, a mechanism which allows

that organism to learn only from potentially fatal consequences does not suffice: if the organism dies as the result of an experience, it does not benefit from the experience (or, at least, not for long). An effective strategy here lies at a meta-level with respect to a learned body of experience: if an organism is aware that it is in circumstances that it cannot predict reliably, it can behave more cautiously, its metabolism can be aroused to prepare for flight, and it can devote more attention than normal to its surroundings; thus, the effective strategy is also *a*ffective. Huron convincingly argues that this process is exapted to produce part of the aesthetic effect of music; however, for the purposes of the current section, the mere *ad*aptation suffices: self-evidently, there is a mechanism that allows uncertainty to affect behaviour in humans and other animals, and that mechanism does not rely on explicit reasoning: indeed, the converse is the case: we feel nervous in uncertain situations, and the feeling serves to make us wonder why, as well as to heighten our attention to appropriate sensory inputs and to prepare for flight. This mechanism, and the associated affective response, is not the same as fear, but can lead there *in extremis*.

Finally, any kind of learning of this nature is inadequate unless it includes *generalisation*. It is necessary to be able to generalise from both co-occurrence and sequence that similar consequences arise from similar events, encounters, etc. Without this, mere tension cannot lead to the fear that is appropriate at the sight of the bared fangs of a previously-unexperienced large animal. This accords with proposals such as that of Gärdenfors (2000), that perceptual learning systems are motivated by the need to understand similarities and differences between perceived entities in the world, and to place observations at the appropriate point between previously experienced referents.

Prediction, Prioritisation and Selection

Given a model of the world, suitably subcategorised into types, situations, etc., one can imagine a set of generators using the model with recent and current perceptual inputs matched against precursors of sequential associations, making predictions, on a basis that is stochastic, and conditioned by the model. Making such predictions quickly, one at a time, would be valuable, but, given the nature of brains, slow, multiple predictions, in parallel, are a more likely candidate for evolutionary success, and the more the better—as in Baars' proposal. But this begs a question: arbitrarily many predictions occurring simultaneously will be an impossible, incomprehensible babble, so how will useful candidates for prediction be selected? Baars' solution is the problematic threshold, described above.

Another shortcoming of the Global Workspace Theory is unclarity about precisely what the notion of generators "recruiting" one another means. The *effect* is something like an additive weight: the more generators that are "recruited", the greater the impact of their output. In my proposal, we will avoid answering this question, by approximating the effect of the recruitment, rather more simply. I return to this below; in the argument that follows, I will use the analogy of sound volume to refer to this property: "loud" predictions come from many generators, "quiet" ones do not.

My proposal here is based on statistical, frequentist notions of learning, and so my reasoning is couched in terms of statistical models; however, I do not think that the reasoning is in principle exclusive to such models, and it should not be supposed that the proposal is *restricted* in this way. In this view of the system surrounding the Global Workspace, there are many independent subsystems, which are making multiple predictions by biased sampling from a predictive statistical model of (assumedly) reasonable quality. It is also appropriate to assume imperfect models: each of these generating subsystems will have a fragmentary, partial view of its world and its predictions, as to model everything all the time in massive-parallel would be prohibitively expensive. It follows from the use of frequentist models that the more expected occurrences are the more likely ones to be predicted: the commonest predictions will be the most expected ones. This means there are relatively "loud" groups of contributions, reinforcing each other. Conversely, extremely unlikely predictions will be proposed by only a very small number of generators, and as such will never be "audible".

In a model of prediction and action based solely on this frequentist principle, an organism will tend do the commonest thing, even when inappropriate, and therefore will be doomed to failure: it will not "imagine" unlikely and surprising situations, and will not therefore prepare itself against necessary eventualities. To see this, consider a territorial animal, on patrol, and let it be a high enough species to learn its reactions. Today, our animal senses the things it usually senses, and the vast majority of things in the world today are the same as they were the last time it passed this way. One tiny difference is a scent that it does not recognise, that it has not experienced before. Since this difference is small in comparison to the rest of the data in the world, and it has not been experienced before, in purely frequentist terms, it will be ignored: it is unlikely, and it has no known consequences and determines little or no probability mass.

In Baar's theory, the pure frequentist approach, where the most likely outcome is chosen, corresponds with multiple generators in coalition generating that outcome. The likelihood of each generator predicting an outcome is proportional to the "volume" of that outcome across the set of generators. Therefore, we can neatly draw a veil over the mechanistic gap left by Baars' idea of coalition formation, and simply use the likelihood of the outcome, p, to model its outcome.

In reality, though, we know well that to carry on as normal will not be the reaction of an animal in these circumstances: it will experience Huron's proposed affective response, described above. Therefore, it is necessary to hypothesise a mechanism to cause that response. In our current simple context of abstracted statistical modelling, the obvious choice for such a mechanism is the notion of *entropy*, as formalised by Shannon (1948). MacKay (2003) makes a distinction between *information content*, h, which is defined as an estimate of the number of bits required to describe an event, e, given a context, c, or its *unexpectedness*:

$$h(e \mid c) = -\log_2 p(e \mid c),$$

and *entropy*, H, which is defined as an estimate of the *uncertainty* inherent in the distribution of the set of events $\mathcal{E}$

from which that e might be selected, given the context, c:

$$H(c) = \sum_{e \in \mathcal{E}} p(e \mid c)\, h(e \mid c) = -\sum_{e \in \mathcal{E}} p(e \mid c)\, \log_2 p(e \mid c).$$

H is maximised when all outcomes are equally likely, and minimised when a single outcome is certain. Both h and H are useful to our hypothetical animal.

First, consider h_t, the unexpectedness of a partial model of the actual on-going experience in a particular state, t. If the experience is likely (in particular, if it is *readily predictable* from what has gone before), it is not unexpected, and therefore h_t is low; if it is unlikely, it is unexpected, and so h_t is high. An experience such as encountering a *new* scent is maximally unlikely, in frequentist terms. To model this, I propose that individual generators are sensitive to their own h_t value, and decrease their notional "volume" when it is low. Thus, the likelihood of models of the experience in which the new scent is included being heard in the theatre is positively related (possibly in a non-trivial way) to its unexpectedness. I call this the *recognition-h* case. It may explain why unexpected things are noticed.

Now, consider, h_{t+1}, the unexpectedness of a predicted situation. It is maximally unlikely that a *prediction* will be made including a scent that has not been encountered before, and, as above, we would therefore expect h_{t+1} to be very high, causing alarm. Excess of such predictions, or repeated occurrence of a single one, would lead to a state of constant anxiety[2]. I call this the *prediction-h* case. It may explain why surprising predictions are more likely to draw attention than prosaic ones.

Of course, in a simplistic frequentist account, predictions introducing new percepts or concepts cannot arise, because they entail the creation of new symbols. This is why it is necessary to include generalisation and/or interpolation in the theory (see above). Gärdenfors (2000) presents a theory that explicates the symbolic representations more commonly used in statistical AI modelling in terms of an underlying, sometimes continuous, geometrical layer, and, at least at perceptual levels, places cognitive semantics at the centre of mind. In particular, an outline mechanism is supplied whereby previously unencountered stimuli may be assigned first non-symbolic, and then symbolic, representations. It is important to understand that the semantics in these theories are internal to the organism experiencing them, and have no *definition* in terms of the external word; rather they have external associations, which can serve to allow intersubjective meaning, but they themselves are ineffable.

The problem of over-active prediction-h is mitigated by the mechanism supplied above, in which prediction is probabilistic and (broadly) additive across predictors, modelled by p. There are two opposing forces here, one of which changes inversely relative to the other, and because they are co-occurrent, their effects should (broadly) multiply. Therefore, the overall outcome audible in the global workspace can be estimated by multiplying the probability, p of an event (which estimates the likely number of generators predicting it) by h (which estimates the volume at which they are predicting). The resulting likelihood is illustrated by the unit-free diagram in Figure 2. This creates a bias away from predictions which are either very likely or very unexpected, reducing the power of the very unlikely or the very obvious to attract attention. This may explain why unlikely possibilities do not prevent action by overwhelming the acting organism with choice.

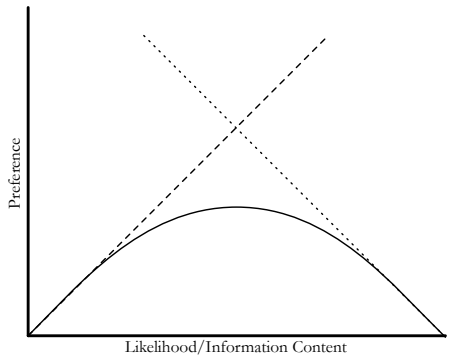


Figure 2: Illustration of the interaction between likelihood and unexpectedness. The overall likelihood (solid) is formed by the multiplication of two monotonic functions: the unexpectedness of a generated item (dashed) and the number of generators likely to agree on it, according to its likelihood (dotted).

It is important to see the difference between recognition-h and prediction-h in the context of the Global Workspace. I propose that generators may generate structures of either kind, and that the two will be in competition for the resource of attention. Thus, clear and present danger or benefit will outweigh predicted likelihoods, because the distribution of *potential* predictions is over a much wider range of possibilities than that over *actual* perceptions, and therefore, comparatively, probability mass is spread more thinly. Conversely, for example, likely but unexpected predicted benefits can outweigh less seriously dangerous present circumstances—thus, prioritising an unusual positive opportunity can be mechanistically explained as an emergent behaviour.

Given that there are now two kinds of generator (or generator output), I must propose a means of distinguishing between them, though this is not a key focus of the current argument. Without such a means, consciousness would be unable to distinguish between the perceived world and the predicted one[3].

Sensing Certainty

Shannon's H is interesting here in a different way. As explained above, H is the expected value of the information

[2]Indeed, some humans who suffer from anxiety, in the clinical sense, report intrusive, repetitive thoughts predicting problems or worries of one sort or another, the anxiety being aroused by fear of what *might* happen. Their situation would be explicable in terms of a breakdown of this mechanism.

[3]Coupled with a deficit in suppression of less likely outcomes, as above, this situation might lead to some root symptoms of schizophrenia: hallucinations, delusions and cognitive disorganisation.

content of a given distribution, so it is different in kind from h, which deals with individual situations, actual or predicted. It is best characterised as the uncertainty inherent in a distribution, and, indeed, a uniform distribution always gives the maximum entropy for a given alphabet size. Unlike h, H really only has meaning in the predictive context: once one knows which possibility of a range is the right one, only information content is really relevant. However, in the predictive context, a predicted outcome of which one is certain is much more useful than one of which one is unconfident: H measures this difference.

I propose, therefore, that, in the predictive generators, higher H also predicts lower volume, so that less certain generated outputs are de-emphasised. This, then, I call *prediction-H*. It may explain how it is possible to *feel certain* about intuitions (as opposed to be convinced of reasoned argument). It also prevents the Global Workspace from being flooded out with predicted information that is not strongly supported, allowing the important material to shine through. A particularly interesting point is this: should a generator make an unlikely prediction, that has sufficient prediction-h to be "audible", in the absence of other explanations, that prediction will have low prediction-H, and so will not be suppressed by this final mechanism. Increasing the range of possibilities over which the distribution holds, even if they are unlikely, increases prediction-H and thus decreases certainty. Under this régime an organism that has less experience is more likely to admit unlikely predictions to consciousness; this might be taken to account for the tendency, for example, of children to be more affected by imagined fears than adults.

No straightforward diagram can be drawn of the effect of prediction-H on the overall likelihood of a generator taking over the Global Workspace, because the numbers depend heavily on the multidimensional distributions from which the various Hs are calculated.

This leaves us with a "volume" value for each generator, T, which is estimated by the following, for either kind of h, above:

$$T = \frac{p \times h}{H}.$$

I propose that, at any given moment, this "volume" value is used in deciding which of the range of possible inputs, derived from matching sensory input to statistical models in memory, enters the Global Workspace. This is illustrated in Figure 3.

Generation, Creativity and Intuition

In the previous section, I outlined a simple, comparative mechanism by which statistically likely and information-theoretically rich structures can emerge from a multi-agent system furnished with high-quality models of a domain of knowledge. With such a mechanism, the Threshold Paradox disappears. I should also note that it is possible that such a mechanism is one of Baars' own proposals; however, if so, it is not clearly specified as such. The remaining question is then: how does this mechanism for choosing access to consciousness help to simulate creativity?

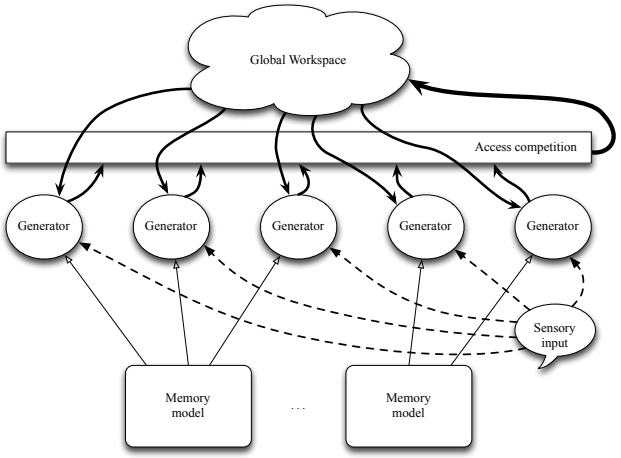


Figure 3: Schematic diagram of my proposal for the Global Workspace. In this version, there is no need for a threshold of access. Instead, the generators compete, one against another, and probability and information content determine the winner.

Perhaps surprisingly, an answer may be found in the writing of Wolfgang Amadeus Mozart (quoted by Holmes, 2009, pp. 317–8):

> When I am, as it were, completely myself, entirely alone, and of good cheer – say traveling in a carriage, or walking after a good meal, or during the night when I cannot sleep; it is on such occasions that my ideas flow best and most abundantly. Whence and how they come, I know not; nor can I force them. Those ideas that please me I retain in memory, and am accustomed, as I have been told, to hum them to myself.

One might paraphrase the opening sentence here as "when I am not being bothered, and when I have no worries and no particular goals", which in turn means "when I have no distractions" or "when I have no information-rich input to consciousness from outside or within". This accords with a situation when the Global Workspace is occupied only by weakly-informative ephemera, and when generators are receiving little or no external stimulus.

Recall now my earlier proposal that effective animals will base their actions not only on received stimuli, but on the results of the comparison of received stimuli with predictions about the current state of the world made from the previous state(s). Suppose that those generators continue to generate, even when there is very little informative input. Given the appropriate knowledge and tendencies in a particular individual (music—and, by all accounts, scatology—in Mozart), generators will begin to freewheel, within the same statistical framework as above, but lacking the statistical prior of a particular stimulus. The outputs, one might expect, will be rather more diffuse and perhaps less highly rated than when directly stimulated, but this is not a brake on their progress towards the Global Workspace, because *there is little or no competition*. At this point, the diagram in Figure 2 becomes recognisable as the Wundt curve, as it defines a

sweet spot of balance between dullness and over-complexity in information-theoretic terms.

Mozart, above, describes a particular kind of musical approach, where one essentially enters a quiescent state, in order deliberately to allow Baars' generators to freewheel; I find that the same method works for writing. But this is only one case of many. For example, the mechanism above also accounts for why hearing a musical phrase, or even a non-musical pitch sequence, may give rise to new musical phrase: the percept conditions the generators in a particular way, and so affects the likely outcomes, which are generated all the time. The ones with the right statistical properties make it into the Global Workspace, and so can be further elaborated.

Note that this mechanism can apply to any statistical model available to the generators, so it need not be restricted to music (as it is in the system components summarised in the next section). In principle, the same idea can work with any model from which statistical likelihoods can be computed. This means, for example, that it can account for the generation of sentences, and therefore possibly internal speech. If internal speech is equated with essential thought, as commonly, then the current approach can account for general creative thought and for the emergence of particular thoughts into consciousness as intuition. It can also, via prediction-H, account for the (sometimes inappropriate) feeling of certainty associated with thoughts and intuitions.

Thus, I suggest that "Threshold Paradox" as a name for this issue needs to be reinterpreted. The paradox is not in the nature of the threshold, but in the formulation of the Global Workspace as requiring one. The current theory reformulates entry to the Workspace as purely competitive, without a particular boundary, so, one might say, the paradox arose from the assumption that the Threshold exists.

What is more, in the present theory, there is no longer any need to search for an explanation of creativity as a distinct phenomenon. In my approach, non-conscious creativity is happening all the time as a result of on-going anticipation in all sensory (and other) modalities. When conditions are right, this essential survival mechanism is not so much *exapted* for creativity, but gives rise to creativity as a side effect.

Towards a Creative System

To ground this theory in a technical base, I now summarise research that has already been conducted towards building a system of the kind proposed here, in the domain of musical creativity. Pearce and Wiggins (2006)[4] describe a statistical model of musical learning, based on, but extending, statistical language learning methods. Wiggins (2011) has shown that the extensions to the musical model can also benefit language models. Pearce and Wiggins (2007) showed how the model could generate entire musical melodies, though the requirements of the current proposal are less stringent, as fragmentary musical ideas are all that is required: in this case short sequences of notes that might be consciously assembled into melodies. Most importantly, the model has been used to demonstrate that high information content corresponds with increased beta-band synchrony in human listeners (Pearce et al., 2010), providing at least circumstantial evidence that cognitive resource (i.e., attention) does indeed follow information content, which would accord with that information's entry into the Global Workspace when the circumstances, as described above, are right.

Ponsford, Wiggins, and Mellish (1999), Whorley, Pearce, and Wiggins (2008), Whorley, Wiggins, and Pearce (2007) and Whorley et al. (2010) have presented more complex models for dealing with deeper aspects of music than melody.

A crucial piece of evidence for the model of creativity proposed above is embedded in the workings of Pearce's perceptual model—recalling that perception and prediction are closely linked in the view of the world presented here. There are two sub-models, both of which contain multiple predictors. The distributions output by the two sub-models are combined multiplicatively, with weightings derived their relative information entropy. The distributions output by the multiple predictors *within* each of the two sub-models are combined in the same way. Other configurations (for example, a one-stage combination of all of the distributions, instead of this two-stage combination) produce a less successful model of human behaviour. This system matches exactly against the multi-stage version of Baars' Global Work Space, described above, coupled with my proposal for a competition mechanism based on information content and entropy.

There is still substantial work to be done on this model before the simulation of creativity can be claimed. The next threshold to cross is not a paradox, but the engineering task of implementing the integrated multiple generators in the model described above, to test out the this particular approach to competitive generation in the Global Workspace.

[4] A fuller presentation of the modelling work published up to 2007 is given in Pearce's (2005) PhD thesis.

Acknowledgments

I gratefully acknowledge the contribution of the ISMS group, most particularly Roger Dean, Ollie Bown, Jamie Forth and Marcus Pearce, of Joydeep Bhattacharya, and of three anonymous referees, to the thinking presented here. Funding was provided by EPSRC Research Grant EP/H01294X/2, "Information and neural dynamics in the perception of musical structure".

References

Baars, B. J. 1988. *A cognitive theory of consciousness*. Cambridge University Press.

Chalmers, D. J. 1996. *The Conscious Mind: in search of a fundamental theory*. OUP.

Colton, S. 2009. Seven catchy phrases for computational creativity research: A position paper. In *Proceedings of the Dagstuhl Workshop on Computational Creativity*. Germany: Schloss Dagshtul.

Corkill, D. D. 1991. Blackboard systems. *AI Expert* 6(9):40–47.

Edelman, G. M.; Gally, J. A.; and Baars, B. J. 2011. Biology of consciousness. *Frontiers in Psychology* 2.

Gärdenfors, P. 2000. *Conceptual Spaces: the geometry of thought*. Cambridge, MA: MIT Press.

Holmes, E. 2009. *The Life of Mozart: Including his Correspondence*. Cambridge Library Collection. Cambridge University Press.

Huron, D. 2006. *Sweet Anticipation: Music and the Psychology of Expectation*. Bradford Books. Cambridge, MA: MIT Press.

Lakatos, I. 1970. Falsification and the methodology of scientific research programmes. In Lakatos, I., and Musgrave, A., eds., *Criticism and the Growth of Knowledge*. Cambridge, UK: Cambridge University Press. 91–196.

MacKay, D. J. C. 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge, UK: Cambridge University Press.

Minsky, M. 1985. *The Society of Mind*. New York, NY: Simon and Schuster Inc.

Moffat, D., and Kelly, M. 2006. An investigation into people's bias against computational creativity in music compositio. In *Proceeings of the International Joint Workshop on Computational Creativity*.

Pearce, M. T., and Wiggins, G. A. 2006. Expectation in melody: The influence of context and learning. *Music Perception* 23(5):377–405.

Pearce, M. T., and Wiggins, G. A. 2007. Evaluating cognitive models of musical composition. In Cardoso, A., and Wiggins, G. A., eds., *Proceedings of the 4th International Joint Workshop on Computational Creativity*, 73–80. London: Goldsmiths, University of London.

Pearce, M. T.; Herrojo Ruiz, M.; Kapasi, S.; Wiggins, G. A.; and Bhattacharya, J. 2010. Unsupervised statistical learning underpins computational, behavioural and neural manifestations of musical expectation. *NeuroImage* 50(1):303–314.

Pearce, M. T. 2005. *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition*. Ph.D. Dissertation, Department of Computing, City University, London, UK.

Ponsford, D.; Wiggins, G. A.; and Mellish, C. 1999. Statistical learning of harmonic movement. *Journal of New Music Research* 28(2):150–177.

Ritchie, G. 2001. Assessing creativity. In *Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in the Arts and Sciences*, 3–11. Brighton, UK: SSAISB.

Russell, S., and Norvig, P. 1995. *Artificial Intelligence – a modern approach*. New Jersey: Prentice Hall.

Shannon, C. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27:379–423, 623–56.

Tononi, G., and Edelman, G. M. 1998. Consciousness and complexity. *Science* 282(5395):1846–1851.

Wallas, G. 1926. *The Art of Thought*. New York: Harcourt Brace.

Whorley, R.; Wiggins, G. A.; Rhodes, C.; and Pearce, M. 2010. Development of techniques for the computational modelling of harmony. In Ventura, et al., eds., *Proceedings of the First International Conference on Computational Creativity*.

Whorley, R. P.; Pearce, M. T.; and Wiggins, G. A. 2008. Computational modelling of the cognition of harmonic movement. In *Proceedings of the 10th International Conference on Music Perception and Cognition*.

Whorley, R. P.; Wiggins, G. A.; and Pearce, M. T. 2007. Systematic evaluation and improvement of statistical models of harmony. In A. Cardoso, and G. A. Wiggins., eds., *Proceedings of the 4th International Joint Workshop on Computational Creativity*, 81–88.

Wiggins, G. A. 2011. "I let the music speak": cross-domain application of a cognitive model of musical learning. In Rebuschat, P., and Williams, J., eds., *Statistical Learning and Language Acquisition*. Amsterdam, NL: Mouton De Gruyter.

Wiggins, G. A. 2012. Defining inspiration? Modelling non-conscious creative process. In Collins, D., ed., *The Act of Musical Composition – Studies in the Creative Process*. Aldershot, UK: Ashgate.

Brainstorming in Solitude and Teams: A Computational Study of Group Influence

Ricardo Sosa
Singapore University of Technology and Design
20 Dover Drive, Singapore 138682
ricardo_sosa@sutd.edu.sg

John S. Gero
Krasnow Institute for Advanced Study and
Volgenau School of Engineering
George Mason University
john@johngero.com

Abstract

Early studies of creative ideation showed that individuals brainstorming in isolation tend to generate more and better ideas than groups. But recent studies depict a more complex picture, reinforcing the need to better understand individual and group ideation. Studying group influence is one way to address the complex interplay between ideas in different brainstorming scenarios. We define group influence as the degree to which individuals are influenced by ideas coming from other team members. This paper presents results from a multi-agent simulation of the role of group influence in brainstorming groups. The results from the simulations indicate that the findings from previous laboratory studies tend to be misinterpreted, and that both isolation and teamwork present opportunities and challenges for creativity.

Introduction

Is it better to generate new ideas in solitude or in teams? Creativity research has shown that this distinction is not trivial. Early studies showed that individuals working in privacy tend to generate superior results along three criteria: total number of ideas, number of unique ideas, and quality of ideas [1]. But a more complex picture is portrayed by subsequent studies, reinforcing the need to better understand the interplay between individual and group ideation as well as the importance of facilitation dynamics [2].

The term 'brainstorming' refers to the method of problem solving based on timed sessions where participants are instructed to address a problem by freely generating a large number of ideas irrespective of their apparent value [3]. The aim of brainstorming sessions is to generate as many different alternative solutions to a given problem as possible. Whilst many variants of brainstorming have been proposed, the basic premises are: a) to maximize the number and the originality of ideas, b) to combine or improve ideas suggested, and c) to avoid critical evaluation of ideas [4].

Individual brainstorming consists of engaging subjects in idea generation sessions isolated from others. Group or team brainstorming refers to the more typical scenario where individuals interact to generate and evaluate possible solutions to a common problem. Following the literature, we use the term *nominal group* to refer to the former and *interactive group* to refer to the latter condition [2].

Recent studies of idea fluency in brainstorming show that nominal groups outperform non-facilitated interactive groups both in gross and net fluency of ideas; but are considerably outperformed by facilitated interactive groups [2]. As with other factors related to team dynamics, such as diversity and leadership, group influence as a construct and its effects on creative ideation are yet to be fully understood. This is a relevant topic in the still incipient research stream on multi-level approaches to team creativity [5].

The general process by which individuals in isolation consistently surpass group creativity has been explained as 'ideational productivity loss' and appears to have a series of likely cognitive and group-level causes [6]. Cognitive factors that may interfere with ideational productivity include production blocking, interruptions, forgetting ideas, and distraction by task-irrelevant processes. A higher cognitive load is also often cited as a source of ideational loss, typically caused by attending to other's ideas.

Group factors that may account for productivity loss include team structure and diversity, turn-taking, awareness of public evaluation, disposition to converge with others' judgments, lower motivation due to shared responsibility, and a tendency to free-ride [7]. Multi-level approaches are required to understand, for instance, what is the appropriate degree of accessibility to others' ideas when brainstorming in teams in order to ensure that individuals are able to both build upon their own ideas as well as upon the ideas of their teammates.

Teamwork in creativity enables the important process of sharing ideas; however this freedom may have two different effects on creative ideation: one possibility is that teammates generate a wide range of diverging ideas thus obstructing the connection and refinement of coherent

'trains of thought'. The noise generated in this imagined scenario would more likely produce incomplete and incompatible ideas of low quality not to mention dissatisfaction from the participants. A second possibility is that teammates rapidly converge in agreement around one or just a few dominant ideas without exploring other alternatives.

Group influence can be one way to address this interplay between ideas in brainstorming. We define group influence in this paper as the degree to which individuals are influenced by ideas coming from other team members. Here, group influence is a group-level rather than an individual construct. Groups with high influence levels are those where all ideas by all participants are always available to every group member. Groups with low influence levels are those where individuals are only exposed to their own ideas. Between these extremes, group influence indicates the ratio of ideas available to brainstormers.

In this paper we present results from a multi-agent simulation of the role of group influence in brainstorming groups. Our aim here is to model the interactions between agents engaged in a simple task of divergent reasoning in order to inspect the beneficial and detrimental effects of different team structures in idea generation. In defining this model, we follow the distinctions between ideas, agents and societal factors of the IAS framework for the computational modeling of creativity and innovation as explained below [5]. The rest of this paper is organized as follows: the next section presents precedent work on the computational modeling of group brainstorming, the following section introduces our own modeling approach to group influence in brainstorming, then the simulations results are presented and the paper concludes with a discussion of the results and their implications in computational creativity research.

Models of Group Brainstorming

This paper presents an approach to the study of creativity using computational social science [7] in order to inspect the mechanisms behind the apparent paradox of ideational productivity loss in brainstorming groups. Computational social science utilizes multi-agent simulations that are useful to explore hypotheses, test assumptions and understand fundamental issues in complex social systems. These systems are also useful to generate predictions for future laboratory experiments or case studies.

Semantic and Social Models

Iyer et al [8] propose a connectionist framework of idea generation in order to inspect experimental data from laboratory studies on ideation and idea priming. In particular they explore the interaction between 'irrelevant primes' and context familiarity; irrelevant cues are defined as sets of ideas of which only a fraction are related to the task at hand, while context familiarity is given by the pre-existing classification of ideas defined in the system.

With this model, the researchers emulate the laboratory results and provide hypotheses as to why even irrelevant primes can increase idea quality and fluency. By manipulating the degree of familiarity between contexts, they show that when irrelevant primes are used between two completely unfamiliar contexts, there is no benefit, whilst irrelevant priming is useful only when partial information about semantic relationships is shared between search contexts. In this vein, the authors suggest future experimental studies on the creative capacity to create 'short-cut linkages' between features, concepts or semantic categories that are typically not related.

In an extension of this work, Paulus et al propose an approach to modeling group creativity by vertically integrating neural and social networks [9]. They define agents as simplified versions of the connectionist model described above, and account for individual differences in semantic contexts, idea association, domains, cognitive strategies and responses to cues. Through what they define as a *parameterized interaction protocol* (PIP), their proposed model accounts for turn-taking between agents and, more relevant to our approach, the accessibility of ideas by either the entire group or a selected few. With this model still under development and testing, the authors aim to address a range of research questions, including the efficiency of certain interaction structures and scheduling protocols for group ideation.

Group Influence in a Design Task

From the perspective of computational social science, creative systems are modeled by multiple generators and evaluators of ideas linked in a social system. In such systems, creativity is explained as an emergent outcome, i.e. a global effect that 'grows' from simple local interactions [10]. The model presented here is defined using the channels of interactions specified in the IAS framework (ideas, agents, society) [10]. Agents (*A*) engage in a simple designing task that constitutes the agent-idea channel (*Ai*) where the resulting designs belong to the set of Ideas (*I*); social structures (*S*) determine the availability of ideas (*Si*); ideas are used by agents (*Ia*) to build design concepts (*Aa*) that are further applied in the design of new ideas (*Ai'*).

In this model, *Ai* is implemented as a shape search process starting from an initial set of polygons and affine transformations, *I* is the set of final shape representations produced by the agents, *S* is the arrangement of agents in groups, *Si* is the experimental variable of group influence, *Ia* is a transmission mechanism of ideas to agents, and *Aa* is modeled as an inference process of design concepts. At the moment, this model is limited to only four of the nine channels of interaction in the IAS framework, namely: *Ai, Si, Ia* and *Aa*. In the future, we plan to integrate and examine more IAS processes in this model including leadership styles (*As*), compliance to group majority (*Sa*), group agreement to adjust idea influence (*Is*), etc.

A description of the simplified design task implemented in this system can be formulated as: "within a fixed time

period, generate as many different shape compositions as possible by combining a set of initial shapes". Shape compositions are defined as arrangements of *n* final shapes created from the combination of less than *n* initial shapes. New shapes are created by the superposition of existing shapes which lead to the identification of new vertices in the intersections of line segments. This enables the emergence of new shapes as the set of paths {*LM*} from between the start and end points of figures *L* and *M* that lead through each intersection point, traversing each segment no more than once [11]. This shape arithmetic task provides a relative quantitative measure to compare two or more sets of results. A quality criterion is defined for this task as a function of the total number of new shapes created and their number of sides.

Figure 1 illustrates one composition created by this generative program. Further details on the complexity of this type of tasks are found elsewhere [13]. This two-dimensional shape representation is used to model divergent visual reasoning and is similar to those typically used in brainstorming research [6]. Whilst this design task is fairly straightforward to implement in a computer system, the results are varied enough to capture some of the key properties of design situations such as open-ended problem formulations with many appropriate solutions, and incremental development of solutions. A measure of task difficulty is defined by the number of initial shapes and the number of sides of these initial shapes. In this paper we present results using two initial shapes of three sides each.

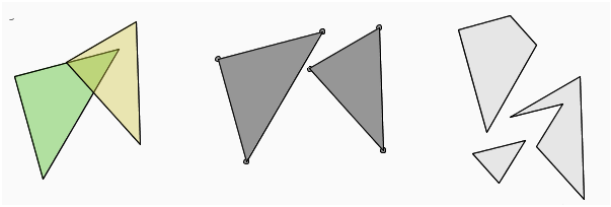


Figure 1 A random composition of 2 initial shapes where overlapping triangles are detected forming 3 emergent subshapes of 3, 4 and 5 sides, respectively. Shape compositions with more subshapes and subshapes with more sides are ranked higher.

The task is used to study group brainstorming by implementing a multi-agent system where agents are automated shape generators that search for new solutions, derive design concepts and interact in this process over a fixed time period. Agent behavior in this simplified model of brainstorming consists of the following behaviors: exploration function (random shape drawing and transformation), evaluation function (concept formation from topology relationships of shapes), and exploitation function (shape drawing and transformation by application of learned concepts).

Shape exploration in this program can be considered potentially creative inasmuch as emergent shape semantics "exists only implicitly in the relationships of shapes, and is never explicitly input and is not represented at input time" [12].

A design concept is defined here as a topology relationship between the initial shapes associated to the fitness of the final shape composition. More details are provided below. After a designer agent has generated one or more concepts, it can use them to generate new shapes. Exploitation strategies consist of random variations to existing design concepts. New compositions can then be obtained as a result of applying the modified rule and evaluating whether its outcome yields a new shape composition.

The following pseudo-code shows the algorithm to generate initial shapes and new emergent subshapes in this task (exploration function):

```
for(initialShapes) {
  select n random (x,y) points
  connect all pairs of points with lines
  build a polygon with resulting lines
}
for(every polygon) {
    for(every line i of every polygon) {
        find intersection point(line_i-line_n)
        store all vertex in a set
    }
}
for(all vertex in set) {
  build all subshapes via graph search (dijkstra)
  store new subshape in a set
}
eliminate duplicate subshapes
```

The following pseudo-code shows the algorithm to assign a qualitative measure to shape compositions (first part of evaluation function):

```
for(finalShapes) {
  fitness += (sides of subshape * finalShapes)
}
```

The following pseudo-code shows the algorithm to build design concepts in this task (second part of evaluation function):

```
for(all initialShapes s) {
  s.insideVertex += (vertex is within boundaries of shape s+1)
  s.outsideVertex += (!vertex is within boundaries of shape s+1)
  s.inLine += (vertex intersects line of shape s+1)
 s.coincidentVertex += (vertex is coincident with vertex of shape
s+1 }
 designConcept = { {insideVertex, outsideVertex, inLine,
     coincidentVertex} , fitness}
 store designConcept in a set
```

The exploration and exploitation mechanisms used here are inspired in the classic notions of divergent or 'horizontal' and convergent or 'vertical' thinking processes [3]. During brainstorming sessions, one may assume that exploration

enables the discovery of new types of solutions, whilst exploitation allows for the generation of alternatives or new instances –a kind of tradeoff in a multi-armed bandit problem.

In this system, designer agents start with exploration and transition to exploitation given a variable defined by the experimenter. The following pseudo-code shows the algorithm for selecting between exploration and exploitation:

```
for(designerAgents) {
  if (timeStep < exploreLength) strategy = "exploration"
  else {
    strategy = "exploitation"
    select a random designConcept  from set of concepts
    switch (designConcept) {
      case (insideVertex): initialShapes(insideVertex)
      case (outsideVertex): initialShapes(outsideVertex)
      case (inLine): initialShapes(insideVertex)
      case (coincidentVertex): initialShapes(coincidentVertex)
    }
  }
}
```

Group influence γ is defined as a sharing ratio of concepts: in the extreme case where $\gamma = 0$, agents have no access to the concepts generated by other agents; for cases $\gamma > 0$, agents have access to a fraction of the concepts generated by other agents up to $\gamma = 1$, where all agents have access to all concepts generated in the group. This experimental variable γ enables the modeling of both nominal and interactive groups in the brainstorming research literature, as well as scenarios similar to computer-mediated brainstorming where the researcher can control the level of interaction between participants [2].

Group influence γ is implemented in two sections of the code. First, agents store new design concepts in a shared team pool of concepts with a probability γ. Second, γ is also used in turn-taking on each simulation step. This is to account for the differential conditions in which nominal and interactive teams operate: when individuals work alone there is a type of allocation of turns in parallel, while teammates work in sequential turns. In this paper we inspect four γ scenarios: $\gamma = 0$, 0.33, 0.66 and 1.0.

Exploration length ε is defined as a ratio of total simulation time during which agents activate exploration behavior. This variable is used to model the timing at which brainstorming participants switch from exploration to exploitation behaviors. Although we acknowledge that such transition may take more complex patterns in real brainstorming sessions, in this paper we adopt a parsimonious approach as a foundation for future models. Exploration lengths $\varepsilon = 0.2$ to 1.0 are inspected in this paper in 0.2 increments.

In this paper we present and discuss results of four and sixteen-member groups where both group influence γ and exploration length ε are the experimental variables and both quantity and quality of generated ideas is the dependent variable. Gross fluency refers to the total number of design concepts generated during a simulation, while net fluency refers to the number of original or unique design concepts produced.

The impact of varying the level of group influence in idea fluency at different stages of a brainstorming session is likely to provide a possible explanation of the mechanisms behind the well-documented yet poorly understood phenomenon of 'ideational productivity loss' in group brainstorming.

Results

All results are mean values of 30 runs for every experimental condition. Control random-generator seeds are used in order to compare the effects of the independent variables. The trend is clear: as the scope of influence of ideas increases, fluency decreases across all exploration lengths. Table 1 shows the results for all 20 experimental conditions in gross and net fluency in four-member teams. When $\gamma = 1$, agents are activating the exploration strategy during 100% of the simulation; therefore no advantage from exploitation behavior is possible.

Table 1. Results in gross and net fluency from varying group influence γ in teams of 4 agents across a range of exploration lengths ε.

Exploration length ε	Group γ	Gross fluency	Net fluency
$\varepsilon = 0.2$	$\gamma = 0$	40.9	19.63
	$\gamma = 0.33$	53.86	19.46
	$\gamma = 0.66$	41.23	16.9
	$\gamma = 1$	18.4	9.2
$\varepsilon = 0.4$	$\gamma = 0$	48.46	22.76
	$\gamma = 0.33$	63.46	23.4
	$\gamma = 0.66$	47.66	19.23
	$\gamma = 1$	23.26	11.63
$\varepsilon = 0.6$	$\gamma = 0$	56.96	25.9
	$\gamma = 0.33$	69.46	25.43
	$\gamma = 0.66$	47.16	19.6
	$\gamma = 1$	29.13	14.56
$\varepsilon = 0.8$	$\gamma = 0$	57.56	25.03
	$\gamma = 0.33$	64.86	23.03
	$\gamma = 0.66$	44.9	18.26
	$\gamma = 1$	27.6	13.8
$\varepsilon = 1.0$	$\gamma = 0$	44.36	14.93
	$\gamma = 0.33$	44.43	14.13
	$\gamma = 0.66$	33.9	13
	$\gamma = 1$	22.06	11.03

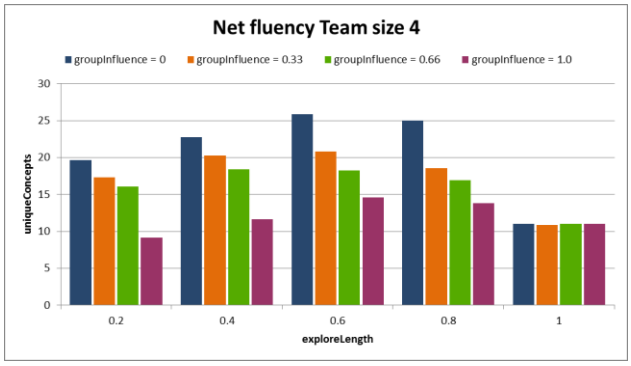


Figure 2. Group influence γ has a negative effect on net fluency across all exploration lengths ε. The differential effects of γ are higher when ε is low, and net fluency is higher across all γ when ε is medium. With high ε, the effects of γ are less significant.

Group influence γ has a clear effect on the generation of unique design concepts or net fluency, Figure 2. In this model, agents brainstorming in isolation do produce more original ideas than the *same* agents brainstorming in teams. These results are consistent across different team sizes from 4 to 16 members in our model, Figure 3.

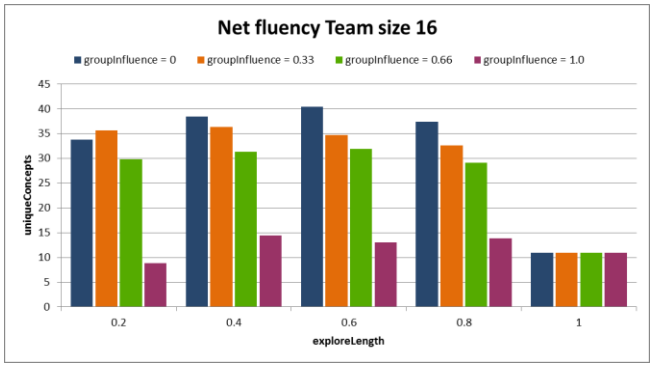


Figure 3. In large teams (N = 16), group influence γ has a more significant effect on net fluency across all exploration lengths ε.

When group influence is zero, agents contribute no solutions to the common pool of team concepts, and they only store and have access to their own concept pool. Gross fluency is the total sum of individual concepts, while net fluency is the count of original concepts in this set. Consistently across different team sizes, when group influence is zero, net fluency is highest indicating that agents in isolation generate more unique solutions than when they share solutions with others.

Although this outcome is consistent with the brainstorming literature, it still seems counter-intuitive; how can teams of agents in this model be less efficient than same-size groups of separate individuals? This result may seem paradoxical particularly when we consider the amplifying effects that the exploitation strategy has in this model, as evidenced by scenarios where agents explore the entire simulation time span (ε = 1.0). If exploitation is so productive (particularly when balanced with similar rates of exploration in this model), where does the advantage of isolated agents come from despite the fact that they have access to smaller solution sets during exploitation? In other words, one could expect that teams of agents in this model would be more productive given that each individual agent has access to a larger pool of concepts from which it can retrieve a higher diversity of solutions in order to build more concepts. In contrast, we observe that as group influence increases and agents contribute more and have more access to a larger pool of solutions, both gross and net fluency decrease. The gap between net fluency of nominal and interactive groups varies in this model as a function of exploration length, i.e., how early or late is exploitation activated during the simulated brainstorming session.

In larger groups the effects of group influence are more significant. Agents in large teams appear rather inefficient in high group influence conditions: their net fluency is equivalent to that of teams four times smaller. In this respect, it would be tempting to conclude that working in isolation is more efficient for creative ideation.

However, there is a fundamental distinction that is made clear in this model, which has largely remained implicit across studies that compare the performance of nominal versus interactive teams: the total output of these two types of groups is incommensurable. The key is turn-taking; the comparison is inadequate when measured in number of turns rather than in minutes or hours. The difference is that in isolation, although in theory the same number of individuals are generating and recording ideas, in fact the number of turns is *n* times higher than in interactive groups since turn-taking occurs in parallel. In principle, no idle time exists for individuals in nominal groups. In contrast, teams follow some type of sequential order (skewed or not) by which all team members except one are idle at every turn or intervention. Therefore, this natural 'bottleneck' in team interaction (*production blocking*) is a sufficient cause for the relative poor performance of teams when compared to the aggregate results of individuals in isolation.

In order to account for this inequality, turn adscription is manipulated in our model to ensure that all agents in nominal and interactive groups have access to an equal number of turns over the simulated time. The result is an increase in gross fluency as group influence increases, Figure 4.

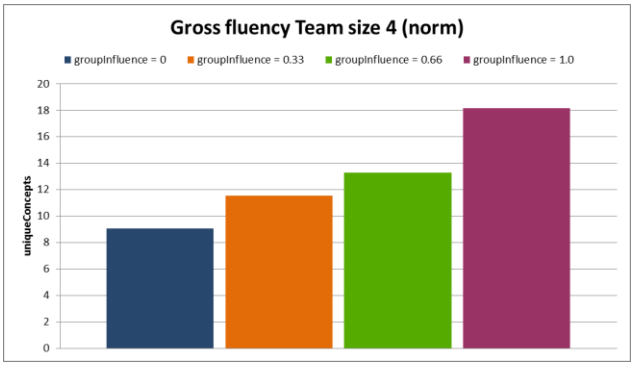


Figure 4. Teams outperform same-number of individuals in gross fluency as group influence γ increases (ε = 0.2).

Discussion

Is it better to generate ideas in solitude or in teams? The work presented in this paper suggests that each condition may present certain advantages and judging performance merely by measuring output is a limited approach. Although no definite answer can be expected from this simple model of brainstorming, it does capture interesting observations related to one of the key causes associated to ideational productivity loss, i.e. production blocking in groups. Within its limitations, this model supports a number of insightful hypotheses to consider:

- The balance between divergent and convergent thinking in a brainstorming session is important, and the time at which ideation is switched between these two modes of thinking is likely to have important an effect in the productivity of brainstorming groups.

- Individuals brainstorming in isolation are more productive than teams over similar time periods as a result of their increased *intensity* of participation. Teams present a 'bottleneck' in the form of sequential turn-taking, which is avoided by individuals in isolation who are –in principle– constantly active in generating new ideas and building up on previous ideas.

- The increased fluency of isolated brainstormers over teams may be a feature of easy tasks. It is possible that in difficult tasks, group diversity is more advantageous than individual ideational intensity. If this is the case, then transformative creativity may be more appropriate for brainstorming in solitude, whilst combinatorial creativity may be a more suitable objective of teams.

- Turn allocation can be optimized via facilitation techniques or technological means so that an adequate balance exists between having access to others' ideas and avoiding interruptions. This balance may turn out to be a key factor in the performance of brainstorming groups.

The work presented here focuses on the effect of influence over ideas; it is natural to expect a more complex picture that includes individual diversity and other situational conditions. Nonetheless, our results can be cautiously compared to those from laboratory studies. For instance, a widely-cited study of 4-person groups in the same two assessment conditions shows a productivity gain of around 60% from interactive to nominal groups [14]. On the other hand, another study where the total number of ideas is considered in 4-people groups but in a simpler task, reports a mean difference of 38% between nominal and interactive groups [15]. In our system these differences range between 40% and 100% depending on certain task factors.

Nevertheless, the aim of this system is not to replicate a particular task or set of results, but rather to demonstrate the nature and effects of production blocking in teams or interactive groups. In addition, these findings provide a possible explanation as to why people may enjoy more working in groups than in isolation [16, 17]. Apart from a number of social reasons, in terms of idea generation, our experiments suggest that individuals find it easier to operate in groups as they have access to a large number of ideas generated by others. Namely, significantly less individual effort is required to generate solutions.

If the results of this experiment were able to be generalized, then facilitators of brainstorming sessions should consider the aim of the session in relation to the type of demands imposed over the search of solutions, the degree of transformative or combinatorial creativity required, the social influence of the group (as a sum of paired influences between team members), and the resulting hierarchical interactions between brainstormers.

Brainstorming has been treated in general as a 'black box' method of problem solving. People are allocated into teams and they are expected to come up with solutions in a period of time with the general rule that they generate ideas without constraints. The importance of these simple computational experiments is that they show that the results of brainstorming sessions can be qualitatively different between independent individuals and groups, and also between different types of groups. Further modeling will be necessary in order to formulate and evaluate research-based instructions for adequate brainstorming sessions [18]. Future work with this model will account for individual agent diversity.

Acknowledgements

This research is supported in part by the National Science Foundation under Grant Nos. NSF IIS-1002079 and NSF SBE-0915482. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. Taylor, D. W., Berry, P. C., and Block, C. H. 1958. Does group participation when using brainstorming facilitate or inhibit creative thinking. *Administrative Science Quarterly*, 3, 23–47.

2. Isaksen, S. G. and Gaulin, J.P. 2005. A Reexamination of Brainstorming Research: Implications for Research and Practice. *Gifted Child Quarterly*, 49 (4): 315-329.

3. Osborn, A. F. 1957. *Applied Imagination*, Scribner, New York.

4. West, M.A. 2002. Sparkling fountains or stagnant ponds: an integrative model of creativity and innovation implementation in work groups, *Applied Psychology An International Review*, 51(3): 355-424.

5. Reiter-Palmon, R., Herman, A. E., and Yammarino, F. J. 2008. Creativity and cognitive processes: Multi-level linkages between individual and team cognition. In Michael D. Mumford, Samuel T. Hunter, Katrina E. Bedell-Avers, eds., *Multi-Level Issues in Creativity and Innovation*, Emerald, 203-267.

6. Paulus, P. B. and Nijstad, B. A. 2003. *Group Creativity: Innovation through Collaboration*, Oxford University Press, Oxford.

7. Paulus, P. B. and Brown, V. 2003. Ideational creativity in groups: Lessons from research on brainstorming. In PB Paulus and B Nijstad, eds., *Group Creativity: Innovation Through Collaboration*, Oxford University Press, New York, 110-136.

6. Brown, V. and Paulus, P. B. 2010. A Simple Dynamic Model of Social Factors in Group Brainstorming, *Small Group Research*, 27(1): 97-114.

7. Gilbert, G. N. and Conte, R. 1995. *Artificial Societies: The*

Computer Simulation of Social Life, UCL Press, London.
8. Iyer, L. R., Minai, A. A., Brown, V. R., Paulus, P. B. and Doboli, S. 2009. Effects of Relevant and Irrelevant Primes on Idea Generation: A Computational Model. *Proceedings of International Joint Conference on Neural Networks*, Atlanta, Georgia, USA, June 14-19, 2009.
9. Paulus, P. B., Levine, D.S., Brown, V., Minai, A. A., and Doboli, S. 2010. Modeling Ideational Creativity in Groups: Connecting Cognitive, Neural, and Computational Approaches. *Small Group Research* 41: 688-724.
10. Sosa, R., Gero, J. S. and Jennings, K. 2009. Growing and destroying the worth of ideas, *C&C'09 Proceedings of Conference on Creativity and Cognition*, ACM, 295-304.
11. Gross, M.D. 1996. Emergence in a recognition based drawing interface, *Visual and Spatial Reasoning II*, J. Gero, B. Tversky, T. Purcell, eds., Key Centre for Design Cognition and Computing, Sydney Australia, 51-65.
12. Gero, J. S. and Jun, H. 1995. Getting computers to read the architectural semantics of drawings, *in* L. Kalisperis and B. Kolarevic (eds), *Computing in Design: Enabling, Capturing and Sharing Ideas*, ACADIA, 97-112.
13. Stouffs, R. and Krishnamurti, R. 2006. Algorithms for classifying and constructing the boundary of a shape, *Journal of Design Research* 5(1): 54-95.
14. Diehl, M. and Stroebe, W. 1987. Productivity loss in brainstorming groups: Toward the solution of a riddle, *Journal of Personality & Social Psychology*, 53(3): 491-509.
15. Paulus, P. B. and Dzindolet, M. T. 1993. Social influence processes in group brainstorming, *Journal of Personality & Social Psychology*, 64(5): 575-586.
16. Nijstad, B. A. and De Dreu, C. K. W. 2002. Creativity and group innovation, *Applied Psychology An International Review*, 51(3): 400-406.
17. Paulus, P. B., Dzindolet, M. T., Poletes, G. and Camacho, L. M. 1993. Perception of performance in group brainstorming: The illusion of group productivity, *Personality & Social Psychology Bulletin*, 19(1): 78-89.
18. Paulus, P. B. and Dzindolet, M. 2008. Social influence, creativity and innovation. *Social Influence*, 3(4): 228-247.

Representational affordances and creativity in association-based systems

Kazjon Grace
Faculty of Architecture, Design and Planning
Sydney University
NSW, Australia
kazjon@arch.usyd.edu.au

John Gero
Krasnow Institute
for Advanced Study
George Mason University
Fairfax, VA, US
john@johngero.com

Rob Saunders
Faculty of Architecture, Design and Planning
Sydney University
NSW, Australia
rob.saunders@sydney.edu.au

Abstract

This paper describes ongoing research into association-based computational creative systems. The necessary components for developing association-based creative systems are outlined, and the challenges in measuring the creativity of such a system are discussed. An approach to operationalising creativity metrics for association-based systems based on representational affordances is described. This approach is then demonstrated through an analysis of results produced by a system for constructing associations between visual designs.

Association-based creative systems

Association, or the construction of a new relationship between two concepts or ideas, is a cognitive process at the heart of many creative endeavours. Its presence is most obviously felt in analogy and metaphor, but associative reasoning is also a component of complex similarity judgement, recognition and simplification tasks (Markman and Gentner 1993; Balazs and Brown 1998; Goel 2008) that are critical to the appreciation of creative works. Given the creative potential of analogical processes (Goel 1997; Hofstadter and the FARG 1995; Kuhn 1962) and the importance of understanding and appreciating creative works (Jennings 2010; Wiggins 2006; Colton 2008) to a computational creative system, it is clear that an operationalised understanding of the process of association that underlies these and other acts is of value to the field of computational creativity. Furthering that understanding is a twofold endeavour: computational models of association that are general, extensible and powerful must be developed, and metrics by which the creativity of those models can be assessed must be devised.

Association involves representing objects in a manner that enables a new relationship between them. A mapping is then constructed between the objects which embodies that relationship. These two component processes - representation and mapping - cannot be modelled serially or discretely, as representation depends on mapping and mapping depends on representation (Kokinov 1998). This complex relationship between the mapping and the representations used in mapping creates a 'chicken-or-egg' problem that must be addressed by any computational model. Not only must a computational model of association possess representational flexibility, but the search for representations must be informed by feedback from the ongoing search for mappings, just as the search for mappings is influenced by the construction of new representations.

Notably the process of association does not incorporate the use or evaluation of the relationships that it constructs. This is the primary addition of processes like analogy that extend association - analogy adds the transfer of knowledge between the associated object, the use of that knowledge to achieve some goal, and the evaluation of the analogy in terms of its utility at achieving that goal (French 2002). Association-based similarity judgement also extends association, in this case by evaluating mapped and unmapped attributes to construct a notion of similarity between the objects and then using that similarity in some categorisation or comparison task (Markman and Gentner 1993). Models of association must be capable of supporting this variety of applications.

This research has developed the notion of interpretation-driven search as a general framework for computational association. We investigate this approach for its potential to exhibit creative behaviours.

Interpretation-driven association

Donald Schön (1983) proposed a theory, 'reflection-in-action', to explain the cyclical interactions of evaluation and synthesis processes that had been observed in studies of designers. Schön suggests that designers change the design representations with which they are working, then observe and reflect on the effects of those changes. As a result of that reflection, the designer again acts to change the emerging design representation. This iterative interaction is enabled by the designer's ability to interpret a representation in a new way after it has been produced. Schön posits that the designer's ability to see things in an emerging design that were not consciously put there is the core of the creative design process.

The framework for computational association developed in this research draws a parallel between Schön's theory of creative design and Boden's (1990) notion of creativity as exploring (and potentially transforming) a conceptual space. The actions taken by a designer to modify their design may translate that design to a new position within the designer's conceptual space, or they may transform the space itself, re-

formulating the designer's understanding of the problem and producing a novel and surprising design. The genesis of both exploratory and transformative reformulations is the reconceptualisation of the representation the designer had constructed previously. This produces a new interpretation of the design on which previously impossible actions are rendered possible.

Schön sees the process of reflection-in-action as itself being based on analogical reasoning (Schön and Wiggins 1992), but this research inverts that relationship, putting forward a framework for association that is based on Schön's iterative cycle of reflection and action. This framework is referred to as interpretation-driven search. While inspired by the design process, the interpretation-driven search approach can be generalised beyond design tasks to any domain in which potentially creative associations are constructed.

Interpretation-driven association uses iterative transformation and exploration of the objects being associated to produce a representation that enables a new mapping to be constructed. An interpretation is a transformation of the representation of the objects being associated. These transformations affect the object representations and enable potential mappings between them to be explored. In this approach, interpretations are explicitly represented elements of system knowledge, allowing them to be constructed, evaluated, stored and retrieved. The interpretation process iteratively interacts with the process of searching for mappings and operates in parallel with it. Interpretation influences mapping search and mapping influences the construction, application and evaluation of interpretations.

A model of association that implements these principles can broadly be viewed a consisting of three processes: Representation, Interpretation and Mapping. Representation produces the 'original' representations of the objects that are then iteratively searched, transformed and mapped by the Interpretation and Mapping cycle. This framework can be seen in Figure 1. The benefits of this parallel, interactive approach are discussed in Grace et. al. (2012), along with a more detailed elaboration of the framework.

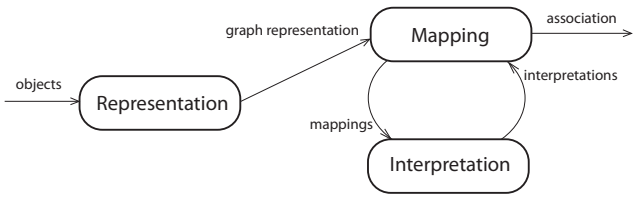


Figure 1: Interpretation-driven search, a high-level framework for computational association.

The creativity of associations

The definition of, and criteria for, creativity have been the subject of considerable debate. One broad definition that has attained some consensus is that of creativity as the union of *novelty* and *value* (Sternberg and Lubart 1999). Novelty is a metric based on the difference between the artefact and other, existing artefacts in the same domain. Value is a metric based on the artefact's performance at whatever tasks to which it is applied, when compared to the performance of existing artefacts. Both of these qualities are highly contextualised, as novelty can only be assessed from the perspective of a viewer and usefulness can only be assessed in the context of an application.

Some challenges arise in applying this pair of creativity criteria to the domain of computational association. Firstly, the novelty of an association is on some level guaranteed, as by definition an association must be a new relationship that did not exist previously. Recalling a relationship of which a system was already aware is a memory task, not an association one. This makes an association always at least P-novel (novel to the system itself, as defined by Boden (1990)).

A significant challenge in applying the "novelty and value" framework for evaluating creativity to a model of association is in assessing an association's value. Association does not necessarily incorporate an evaluative component and it is not necessary that an association be constructed to serve some purpose. We refer to this goal-agnostic form of association as 'free' association, which may incorporate evaluative components but in which the associations are not used to accomplish some purpose. Evaluation and purposefulness are instead features of association-derived processes that incorporate additional components. This does not mean, however, 'free' association has no effect on the system that constructed it, and therefore an alternative assessment for value can be derived. Different associations produce different transformed and mapped representations of the associated objects, and their value can be assessed based on the degree to which those representations go on to affect the system. This research focusses on this representational affordance model of association value as a way by which the model of association that has been developed could be further developed into an association-based creative system.

Representational affordance as a utility metric

The "affordances" of an object or environment were first defined by the psychologist James Gibson (1979), referring to the opportunities it offers to a user. As applied to the design of objects (Norman 2002) affordances refer to the possibilities for action that a user perceives when interacting with an object. Affordances do not require instruction, they emerge implicitly from the interaction of an object, its user and the situation (Maier and Fadel 2009).

A representation is an internal surrogate that encapsulates knowledge about an entity, enabling the agent or system to reason about that thing (Davis, Shrobe, and Szolovits 1993). In any system that permits the construction of different representations of an object, those representations will facilitate the performance of different actions by that system. Different representations of objects within a system open up different action possibilities for that system. Gero and Kannengeisser (2012) refer to this as representational affordance: the cognitive actions that are enabled by a representing an object in a particular way. During the design process a representation may afford the construction of a new representation with its own, different set of affordances. Gaver

(1991) refers to this as "sequential affordance" and it is consistent with the notion of reflection-in-action (Schön 1983).

Representations can provide affordances based on their syntax or based on their semantics. The structure of a graph representation can provide syntactic affordances, such as path following or matching. However, a representation can also provide semantic affordances based on how its content can interact with other system knowledge. This paper focusses on semantic representational affordances.

In modelling the creativity of an association, a key question arises: what is the value of a new mapping and the representations that underlie it? We define an association's value in terms of what activities the possession of that association enables the system to do. An association can be said to be of value if the interpretation of the associated objects it contains provides the system with different representational affordances than it previously possessed. Furthermore, associations can be compared and contrasted by the affordances they provide.

Value can be defined using representational affordances in the absence of any specific objectives or purpose of the association construction process, making it apt for use in a general model of creative association. In the case of an analogy-making system built on a model of association, the affordances that would be most relevant would be those that enable acts of knowledge transfer between the object domains. By contrast, in a model of design style the most relevant affordances would be those that permitted the detection of new patterns that connect stylistically similar objects.

A model of 'free' association that does not extend the process to incorporate a use for the mappings it constructs can also be assessed using the representational affordance metric for value. If the goal of a free association system is to construct as many different associations as possible, then valuable associations are those that afford the possibility of future, different associations. This kind of sequential affordance of association is made possible by association models that incorporate the effects of a system's past experiences in constructing associations into new association tasks.

In this research we use the notion of representational affordances as a value metric for association models to discuss the potential creativity of results from a computational model of association.

Experimenting with interpretation-driven association

A computational model based on the interpretation-driven framework for association has been developed. An implementation of that model which constructs 'free' associations (in that the associations it constructs are not used for any explicit goals) between ornamental designs is described here. The structure of the model and its prototypical implementation are presented here, along with selected association results produced by the system. The potential representational affordances of the results presented are discussed as a first step towards extending this model towards an association-based creative system.

Computational model

Interpretation-driven search builds on the model of analogy as Structure Mapping (Gentner 1983), in which the relationships within two objects are mapped, rather than their features. The search for these relationship mappings is integrated with an iterative process of re-representation.

The model of interpretation-driven association (see Grace et. al. (2012) for a detailed description) is comprised of five processes. The first three processes: concept formation, relation formation and graph construction collectively form the "representation" process of the interpretation-based framework, while the latter two processes, mapping and interpretation, are direct implementations of that framework.

The system begins with an image-based representation of the objects, extracts a set of features to describe them and then categorises those features into concepts. Relationships between these features within each object are then constructed based on both topological information (such as relative size, bearing or symmetry) from the feature sets and typological information from the conceptual categorisation (such as conceptual similarity or conceptual sameness). The features and relationships are then compiled into a graph representation that serves as the basis for the iterative mapping and interpretation. The mapping process then searches these graphs for subgraphs that contain common edge labels. These subgraphs represent regions of the two images that possess a consistent relational structure.

The transformations that are applied by the interpretation process affect the structure or content of the object graph representations. Implementations of this model could utilise a variety of transformational approaches, such as transforming the graph objects directly, transforming the features or concepts directly and then re-constructing the graphs, or even transforming the process by which one or more representational stages are constructed.

At any given time, a single transformation is applied to the graph representations, this is referred to as the 'current' interpretation. This interpretation changes the structure of the graphs, altering the trajectory of the mapping search operating on those graphs. The mapping search process produces candidate mappings as it searches, and these are used to construct new interpretations. New interpretations are constructed by examining what features-to-feature mappings in those candidates cannot currently be successfully mapped, and extrapolating what transformations would be necessary to cause those to be successful.

Implementation

The implementation of the model uses vector images as its input, calculating object features from the minimal closed shapes formed by vector lines. The kinds of relationship implemented in the system are 'same concept', 'similar concept', 'relative scale', 'linear distance', 'horizontal distance', 'vertical distance', 'relative orientation', 'bearing', 'contains', 'reflection of', 'shared vertex' and 'shared edge'. The implementation is provided with the knowledge necessary to detect these relationships and categorise them into groups such as "slightly smaller than" or " 120 degrees of

difference in orientation". Instantiations of these relationships form the edge labels on the graph representations of each object being associated.

Mapping search is implemented as a genetic algorithm that searches for subgraph isomorphisms between the graph representation of each object. Each individual in the population of the genetic algorithm is a set of mappings between a feature in one object and a feature in the other. The fitness for this algorithm is the largest contiguous subgraph that can be constructed out of those feature-to-feature mappings in both objects. This use of a powerful, general search algorithm reflects the fact that we are not attempting to implement association in a biologically or cognitively plausible way, rather we are demonstrating the feasibility of the interpretation-driven approach.

The interpretation process is implemented as the substitution of relationships between features. Replacing relationships effectively causes the system to perceive two disparate relationships as being alike. Interpretations in this system can be expressed as "in this situation, relationship X in the first object is the same as relationship Y in the second object'. An interpretation is therefore a set of rules for replacing relationships, where relationships are represented as edge labels in the graphs. Which interpretation is being applied to the objects is able to change every iteration, providing the parallelism between mapping and interpretation that characterises the interpretation-based framework.

Methodology

A total of 31 ornamental designs were inputted into the system as part of a series of experiments to demonstrate the application of interpretation-based association. Objects were drawn from a broad variety of design domains, including symbols, architectural ornamentation and decorations and object designs. These objects were drawn from a variety of cultures and historical periods. From this library of designs a subset of objects were selected for which interesting associations could be produced and the capabilities of the system could be documented.

A set of associations constructed by repeatedly associating a single pair of objects is presented. These associations are presented as a demonstration of the interpretation-based model, but also as a starting point from which the use of representational affordances as a metric for utility in association-based creative systems can be discussed.

The two objects associated here are presented in Figure 2. Object 1, on the left, is a Hittite sun symbol, while Object 2, on the right, is a Japanese floral symbol. Both are vector line drawings produced manually from black and white images by the authors. For the purposes of this experiment the system has been restricted so that the only type of relationship which connects the features of these two objects is relative orientation.

Results

Three associations between the two objects in Figure 2, along with the interpretations used to produce them, are shown in Figures 3, 4 and 5. All three associations constructed between these two objects utilised the 'relative orientation' relationship type, but each involved a different interpretation. These differing interpretations permitted the construction of different mappings.

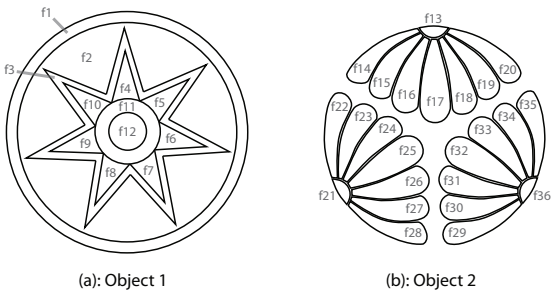


Figure 2: The two objects used in the example associations. The minimal closed shapes extracted from the images by the system are numbered. The relative orientations of these features are the relationships relevant to these examples. Designs sourced from (Humbert 1970).

In each of these figures the associated objects are presented side by side, with the features involved in the mapping being highlighted. The mapping between features in one object and features in the other is shown as solid lines joining the two images. The common set of relationships between the features within each object is shown as thick dashed lines. Only the relationships that are used in the mapping are shown, pairs of features can have many relationships connecting them. Each of these relationships is labelled with its uninterpreted description. Interpretations are an imposed equality between different labels and are shown at the bottom of each image. Mappings can be constructed between sets of features that share patterns of relationships after this interpretation is applied.

The first association, seen in Figure 3, is constructed without the use of an interpretation. Within the representations of the two objects there exists a pattern of seven nodes in each that share a consistent pattern of relationships. All seven objects in both objects, in the order indicated by the thick dashed lines, are consecutively separated by approximately 150 degrees of orientation. This relationship is present between every second point in the seven-pointed star in Object 1, starting from Feature $f4$ and proceeding twice around the star to Feature $f9$. The same relationship of relative orientation is present between every eighth petal in Object 2 - that is between each petal and the petal one spot to its left in the floret to its left, starting from Feature $f35$ and proceeding in a spiralling pattern twice around the design to $f29$. The 'null' interpretation $i_\emptyset$ is shown as this mapping can be constructed from the base representations produced by the system without any transformation.

This association is included to demonstrate the capability of the representation construction and mapping search elements of our model of association. Without the use of interpretations the system is capable of producing representations of visual objects comprised of networks of abstract

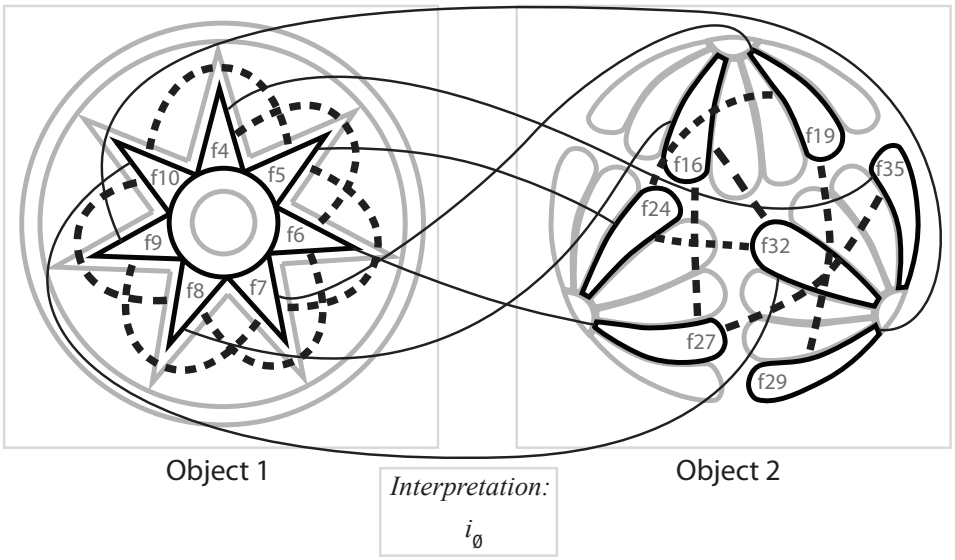


Figure 3: An association constructed between the two objects without the use of an interpretation. All the relationships incorporated into this mapping (depicted by thick dashed lines) are of the type "∼150 degrees of difference of orientation". These relationships join the seven points of the star in Object 1 (by traversing the star twice) and seven of the petals in Object 2 (in a spiral pattern joining each eighth petal). The empty or "null" interpretation means that these relationships are present in the default object representations the system constructs.

relationships and features. These representations can then be searched for common patterns of relationships, allowing the features which those relationships join to be mapped.

Without the ability to transform representations, this association (and others trivially different from it) are all that the system can construct. With mappings limited to those relationships already present in both objects, the potential for constructing an association with relevant affordances is slim. The only way to give a system with a single representation of each object the ability to construct additional associations is to incorporate more information into those representations. In this case that information would take the form of additional types of relationship between features other than differences of orientation.

Utilising the capacity to reinterpret object representations, the system is not limited to "literal" associations as seen in Figure 3. Figure 4 shows the mapping produced by an association between the same two objects that was produced through interpretation-driven search. During the search for mappings between the objects, the system constructed (initially by chance) some fragment of a mapping similar to the one shown in Figure 4. This mapping candidate would not have been successful in the absence of an interpretation. The mapping was selected by the interpretation construction process, which reverse-engineered one or more interpretations from it. Interpretations are generated that would improve the size of the largest common subgraph of the mapping specified by the candidate. That interpretation is then likely to become the "active" one if the mapping search reaches a point where the current interpretation is significantly outperformed by the new interpretation. The search for mappings influences the construction of interpretations and then those interpretations in turn influence mapping.

The mapping expressed in Figure 4 is based on an interpretation that effectively treats the orientation difference between adjacent points on the star in Object 1 the same as the orientation difference between adjacent points on the star in Object 2. Thick dashed lines are shown representing the "approximately 50 degrees of orientation difference" relationship in Object 1 and the "approximately 20 degrees of orientation difference" relationship in Object 2. The resultant mapping connects each feature in the star in Object 1, starting with $f4$ and proceeding sequentially to $f10$ with each feature in a floret in Object 2, starting with $f14$ and proceeding sequentially to $f20$. This mapping was constructed from low-level relationships extracted from a visual representation of these objects and then interpreted to make those relationships situationally alike.

The notion of representational affordances as a tool for assessing value in association models can be applied to the association shown in Figure 4. The interpreted representations of the two objects are useful to the extent that the new interpretations more aptly afford actions to the system. In the case of a free association model like this implementation, the only actions available are the construction of different associations, and therefore the value of an association can only defined by the degree to which it enables that. In this system interpretations are remembered and re-used, which causes the system's past experiences to affect future interpretations. This provides a mechanism by which this association can guide the system's future actions.

Representational affordances provided by the association

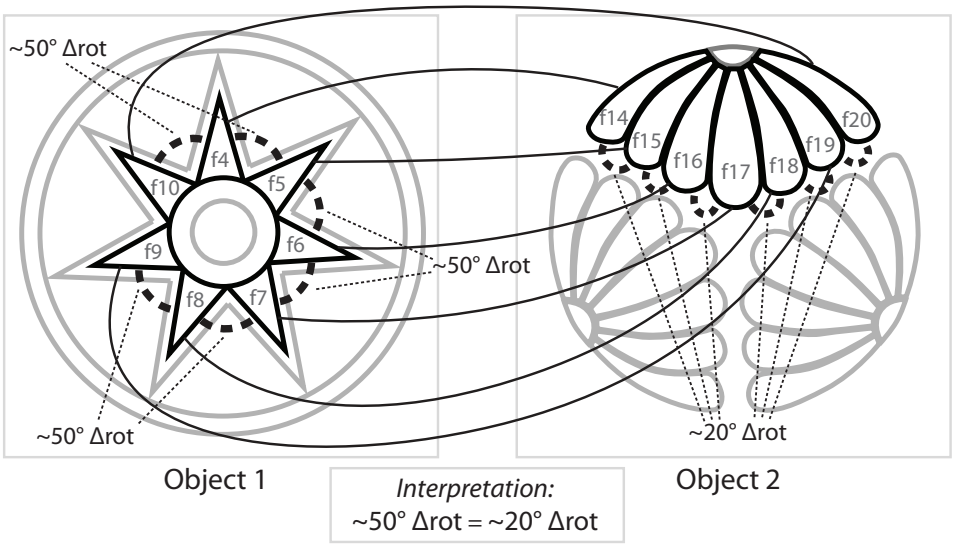


Figure 4: An association constructed between the two objects through the application of an interpretation which equates the relationship "~50 degrees of difference of orientation" in Object 1 to the relationship "~20 degrees of difference of orientation" in Object 2. These relationships join the adjacent points of the star in Object 1 with eight adjacent petals in one of the florets in Object 2. The interpretation that enabled this association is shown in the box beneath the objects.

in Figure 4 could be used to define the value of that association if the system were extended to perform purposeful association construction. If associations were being constructed for use by an ornamental design classification system, then the transformed representation may afford the possibility of classifying Object 2 as being based on radial adjacency, which a naive classification system would have been able to do. If instead associations were being constructed for use in an analogy-based design system, then the mapping of $f4$ through $f10$ to $f14$ through $f20$ may afford the transfer of knowledge about $f11$ (the centre circle which all of the mapped features in Object 1 touch) to $f13$, which could then be considered the "centre" of Object 2. The actions permitted by the representational affordances of the association are used by the system to achieve its goals, so those affordances can be said to be of value because of that use.

Figure 5 shows a different association constructed by the system. This time the interpretation equates the difference in orientation between every third point in the star in Object 1 with the difference in orientation between the edge petals in Object 2. Viewing the "edges" of a compound object such as Object 2 as being part of a rotating sequence may be a valuable affordance for a creative system. The association system implemented in this research was able to find a broad variety of such associations using just two objects and considering just one type of relationship. Different combinations of the relationships that were mapped in Figures 4 and 5 were also found, such as mapping every adjacent point on the star in Object 1 to the outermost petals of the three florets in Object 2. Interpretation construction provided the system with the ability to produce a variety of divergent mapping from a single association problem.

Discussion

The experiments described here demonstrate that it is possible to use the interpretation-driven search approach to construct associations between real-world design objects. The representation construction processes used to produce the graphs on which the iterative mapping and interpretation processes operate have been shown to be viable. Associations were produced based on interpretations that were constructed by the system that transformed graph representations that had also been constructed by the system from features and concepts that had been extracted from low-level visual input. These results serve as an initial proof of concept of the interpretation-driven model of association.

The associations presented in this paper could not have been constructed from the information that was provided to the system without the ability of the system to transform its representations through the interpretation process. These associations could be constructed without the use of interpretation if the system were provided with additional information about the relationships between the objects, but this reduced representational autonomy would have a deleterious effect on novelty. As assessed in the context of a hypothetical society of individuals with access to the same information and possessed of comparable perceptual abilities, associations produced using interpretation will be P-novel to any individual that has not constructed the same interpretation, while associations produced using additional information would be apparent to any other individual with access to that information.

The P-novelty of an association is guaranteed as the system by definition did not know of the relationship expressed in an association before its construction. However, the inter-

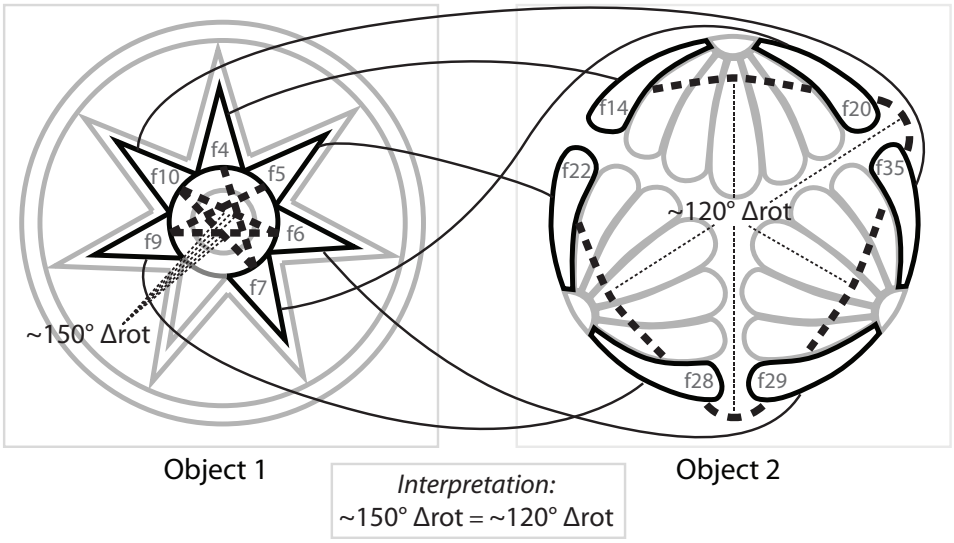


Figure 5: An association constructed between the two objects through the application of an interpretation which equates the relationship "~150 degrees of difference of orientation" in Object 1 with the relationship "~120 degrees of difference of orientation" in Object 2. These relationships join every third point of the star in Object 1 with the two outermost petals in each floret in Object 2. As there are only six such petals in Object 2, only six of the seven points in Object 1 have been mapped. The interpretation that enabled this association is shown in the box beneath the objects.

pretations used in the model of association described in this research may not themselves be novel as they can be learnt and re-used. It is possible to apply a known interpretation to a different object and still produce a novel representation, as a known transformation can produce a novel result.

It is also possible that a P-novel association can be constructed from a representation that has been used before. For example, if the representation of Object 1 present in 4 (and the interpretation used to produce it) were associated with some other, different object, then the resulting association would be P-novel. The known representation of Object 1 is novel in the current circumstances, but is not novel to the system as a whole. This is referred to as "situational" or "S" novelty, after the definition of "S-creativity" in Suwa et al. (1999).

Figure 3 demonstrates an inherent weaknesses of a model of creative association that does not incorporate the ability to reinterpret its object representations. The mapping it expresses is quite likely H-novel, in that it is not expected that the majority of human observers would have identified that mapping. However, any sufficient pattern matching system provided with the two graph representations used by the system would come to the exact same conclusion. The mapping used in this association may be H-novel in a society composed of individuals with human perceptual biases, but to a hypothetical society composed of individuals with a perceptual system like that of this model, the mapping is obvious.

The use of an interpretation in the construction of associations acts to redress the weakness present in models of creative association based on static representations. The association expressed in Figure 4 could not be constructed from the graph representations the system built without the use of a representation transformation process. In a hypothetical society of individuals with the same perceptual biases as this system, the mapping used in this association would be P-novel to any individual that had not constructed the same interpretation. If this same mapping were to be constructed by incorporating a new relationship type in the default representations that made the mapping possible without interpretation - "radial adjacency" for example - then the resultant mapping would, like the one in Figure 3, be trivially deducible by any other pattern matching system with access to the same information.

The utility component of evaluating creativity can be aided by the use of representational affordances. In the interpretation-driven approach to association, mappings are produced between transformed representations that can reveal structures and connections that were not apparent in the original representations. The utility of those mappings can then be assessed by what actions those new structures enable the system to take. For example, the view of the outermost petals in Object 2 as a sequence of rotated features seen in Figure 5 was not expressed in the uninterpreted representation. The notion of representational affordance also allows the value of an association to be defined in the abstract for models of association that do not use the associations they construct to accomplish any objectives.

Maher (2010) frames the evaluation of a creative artefact as requiring three criteria; not just novelty and value but also unexpectedness. Unexpectedness (also referred to as 'surprisingness') is a metric based on how different the artefact is to what was expected to be the next artefact produced.

Unexpectedness, writes Maher, differs from novelty in that it relates to the expected trajectory of the domain or field in which the artefact is being produced, which is distinct from the existing set of artefacts within that domain.

Assessing the unexpectedness of an association using Maher's criteria requires identifying abstract patterns in the sequence of recently constructed associations that can be used to project a trajectory of expected associations. Associations can certainly be surprising in a variety of ways - much humour depends on setting the recipient up to expect that a certain association is being proposed, then subverting that expectation and instead constructing a very different association. However, unexpectedness as defined by Maher specifically refers to identifying emerging trends in the output of a system over multiple iterations. This is a challenging task in the domain of association as it is difficult to define a similarity metric that could then be used to find patterns in association output.

The interpretation-based model of association could provide a method by which the unexpectedness of associations could be assessed. Interpretation-based associations could be characterised by the interpretations used to construct them, which are significantly more generalisable than the mappings that comprise those associations. The explicit representation of interpretations permits further investigation of expectation and unexpectedness in computational models of association. This would address a challenging issue in the development of creative models of association and methods by which they can be evaluated.

The representational affordance framework for assessing the value of an association allows us to consider the associations constructed by an interpretation-based system as a creative artefact. The utility of that artefact is the degree to which it has an effect on the system that constructed it. In a "free" association system where experience plays a role in future associations then that effect can be defined as the influence an association has on the construction of future associations.

References

Balazs, M. E., and Brown, D. C. 1998. A preliminary investigation of design simplification by analogy. In *AI in Design Conference*. Kluwer.

Boden, M. 1990. *The Creative Mind*. Abacus.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *Proceedings of the AAAI Spring Symposium on Creative Systems*.

Davis, R.; Shrobe, H.; and Szolovits, P. 1993. What is a knowledge representation? *AI Magazine* 14(1):17–33.

French, R. 2002. The computational modeling of analogy-making. *Trends in Cognitive Sciences* 6(5):200–205.

Gaver, W. 1991. Technology affordances. In Robertson, S.; Olson, G.; and Olson, J., eds., *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Reaching through Technology*, 79–84.

Gentner, D. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science* 7:155–170.

Gero, J., and Kannengiesser, U. 2012. Representational affordances in design, with examples from analogy making and optimization. *Research in Engineering Design* (to appear).

Gibson, J. 1979. *The Ecological Approach to Visual Perception*. Houghton Mifflin.

Goel, A. 1997. Design, analogy, and creativity. *IEEE Expert* 12(3):62–70.

Goel, A. 2008. Analogical recognition of shape and structure in design drawings. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 22:117–128.

Grace, K.; Gero, J.; and Saunders, R. 2012. Constructing computational associations between ornamental designs. In *CAADRIA 2012*. to appear.

Hofstadter, D., and the FARG. 1995. *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. New York, NY: Basic Books.

Humbert, C. 1970. *Ornamental Design*. Thames and Hudson.

Jennings, K. E. 2010. Developing creativity: Artificial barriers in artificial intelligence. *Mind and Machines* 20(4):489–501.

Kokinov, B. 1998. Analogy is like cognition: dynamic, emergent, and context-sensitive. In Holyoak, K.; Gentner, D.; and Kokinov, B., eds., *Advances in analogy research: Integration of theory and data from the cognitive, computational, and neural sciences*. NBU Press. 96–105.

Kuhn, T. 1962. *The structure of scientific revolutions*. University of Chicago Press.

Maher, M. L. 2010. Evaluating creativity in humans, computers, and collectively intelligent systems. In *Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design*. Lancaster, UK: Design Network.

Maier, J., and Fadel, G. 2009. Affordance based design: A relational theory for design. *Research in Engineering Design* 20(1):13–27.

Markman, A., and Gentner, D. 1993. Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language* 32(4):517–535.

Norman, D. 2002. *The Design of Everyday Things*. New York: Basic Books.

Schön, D., and Wiggins, G. 1992. Kinds of seeing and their functions in designing. *Design Studies* 13(2):135–156.

Schön, D. 1983. *The Reflective Practitioner*. Basic Books.

Sternberg, R., and Lubart, T. 1999. The concept of creativity: Prospects and paradigms. In Sternberg, R., ed., *The Handbook of Creativity*. Cambridge University Press. 3–15.

Suwa, M.; Gero, J.; and Purcell, T. 1999. How an architect created design requirements. In Goldschmidt, G., and Porter, W., eds., *Design Thinking Research Symposium: Design Representation*, volume 2. MIT Press. 101–124.

Wiggins, G. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Journal of Knowledge Based Systems* 19(7):449–458.

How Did Humans Become So Creative? A Computational Approach

Liane Gabora
Department of Psychology
University of British Columbia, 3333 University Way
Kelowna BC, CANADA, V1V 1V7
liane.gabora@ubc.ca

Steve DiPaola
Department of Cognitive Science / SIAT
Simon Fraser University, 250-13450 102 Ave
Surrey BC, CANADA, V3T 0A3
sdipaola@sfu.ca

Abstract

This paper summarizes efforts to computationally model two transitions in the evolution of human creativity: its origins about two million years ago, and the 'big bang' of creativity about 50,000 years ago. Using a computational model of cultural evolution in which neural network based agents evolve ideas for actions through invention and imitation, we tested the hypothesis that human creativity began with onset of the capacity for recursive recall. We compared runs in which agents were limited to single-step actions to runs in which they used recursive recall to chain simple actions into complex ones. Chaining resulted in higher diversity, open-ended novelty, no ceiling on the mean fitness of actions, and greater ability to make use of learning. Using a computational model of portrait painting, we tested the hypothesis that the explosion of creativity in the Middle/Upper Paleolithic was due to onset of contextual focus: the capacity to shift between associative and analytic thought. This resulted in faster convergence on portraits that resembled the sitter, employed painterly techniques, and were rated as preferable. We conclude that recursive recall and contextual focus provide a computationally plausible explanation of how humans evolved the means to transform this planet.

Introduction

To gain insight into the mechanisms underlying creativity, one might start by testing peoples' creative abilities, perhaps using technologies such as fMRI, or dissect the brains of people who were known to be particularly creative during their lifetimes. However, to gain insight into the *evolution* of creativity, these options do not exist. All that is left of our prehistoric ancestors are their bones and artifacts such as stone tools that resist the passage of time. Thus to understand the evolution of creativity, computational modeling is virtually the only scientific tool we have.

Humans are not only creative; we put our own spin on the inventions of others, such that new inventions build cumulatively on previous ones. This cumulative cultural change is referred to as the *ratchet effect* (Tomasello, Kruger, & Ratner, 1993), and it has been suggested that it is uniquely human (Donald, 1991). A mathematical model of two transitions in the evolution of the cognitive mechanisms underlying creativity has been put forward (Gabora & Aerts, 2009). Computational models of these mechanisms have also been developed (DiPaola & Gabora, 2007, 2009; Gabora, 1995, 2008a,b; Gabora & Leijnen, 2009; Leijnen & Gabora, 2009, 2010; Gabora & Saberi, 2011). However, these efforts used different modeling platforms, and because the aims underlying them have been part scientific and part artistic, their relevance to each other, and to an overarching research program has not previously been made clear. The goal of this paper is to explain how, together, they constitute an integrated effort to computationally model the evolution of human creativity.

First Transition: The Earliest Signs of Creativity

The minds of our earliest ancestors, *Homo habilis*, have been referred to as *episodic* because there is no evidence that they deviated from the present moment of concrete sensations (Donald, 1991). They could encode perceptions of events in memory, and recall them in the presence of a cue, but had little voluntary access to memories without environmental cues. They were therefore unable to shape, modify, or practice skills and actions, and unable to invent or refine complex gestures or vocalizations.

Homo erectus lived between approximately 1.8 and 0.3 million years ago. The cranial capacity of the *Homo erectus* brain was approximately 1,000 cc, about 25% larger than that of *Homo habilis*, at least twice as large as that of living great apes, and 75% that of modern humans (Ruff et al., 1997). This period is widely referred to as the beginnings of cumulative culture. *Homo erectus* exhibited many indications of enhanced intelligence, creativity, and ability to adapt to their environment, including sophisticated, task-specific stone hand axes, complex stable seasonal home bases, and long-distance hunting strategies involving large game, and migration out of Africa.

This period marks the onset of the archaeological record and it is thought to be the beginnings of human culture. It is widely believed that this cultural transition reflects an underlying transition in cognitive or social abilities. Some have suggested that they owe their achievements to onset of *theory of mind* (Mithen, 1998) or the capacity to imitate (Dugatkin, 2001). However, there is evidence that other species possess theory of mind and the capacity to imitate (Heyes, 1998), yet do not compare to modern humans in intelligence and cultural complexity.

Evolutionary psychologists have suggested that the intelligence and cultural complexity of the *Homo* line is due to the onset of *massive modularity* (Buss, 1999, 2004; Barkow, Cosmides, &Tooby, 1992). However, although the mind exhibits an intermediate degree of functional and anatomical modularity, neuroscience has not revealed vast numbers of hardwired, encapsulated, task-specific modules; indeed, the brain has been shown to be more highly subject to environmental influence than was previously believed (Buller, 2005; Byrne, 2000; Wexler, 2006).

A Promising and Testable Hypothesis

Donald (1991) proposed that with the enlarged cranial capacity of *Homo erectus*, the human mind underwent the first of three transitions by which it—along with the cultural matrix in which it is embedded—evolved from the ancestral, pre-human condition. This transition is characterized by a shift from an *episodic* to a *mimetic mode* of cognitive functioning, made possible by onset of the capacity for voluntary retrieval of stored memories, independent of environmental cues. Donald refers to this as a *self-triggered recall and rehearsal loop*. Self-triggered recall enabled information to be processed recursively with respect to different contexts or perspectives. It allowed our ancestor to access memories voluntarily and thereby act out[1] events that occurred in the past or that might occur in the future. Thus not only could the mimetic mind temporarily escape the here and now, but by miming or gesture, it could communicate similar escapes in other minds. The capacity to mime thus ushered forth what is referred to as a *mimetic* form of cognition and brought about a transition to the mimetic stage of human culture. The self-triggered recall and rehearsal loop also enabled our ancestors to engage in a stream of thought. One thought or idea evokes another, revised version of it, which evokes yet another, and so forth recursively. In this way, attention is directed away from the external world toward one's internal model of it. Finally, self-triggered recall allowed for voluntary rehearsal and refinement of actions, enabling systematic evaluation and improvement of skills and motor acts.

Computational Model

Donald's hypothesis is difficult to test directly, for if correct it would leave no detectable trace. It is, however, possible to computationally model how the onset of the capacity for recursive recall would affect the effectiveness, diversity, and open-endedness of ideas generated in an artificial society. This section summarizes how we tested Donald's hypothesis using an agent-based computational model of culture referred to as 'EVOlution of Culture', abbreviated EVOC. Details of the modeling platform are provided elsewhere (Gabora, 2008b, 2008c; Gabora & Leijnen, 2009; Leijnen & Gabora, 2009).

[1] The term *mimetic* is derived from "mime," which means "to act out."

The EVOC World. EVOC uses neural network based agents that (i) invent new ideas, (ii) imitate actions implemented by neighbors, (iii) evaluate ideas, and (iv) implement successful ideas as actions. Invention works by modifying a previously learned action using learned trends (such as that more overall movement tends to be good) to bias the invention process. The process of finding a neighbor to imitate works through a form of lazy (non-greedy) search. An imitating agent randomly scans its neighbors, and adopts the first action that is fitter than the action it is currently implementing. If it does not find a neighbor that is executing a fitter action than its own action, it continues to execute the current action. Over successive rounds of invention and imitation, agents' actions improve. EVOC thus models how descent with modification occurs in a purely cultural context. Agents do not evolve in a biological sense–they neither die nor have offspring–but do in a cultural sense, by generating and sharing ideas for actions.

Following Holland (1975) we refer to the success of an action in the artificial world as its *fitness*, with the caveat that unlike its usage in biology, here the term is unrelated to number of offspring (or ideas derived from a given idea). The fitness function (FF) was originally chosen because it allows investigation of biological phenomena such as underdominance and epistasis in a cultural context (see Gabora, 1995); the one used here is one over several used in EVOC (see Gabora, 2008 for others). The FF rewards head immobility and symmetrical limb movement. Fitness of actions starts out low because initially all agents are entirely immobile. Soon some agent invents an action that has a higher fitness than doing nothing, and this action gets imitated, so fitness in- creases. Fitness increases further as other ideas get invented, assessed, implemented as actions, and spread through imitation. The diversity of actions initially increases due to the proliferation of new ideas, and then decreases as agents hone in on the fittest actions.

We used was a toroidal lattice with 100 nodes, each occupied by a single, stationary agent, and a von Neumann neighborhood structure (agents only interacted with their four adjacent neighbors). During invention, the probability of changing the position of any body part involved in an action was 1/6. On each run, creators and imitators were randomly dispersed.

Chaining. This gives agents the opportunity to execute multi-step actions. For the experiments reported here with chaining turned on, if in the first step of an action an agent was moving at least one of its arms, it executes a second step, which again involves up to six body parts. If, in the first step, the agent moved one arm in one direction, and in the second step it moved the same arm in the opposite direction, it has the opportunity to execute a three-step action. And so on. The agent is allowed to execute an arbitrarily long action so long as it continues to move the same arm in the opposite direction to the direction it moved previously. Once it does not do so, the chained action comes to an end. The longer it moves, the higher the fitness of this

multi-step chained action. Where n is the number of chained actions, the fitness, F_c, is calculated as follows:

$$F_c = F_{nc} + (n - 1)$$

The fitness function with chaining provides a simple means of simulating the capacity for recursive recall.

'Origins of Creativity' Results

As shown in Figure 1, the capacity to chain simple actions into more complex ones increases the mean fitness of actions in the society. This is most evident in the later phase of a run. Without chaining, agents converge on optimal actions, and the mean fitness of action reaches a plateau. With chaining, however, there is no ceiling on the mean fitness of actions. By the 100th iteration it reached almost 15, indicating a high incidence of chaining.

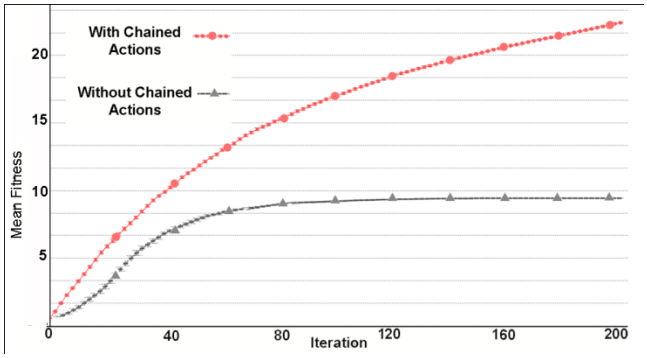


Figure 1. Mean fitness of actions in the artificial society with chaining versus without chaining.

As shown in Figure 2, chaining also increases the diversity of actions. This is most evident in the early phase of a run before agents begin to converge on optimal actions. Although in both cases there is convergence on optimal actions, without chained actions, this is a static set (thus mean fitness plateaus) whereas with chained actions the set of optimal actions is always changing, as increasingly fit actions are found (thus mean fitness keeps increasing).

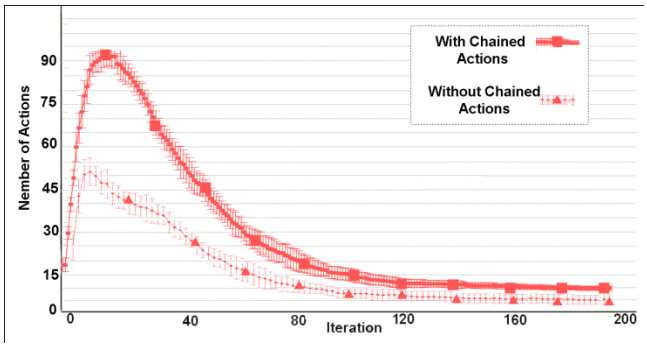


Figure 2. Mean number of different actions in the artificial society with chaining (continuous line) versus without chaining (dashed line).

Recall that agents can learn trends from past experiences (using the knowledge-based operators), and thereby bias the generation of novelty in directions that have a greater than chance probability of being fruitful. Since chaining provides more opportunities to capitalize on the capacity to learn, we hypothesized that chaining would accentuate the impact of learning on the mean fitness of actions, and this too turned out to be the case (Gabora & Saberi, 2011).

Second Transition: The 'Big Bang' of Human Creativity

The European archaeological record indicates that a truly unparalleled cultural transition occurred between 60,000 and 30,000 years ago at the onset of the Upper Paleolithic (Bar-Yosef, 1994; Klein, 1989; Mellars, 1973, 1989a, 1989b; Soffer, 1994; Stringer & Gamble, 1993). Considering it "evidence of the modern human mind at work," Richard Leakey (1984:93-94) describes the Upper Palaeolithic as follows: "unlike previous eras, when stasis dominated, ... [with] change being measured in millennia rather than hundreds of millennia." Similarly, Mithen (1996) refers to the Upper Paleaolithic as the 'big bang' of human culture, exhibiting more innovation than in the previous six million years of human evolution. We see the more or less simultaneous appearance of traits considered diagnostic of behavioral modernity. It marks the beginning of a more organized, strategic, season-specific style of hunting involving specific animals at specific sites, elaborate burial sites indicative of ritual and religion, evidence of dance, magic, and totemism, the colonization of Australia, and replacement of Levallois tool technology by blade cores in the Near East. In Europe, complex hearths and many forms of art appeared, including cave paintings of animals, decorated tools and pottery, bone and antler tools with engraved designs, ivory statues of animals and sea shells, and personal decoration such as beads, pendants, and perforated animal teeth, many of which may have indicated social status (White, 1989a, 1989b).

Whether this period was a genuine revolution culminating in behavioral modernity is hotly debated because claims to this effect are based on the European Palaeolithic record, and largely exclude the African record (McBrearty & Brooks, 2000); Henshilwood & Marean, 2003). Indeed, most of the artifacts associated with a rapid transition to behavioral modernity at 40–50,000 years ago in Europe are found in the African Middle Stone Age tens of thousands of years earlier. However the traditional and currently dominant view is that modern behavior appeared in Africa between 50,000 and 40,000 years ago due to biologically evolved cognitive advantages, and spread replacing existing species, including the Neanderthals in Europe (e.g., Ambrose, 1998; Gamble, 1994; Klein, 2003; Stringer & Gamble, 1993). Thus from this point onward there was only one hominid species: modern *Homo sapien*. Despite lack of overall increase in cranial capacity, the prefrontal

cortex, and particularly the orbitofrontal region, increased significantly in size (Deacon, 1997; Dunbar, 1993; Jerison, 1973; Krasnegor, Lyon, and Goldman-Rakic, 1997; Rumbaugh, 1997) and it was likely a time of major neural reorganization (Klein, 1999; Henshilwood, d'Errico, Vanhaeren, van Niekerk, and Jacobs, 2000; Pinker, 2002).

Given that the Middle/Upper Palaeolithic was a period of unprecedented creativity, what kind of cognitive processes may have been involved?

A Testable Hypothesis

Converging evidence suggests that creativity involves the capacity to shift between two forms of thought (Finke, Ward, & Smith, 1992; Gabora, 2003; Howard-Jones & Murray, 2003; Martindale, 1995; Smith, Ward, & Finke, 1995; Ward, Smith, & Finke, 1999). Divergent or associative processes are hypothesized to occur during idea generation, while convergent or analytic processes predominate during the refinement, implementation, and testing of an idea. It has been proposed that the Paleolithic transition reflects a mutation to the genes involved in the fine-tuning of the biochemical mechanisms underlying the capacity to subconsciously shift between these modes, depending on the situation, by varying the specificity of the activated cognitive receptive field. This is referred to as *contextual focus*[2] because it requires the ability to focus or defocus attention in response to the context or situation one is in. Defocused attention, by diffusely activating a broad region of memory, is conducive to divergent thought; it enables obscure (but potentially relevant) aspects of the situation thus come into play. Focused attention is conducive to convergent thought; memory activation is constrained enough to hone in and perform logical mental operations on the most clearly relevant aspects.

Support from Computational Model

Again, because it would be difficult to empirically determine whether Paleolithic humans became capable of contextual focus, we decided to begin by determining whether the hypothesis is at least computational feasible. To do so we used an evolutionary art system that generated progressively evolving sequences of artistic portraits, with no human intervention once initiated. We sought to determine whether incorporating contextual focus into the fitness function would play a crucial role in enabling the computer system to generate art that humans find "creative" (i.e. possessing qualities of novelty and aesthetic value typically ascribed to the output of a creative artistic process).

We implemented contextual focus in the evolutionary art algorithm by giving the program the capacity to vary its level of fluidity and control over different phases of the creative process in response to the output it generated. The creative domain of portrait painting was chosen because it requires both focused attention and analytical thought to accomplish the primary goal of creating a resemblance to the portrait sitter, as well as defocused attention and associative thought to deviate from resemblance in a way that is uniquely interesting, *i.e.*, to meet the broad and often conflicting criteria of aesthetic art. Since judging creative art is subjective, rather than use quantitative analysis, a representative subset of the automatically produced artwork from this system was selected, output to high quality framed images, and submitted to peer reviewed and commissioned art shows, thereby allowing it to be judged positively or negatively as creative by human art curators, reviewers and the art gallery going public.

Our strategy for modeling contextual focus may raise questions about the ability of computers to "truly" be creative, and the role of the human system designer in the creative output. Several researchers in computational creativity, have addressed such questions by outlining different dimensions of creativity and proposing schema for evaluating a "level of creativity" of a given system, for example (Ritchie, 2007; Jennings, 2010; Colton, Pease, & Charnley, 2011). We are interested in applying such analyses to our portrait-system as a possibility for future work; indeed, the mechanics of contextual focus might be clarified by the computational creativity literature. In particular we are interested in further exploring the link between system-modified fitness constraints and the idea of transformational creativity (Boden, 2003; Wiggins 2006).

However, for the purposes of the current paper, it is less important to address the question of designer involvement in system creativity, or to try and quantify the amount of creativity displayed. Rather, we concentrate on the qualitative impact made by the explicit incorporation of contextual focus into the system as a whole, and its ability to elevate the perceived quality and novelty of system output to a level audiences judged reminiscent of successful "artistic, human-style" creativity.

Generative Art Systems: Creative evolutionary systems are a class of search algorithms inspired by Darwinian evolution, the most popular of which are genetic algorithms (GA) and genetic programming (GP) (Koza, 1993). These techniques solve complex problems by encoding a population of randomly generated potential solutions as 'genetic instruction sets', assessing the ability of each to solve the problem using a predefined fitness function, mutating and/or marrying (applying crossover to) the best to yield a new generation, and repeating until one of the offspring yields an acceptable solution. We are not claiming that contextual focus is Darwinian, but simply that for our computational modeling purposes, Genetic Programming proved a convenient foundational aggregator to support our contextual focus fitness function module.

Typically these systems allow a human user to pick those individuals that will be mated – making the human the creative judge. In contrast, our system used a function trigger mechanism within the contextual focus fitness func-

[2] In neural net terms, contextual focus amounts to the capacity to spontaneously vary the shape of the activation function, flat for divergent thought and spiky for analytical.

tion which allowed the process to run automatically, without any human intervention once the process was started. It was not until the evolutionary art process came to completion that humans looked at and evaluated the art. Others have begun to use creative evolutionary systems with an automatic fitness function in design and music, as well as in a creative invention machine (Bentley, Corne, 2002). What is unique in our approach is that it incorporates several techniques that enable it to shift to processing artistic content in a more divergent or associative manner, and employs a form of GP called Cartesian Genetic Programming (Miller, 2011), detailed in the next section.

Implementation: The GP function set has 13 functions which use unitized *x* and *y* positions of the portrait image as variables and additional parameter variables (noted PM) that can be affected by adaptive mutation. Functions are low level in nature which aids in a large 'creative' search space, and output HSV color space values between 0 and 255. An individual in our population is manifested as one program that runs successively for every pixel in the output image, which is then tested against our creative fitness function. This allows correlated painterly effects as one moves through the image. Functions 1 through 5 use simple logical or arithmetic manipulations of the positions (low level functions create a larger 'creative' search space), whereas 7 through 14 use trigonometric or logical functions that are more related to geometric shapes and color graduations. The 13 functions of the function set are:

```
1: x|y;
2: PM & x;
3: (x ? y) % 255;
4: if (x[y) x - y; else y - x;
5: 255 - x;
6: abs (cos (x) * 255);
7: abs (tan (((x % 45) * pi)/180.0) * 255));
8: abs (tan (x) * 255) % 255);
9: sqrt ((x - PM)2 ? (y - PM) 2); (thresholded at 255)
10: x % (PM? 1) ? (255 - PM);
11: (x ? y)/2;
12: if (x[y)255*((y ? 1)/(x ? 1)); else 255*((x ? 1)/(y ? 1));
13: abs (sqrt (x – PM2? y – PM2) % 255);
```

The contextual focus based fitness function varies fluidly from tightly focusing on resemblance (similarity to the sitter image, which in this case is an image of Charles Darwin), to swinging (based on functional triggers) toward a more associative process of the intertwining, and at times contradicting, 'rules' of abstract portrait painting. Different genotypes map to the same phenotype. This allows us to vary the degree of creative fluidity because it offers the capacity to move though the search space via genotype (small ordered movement) or phenotype (large movement but still related). For example, in one set of experiments this is implemented as follows: if the fittest individual of a population is identical to an individual in the previous generation for more than three iterations, meaning the algorithm is stuck in analytic mode and needs to open up, other genotypes that map to this same phenotype are chosen over the current non-progressing genotype, allowing divergent open movement through the landscape of possibilities.

The automatic fitness function partly uses a 'portrait to sitter' resemblance. Since the advent of photography (and earlier), portrait painting has not just been about accurate reproduction, but also about using modern painterly goals to achieve a creative representation of the sitter. The fitness function primarily rewards accurate representation, but in certain situations also rewards visual painterly aesthetics using simple rules of art creation as well as a portrait knowledge space. Specifically, the divergent painterly portion of the fitness function takes into account: (1) face versus background composition, (2) tonal similarity over exact color similarity, matched with a sophisticated artistic color space model that weighs for warm-cool color temperature relationships based on analogous and complementary color harmony rules, and (3) unequal dominate and subdominant tone and color rules, and other artistic rules based on a portrait painter knowledge domain as detailed in (DiPaola, 2009) and illustrated in Figure 3. The system is biased toward resemblance, which gives it structure, but can, under the influence of functional triggers, exhibit artistic flair.

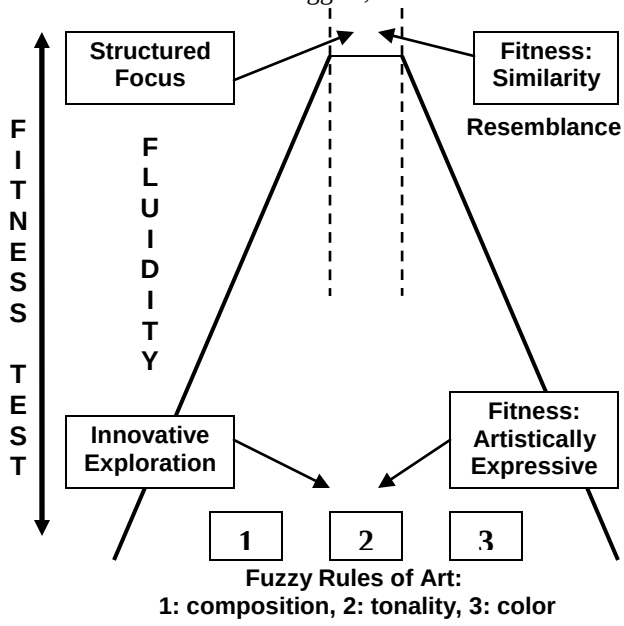


Figure 3. The contextual focus fitness function mimics human creativity by moving between restrained focus (resemblance) to more unstructured associative focus (resemblance + ambiguous art rules of composition, tonality and color theory).

The fitness function calculates four scores (resemblance and the three painterly rules), and then combines them in different ways to mimic human creativity, shifting between unstructured associative focus (rules of composition, tonality and color theory) and restrained focus (resemblance). In its default state, the fitness function uses a more analytic

form of processing, specifically, a ratio of 80% resemblance to 20% non-proportional scoring of the three painterly rules. Several functional triggers can alter this ratio in different ways, but the main trigger is when the system is "stuck". Within any run, for instance as long as an adaptive percentage of 80–20 resemblance bias is maintained (resemblance patriarchs), the system will allow very high scoring of painterly rule individuals to be accepted into the next population. Those with high painterly scores (weighted non-proportionally including for a very high score with respect to just one rule) are saved separately, and mated with the current 80/20 population. Unless other triggers exist, their offspring are still tested with the 80–20 resemblance test. System wide functional changes occur when redundancy triggers affect the default ratio for all individuals. As mentioned previously, when a plateau or local minimum is reached for a certain number of populations, the fitness function ratio switches such that painterly rules are weighted higher than resemblance (on a sliding scale), and work in conjunction with redundancy at the input, node, and functional levels. Similarly, but now in reverse, to the default resemblance situation, high scoring resemblance individuals can pass into the next population when a percentage of painterly rule individuals is met. Using this more associative mode, high resemblance individuals are always part of the mix, and when these individuals show a marked improvement, a trigger is set to return to the more focused 80/20 resemblance ratio.

As the fitness score increases, portraits look more like the sitter. This gives us a somewhat known spread from very primitive (abstract) all the way through to realistic portraits. Thus in effect the system has two ongoing processes: (1) those 'most fit' portraits that pass on their portrait resemblance strategies, making for more and more realistic portraits—the family 'resemblance' patriarchs, and (2) the creative 'strange uncles': related to the current 'resemblance fit', but portraits that are more artistically creative or 'artistically fit'. This dual evolving technique of 'patriarchs and strange uncles' mimics the interplay between freedom and constraint that is so central to creativity. Paradoxically, novelty often benefits from the existence of a known framework reference system to rebel and innovate from. Creative people use some strong structural rules (as in the templates of a sonnet, tragedy, or in this case, a resemblance to the sitter image) as a resource or base to elaborate new variants beyond that structure (in this case, an abstracted variation of the sitter image).

'Big Bang of Creativity' Results

The automatic creative output was generated over thirty days of continuous, un-supervised computer use. The images in Figure 4 show a selection of representative portraits produced by the system. While the overall population improves at resembling Darwin's portrait, what is more interesting to us is the variety of recurring, emergent and merged creative strategies that evolve as the programs in different ways to become better abstract portraitists.

Figure 4. These images have been seen by thousands in the last 2 years and have been perceived as creative art works on their own by the art public, including above at the MIT Museum in Cambridge, MA.

Humans rated the portraits produced by this version of the portrait painting program with contextual focus as much more creative and interesting than a previous version that did not use contextual focus, and unlike its predecessor, the output of this program generated public attention worldwide. Example pieces were framed and submitted to galleries as a related set of work. Care was taken by the author to select representational images of the evolved unsupervised process, however creative human bias obvious exists in the representational editing process. Output has been accepted and exhibited at six major galleries and museums including the TenderPixel Gallery in London, Emily Carr Galley in Vancouver, and Kings Art Centre at Cambridge University as well as the MIT Museum, and the High Museum in Atlanta, all either peer reviewed, juried or commissioned shows from institutions that typically only accept human art work. A typical gallery installation consisted of 40-70 related portraits produced in time order over a given run. Gallery showings focus on "best resemblances" and those that are artistically compelling from an abstract portrait perspective. This gallery of work has been seen by tens of thousands of viewers who have commented that they see the artwork as an aesthetic piece that ebb and flows through seemly creative ideas even though they were solely created by an evolutionary art computer program using contextual focus. Note that no attempt to create a pure 'creativity Turning Test' was attempted. Besides the issues surrounding the validity of such a test (Pease, Colton, 2011), it was not feasible in such reputable and large art venues. However most of the thousands of causal viewers assumed they were looking at human created art. The work was also selected for its aesthetic value to accompany an opinion piece in the journal Nature (Padian, 2008), and was given a strong critical review by the Harvard humani-

ties critic, Browne (2009). While these are subjective measures, they are standard in the art world. The fact that the computer program produced novel creative artifacts, both as single art pieces and as a gallery collection of pieces with interrelated themes, using contextual focus as a key element of its functioning, is compelling evidence of the effectiveness of contextual focus.

Discussion and Conclusions

Many species engage in acts that could be said to be creative. However, humans are unique in that our creative ideas build on each other cumulatively; indeed it is for this reason that culture is widely construed as an evolutionary process (e.g. Bentley, Ormerod, & Batty, 2011; Cavalli-Sforza & Feldman, 1981; Gabora, 1996, 2008; Hartley, 2009; Mesoudi, Whiten & Laland, 2006; Whiten, Hinde, Laland, & Stringer, 2011). Our creativity is evident in all walks of life. It has transformed the planet we live on.

We discussed two transitions in the evolution of uniquely cumulative form of creativity, discussed cognitive mechanisms that have been proposed to underlie these transitions, and summarized efforts to computationally simulate them. Using an agent based computer model of cultural evolution, we obtained support for the hypothesis that the onset of cumulative, open-ended cultural evolution can be attributed to the evolution of a self-triggered recall and rehearsal loop, enabling the recursive chaining of thoughts and actions. Using a generative genetic programming system, we used a computational model of contextual focus to automatically produce a related series of art output that received critical acclaim usually given to human art work supporting the hypothesis that the capacity to shift between analytic and associative modes of thought plays an important role in the creative process.

Our results suggest that the evolution of chaining and contextual focus made possible the open-ended cumulative creativity exhibited by computational models of language evolution (e.g. Kirby, 2001). Note that in chaining versus no chaining conditions the size of the neural network is the same, but how it is used differs. This suggests that it was not larger brain size *per se* that initiated the onset of cumulative culture, but that larger brain size enabled episodes to be encoded in more detail, allowing more routes for reminding and recall, thereby facilitating recursive redescription of information encoded in memory (Karmiloff-Smith, 1992), thereby tailor it to the situation at hand. Our results suggest that it is reasonable to hypothesize that this in turn is vastly accentuated by the capacity to shift between associative and analytic different processing modes.

We wish to acknowledge some limitations of this work. Chaining does not work, as in humans, by considering an idea in light of one perspective, seeing how that perspective modifies the idea, seeing how this modification suggests a new perspective from which to consider the idea, and so forth. We are planning a more sophisticated implementation of that works more along these lines. Second, there is some irony in using an art program based on the genetic algorithm as a starting point to implement contextual focus, which we have claimed is unique to the cultural evolution of ideas and has no counterpart in biological evolution. Our goal here was to see if contextual focus 'works' at all; since this was successful, we will now move on to more cognitively plausible implementations. One of the projects currently underway is to implement contextual focus in the EVOC model of cultural evolution that was used for the 'origin of creativity' experiments. This is being carried out as follows. The fitness function will change periodically, so that agents find themselves no longer performing well. They will be able to detect that they are not performing well, and in response, increase the probability of change to any component of a given action. This temporarily makes them more likely to "jump out of a rut" resulting in a very different action, thereby simulating the capacity to shift to a more associative form of thinking. Once their performance starts to improve, the probability of change to any component of a given action will start to decrease to base level, making them less likely to shift to a dramatically different action. This helps them perfect the action they have already settled upon, thereby simulating the capacity to shift to a more associative form of thinking.

Acknowledgements

We are grateful to Graeme McCaig and grants from *Natural Sciences and Engineering Research Council of Canada* and the Fund for Scientific Research of Flanders, Belgium.

References

Barkow, J. H., Cosmides, L., & Tooby, J., Eds. 1992. *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. New York: Oxford University Press.

Bentley, P., D. Corne D., Eds. 2002. *Creative Evolutionary Systems*, Morgan Kaufmann, San Francisco.

Bentley, R. A., Ormerod, P., & Batty, M. 2011. Evolving social influence in large populations. *Behavioral Ecology and Sociobiology*, 65:537–546.

Boden, M. 2003. *The Creative Mind: Myths and Mechanisms (second edition).* Routledge.

Brown, J. 2009. Looking at Darwin: portraits and the making of an icon. *Isis*. Sept, 100(3):542–70.

Buller, D. J. 2005. *Adapting minds*. MIT Press.

Buss, D. M. 1999/2004. *Evolutionary Psychology: The new science of the mind*. Boston, MA: Pearson.

Byrne, R., & Russon, A. 1998. Learning by imitation: A hierarchical approach. *Behav Brain Sciences*, 2:667–721.

Cavalli-Sforza, L. L., & Feldman, M. W. 1981. *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton: Princeton University Press.

Cloak, F. T. Jr. 1975. Is a cultural ethology possible? *Human Ecology*, 3:161–182.

DiPaola, S. & Gabora, L. 2007. Incorporating characteristics of human creativity into an evolutionary art algorithm. In (D. Thierens, Ed.), *Proc Genetic and Evol Computing Conf* (pp. 2442–2449), July 7-11, Univ College London.

DiPaola S, 2009. "Exploring a Parameterized Portrait Painting Space*", International Journal of Art and Technology*, 2(1-2):82–93.

DiPaola, S. & Gabora, L. 2009. Incorporating characteristics of human creativity into an evolutionary art algorithm. *Genet Prog and Evolvable Machines*, 10(2):97–110.

Donald, M. 1991. *Origins of the Modern Mind: Three Stages in the Evolution of Culture and Cognition*. Cambridge, MA: Harvard University Press.

Dugatkin, L. 2001. A. *Imitation Factor: Imitation in Animals and the Origin of Human Culture.* Free Press.

Gabora, L. 1995. Meme and Variations: A computer model of cultural evolution. In L. Nadel & D. Stein (Eds.) 1993 *Lectures in Complex Systems*. Addison-Wesley, 471−486.

Gabora, L. 1996. A day in the life of a meme. *Philosophica*, 57:901–938.

Gabora, L. 1999. Conceptual closure: Weaving memories into an interconnected worldview. In (G. Van de Vijver & J. Chandler, Eds.) *Proc Closure: Intl Conf Emergent Organizations and Dynamics.* May 3-5, Univ Gent, Belgium.

Gabora, L. 2008a. EVOC: A computer model of cultural evolution. In V. Sloutsky, B. Love & K. McRae (Eds.), *Proc Ann Mtng Cog Sci Soc*, Sheridan Publ, 1466–1471.

Gabora, L. Modeling cultural dynamics. 2008b. *In Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) Fall Symposium 1: Adaptive Agents in a Cultural Context*, AAAI Press, 18–25.

Gabora, L. & Aerts, D. 2009. A mathematical model of the emergence of an integrated worldview. *Journal of Mathematical Psychology*, 53:434-451.

Gabora, L. & Leijnen, S. 2009. How creative should creators be to optimize the evolution of ideas? A computational model. *Electronic Proc Theor Comp Sci*, 9*:*108–119.

Gabora, L. & Saberi, M. 2011. How did human creativity arise? An agent-based model of the origin of cumulative open-ended cultural evolution. *Proceedings:ACM Conference on Cognition & Creativity*, 299-306. Atlanta, GA.

Hartley, J. 2009. From cultural studies to cultural science. *Cultural Science*, 2:1–16.

Higgs, P. 2000. The mimetic transition: a simulation study of the evolution of learning by imitation. *Proceedings: Royal Society B: Biological Sciences*, 267:1355–1361.

Heyes, C. M. 1998. Theory of mind in nonhuman primates. *Behavioral and Brain Sciences, 211*:104–134.

Hinton, G. E. & Nowlan, S. J. 1987. How learning can guide evolution. *Complex Systems*, 1:495–502.

Holland, J. K. 1975. *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: University of Michigan Press.

Hutchins, E. & Hazelhurst, B. 1991. Learning in the cultural process. In Langton, C., Taylor, J., Farmer, D., &Rasmussen, S. (Eds.) *Artificial Life II*. Redwood City, CA: Addison-Wesley.

Jennings, K. E. 2010. Developing creativity: Artificial barriers in artificial intelligence. *Minds and Machines*, 1–13.

Kirby, S. 2001. Spontaneous evolution of linguistic Structure—An iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2):102–110.

Koza, J, 1993. *Genetic Programming*, MIT Press.

Leijnen, S. & Gabora, L. 2010. An agent-based simulation of the effectiveness of creative leadership. *Proc Ann Mtng Cog Sci Soc* 955–960. Aug 11-14, Portland, OR.

Mesoudi, A., Whiten, A., & Laland, K. 2006. Toward a unified science of cultural evolution. *Behavioral and Brain Sciences*, 29:329–383.

Miller, J. 2011. *Cartesian Genetic Programming*, Springer.

Mithen, S. Ed. 1998. *Creativity in Human Evolution and Prehistory*. London, UK: Routledge.

Padian, K. 2008. Darwin's Enduring Legacy, *Nature*, 451:632–634.

Pease, A. and Colton, S. 2011. 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*.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17.

Colton, S., Pease, A, Charnley. J. 2011 Computational creativity theory: The FACE and IDEA descriptive models. *In 2nd Int Conference on Computational Creativity.*

Ruff, C., Trinkaus, E., & Holliday, T. 1997. Body mass and encephalization in Pleistocene Homo. *Nature*, 387:173–176.

Tomasello, M., Kruger, A., Ratner, H. 1993. Cultural learning. *Behavioral and Brain Sciences*, 16:495–552.

Wexler, B. 2006. *Brain and Culture: Neurobiology, Ideology and Social Change*. New York: Bradford Books.

Whiten, A., Hinde, R., Laland, K., Stringer, C. 2011. Culture evolves. 2011. *Philosophical Transactions of the Royal Society B*, 366:938–948.

Wiggins, G. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19(7):449-458.

Creatively Subverting Messages in Posters

Lorenzo Gatti[1], Marco Guerini[1], Charles Callaway[1], Oliviero Stock[2], Carlo Strapparava[2]
[1]Trento-Rise, [2]FBK-Irst
Via Sommarive 18, Povo, Trento - Italy
{l.gatti, marco.guerini, c.callaway}@trentorise.eu, {stock, strappa}@fbk.eu

Abstract

Creativity is widely used in advertisements, and is meant to be appreciated by people. However creativity can also be used as a defense. When we walk in the street we are overwhelmed by messages which try to get our attention with any persuasive device at hand. As messages get ever more aggressive, often our basic cognitive defense – trying not to perceive those messages – is not sufficient. One advanced defensive technique is based on transforming the perceived message into something different (for instance irony or hyperbole) from what was originally meant in the message. In this paper we describe an implemented application for smartphones that creatively modifies the linguistic expression in a virtual copy of the poster. The mobile system is inspired by the subvertising practice of countercultural art, and aims at experiencing aesthetic pleasure that relaxes the cognitive tension of the user.

Introduction

We are surrounded by linguistic expressions on the walls around us. Whenever we walk along a street, posters, signs and other similar advertisements are there trying to attract our attention and in most cases trying to influence our actions, beliefs and behavior. We may try to avoid those ads but it is not easy: even if the characteristics of our perceptive and cognitive system partially help us in being "banner blind" (Pagendarm and Schaumburg 2001; Burke, Gorman, Nilsen and Hornof 2004), pervasive advertising often manages to overcome our barriers (Müller, Alt and Michelis 2011). One strategy to counter messages that forcefully grab our attention is to use our cognitive system to fight back and creatively alter the advertising message itself. This form of "reactive" creativity lies at the root of various phenomena, including some aspects of verbal humor, especially irony. The psychoanalytical approach to humor (Freud 1905), gives an attractive account of the release of energy that results from overcoming our inner censors through the appreciation of humorous expressions. A similarly liberating process can be attributed to other types of variations of linguistic expressions.

From an aesthetic point of view a given variation is more highly appreciated if the change is limited, as for instance suggested by the optimal innovation theory (Giora 2003).

When we humans entertain this creative, reactive modality to defend ourselves, we tend to intervene within our mind. Sometimes people even intervene on the physical object itself, the classic example being the poster, writing over it to correct an expression (or even add graphic symbols to images, such as moustaches added to a face). In countercultural art this is called *subvertising*.

As far as current technology is concerned, a lot of attention is being devoted to figuring out how to exploit smartphones for advertising. Here instead we propose a mainly defensive goal on behalf of the consumer. The aim is to exploit technology for producing linguistic expressions that slightly change the observed advertisement. The goal is to accommodate a message that is biased in a rather different direction. The system produces a new virtual poster so as to help the user relax the cognitive tension produced by the unduly attention-grabbing original message.

In particular we have developed a mobile application that allows users to take a picture of a poster, and then automatically produces a new virtual version with the same layout and visual aspect of the poster, but with a creative variation of the linguistic expression it originally expressed.

In our current prototype the user merely needs to point the camera of the smartphone at the poster, and the image, with the same appearance but the altered linguistic expression, is produced in a few short steps.

An image analysis and reconstruction component takes care of the graphic aspects, and an underlying program is called to obtain the actual variation of the given expression, which can have several different realizations. In this paper we utilize just one of the functionalities of VALENTINO, an affective valence shifting program (Guerini, Strapparava and Stock 2011); the creativity involved in the process is a necessary element for the successful impact of this defensive tool.

Background and relevant Work

The word *subvertising* is a portmanteau of the words "subvert" and "advertising". Subvertising refers to the practice of making spoofs or parodies of corporate and political advertisements in order to make a statement. This can take the form of a new image, an alteration to an existing im-

age, or a modification/re-contextualization of an existing slogan (sometimes called a "meme hack").

According to AdBusters, a Canadian magazine that is a leading proponent of counter-culture and subvertising, "A well produced 'subvert' mimics the look and feel of the targeted ad, promoting the classic 'double-take' as viewers suddenly realize they have been duped. Subverts create cognitive dissonance."

In our work we focus on the creative textual modification task of subvertising. In particular, we want to implement a defense strategy for making the user aware of the subtle presuppositions implicit in advertising messages - by using *exaggeration* (or hyperbole) of the affective content of the message. The main resource used to implement such a defensive strategy is the VALENTINO prototype, a tool for affective modification of existing texts.

Affective variations of pre-existing texts have been studied and implemented in various domains, see for example (Mateas, Vanouse and Domike 2000; Guerini, Strapparava and Stock 2008b, 2011), or similarly funny variations (Stock and Strapparava 2003). The effectiveness of affective variations has also been assessed; in particular, Van Der Sluis and Mellish's (2010) evaluation shows that biased variations of a message work better than the neutral condition. With regard to output quality, Whitehead and Cavedon (2010) demonstrated that adding bigram frequencies for the insertion of valenced modifiers (chosen according according to MAX function) significantly improve the perceived quality of the resulting texts.

Valentino

VALENTINO can modify existing textual expressions towards more positively or negatively valenced versions, given a numeric coefficient that represents the desired valence shifting for the final expression.

Since the system works in an open domain and without lexical restrictions, VALENTINO's linguistic resources are general purpose, and automatically built from large scale corpora and English lexical repositories.

For the task of modifying single words, we automatically built a resource that gathers these terms in vectors (Ordered Vectors of Valenced Terms - OVVTs). We used the WordNet *antonymy* relation as an indicator of terms that can be "graded", and built four groups of terms that can be used (one group for each POS). Moreover, we populated the vectors using other specific WordNet semantic relations (the *similar_to* relation for adjectives, *hyponym* relation for verbs and nouns). Finally the valence of WordNet synsets, taken from SentiWordNet scores (Esuli and Sebastiani 2006), was added to the corresponding lemmata. An example OVVT for the antonymy pair (ugly ↔ beautiful) ordered from most negative to most positive is:

(hideous ... ugly ... unnatural) ↔ (pretty ... beautiful ... gorgeous)

For insertion or deletion of words that play the role of downtoners or intensifiers we created specific OVVTs (which we call Modifier-OVVTs). In this case the words were gathered according to a criterion of contextual, rather than semantic, connection: we used the Google Web 1T 5-Grams Corpus (Brants and Franz 2006) to extract information about co-occurrences. In particular we created resources connecting terms with their modifiers (according to POS), thus obtaining adjective modifiers for nouns, and adverb modifiers both for adjectives and verbs. An example Modifier-OVVT for the term "dish", ordered from most negative to most positive, is:

(disgusting ... mediocre ... tasty ... delicious ... exquisite).

Strategies

We undertook a preliminary qualitative study with human subjects, to understand how people modify the valence of existing texts. The insights gained showed that: (a) people usually modify single words, (b) sometimes add or subtract words that play the role of downtoners or intensifiers and (c) sometimes use paraphrases (Guerini, Strapparava and Stock 2008b).

As a first step VALENTINO performs POS tagging, named entity recognition, morphological analysis and chunking of the existing constituents (NPs, VPs, ADJPs, and so on). This task exploits the TextPro package (Pianta, Girardi and Zanoli 2008). Subsequently the strategies described in points a), b), and c) above are applied to the chunks, following some general guidelines.

Minimal variation: texts (chunks) are slanted as much as needed, but the target score should not be exceeded, limiting the variation as much as possible.

Modification of dependents: A constituent is modified considering first the dependents (from left to right) and then possibly the head. Consider the very positive and the slightly negative variations of the following sentence:

" We ate [a very good dish]$_{NP}$"
"We ate [an incredibly delicious dish]$_{NP}$" (+)
"We ate [a good dish]$_{NP}$" (-)

The rationale is that in a constituent the element that bears the greatest part of the meaning is in the head, and it decreases the further we move into the constituent.

Candidates Selection: The selection of substitutes is a two-step process. Given a term to be modified (e.g. "good" in the example) there can be various candidates for the modification.

- The first step requires filtering out all the terms that do not meet the target score. For example if the target score is higher than +0.5, all terms from -1 to +0.5 are discarded. Further possible constraints can be taken into account (e.g. if the reasoning is about "**good** dish" then only the *similar_to* "good" that co-occur with "dish", and with score > 0.5, should be kept).
- Various strategies can then be used for choosing the best candidate: word persuasive impact (Guerini, Strapparava and Stock 2008a) word or n-gram frequency (Whitehead and Cavedon 2010), mutual-information, etc. Currently, the most used measure in VALENTINO is pointwise mutual information score, which yields modifiers specialized for the given term (e.g. "delicious" co-occurs less fre-

quently than "nice" with "dish", but it is more specialized in this context).

As for metrics that help decide the best quality lexical choice, while we have converged so far on the *best mutual information measure*, we think in different situations different measures should be applied (although this is outside the scope of this paper). Furthermore specific n-gram patterns – see for example (Veale 2011) - for extracting *semantically* exaggerated variations are under development.

In the present scenario, the critical choice was deciding the suitable degree of the *affective* modification amongst those proposed by VALENTINO; i.e. which one is the best for obtaining a defensive effect. In fact, light modifications usually obtain the effect of strengthening the message, while stronger ones can weaken it (Guerini, Strapparava and Stock 2012). Obviously strengthening the message is not the aim of the present tool, which is why we chose maximum target scores for the *affective exaggeration* strategy in subvertising.

Figure 1

Figure 2

Interface

We have implemented SUBVERTISER, a mobile subvertising tool that allows a user to photograph an advertisement they see, select the text they wish to change. The system then replaces that text with a valenced version created by VALENTINO. SUBVERTISER tries to match the font face, size and color of the new slogan to the text in the original image in order to heighten the effect of presenting a message that subverts the original.

SUBVERTISER requires very little interaction on the part of the user: once a photo of a printed ad (e.g., poster, billboard or banner) is taken with the phone, only a few steps are required to swap the original advertising message with a valenced version: selecting the region of the text in the photo to replace, correcting the text after scanning by in-built OCR, and selecting his or her preferred new version from of a list suggested by VALENTINO.

In a typical scenario, the user is walking with friends in a city, perhaps shopping or going to see a movie. When he is interrupted by a poster advertisement that bothers him, he uses his phone to take a snapshot of it, modifies it with SUBVERTISER, and can then show the new ad to his friends.

Algorithm

Behind the scenes, SUBVERTISER performs a number of steps to process both the language and the image of the advertisement. Given the photo taken by the user with the phone's camera (the image can also be chosen from a pre-existing library of images), the user selects the text region he wants to change containing the advertising message by moving and resizing a selection rectangle (Figure 1).

The image area is then passed to an OCR application on the smartphone itself, which scans for text within that rectangle[1]. The OCR both detects the coordinates of the bounding boxes for every individual word as well as returns the recognized text string of the message.

From the bounding box information we obtain the rectangle containing the first line of text, which is then scaled down to 100 pixels in height, and uploaded to an online third-party (multi-step) font recognition service[2] using dedicated APIs. Meanwhile, the program applies an inpainting algorithm to each bounding box in the original text zone. This step reconstructs the background image that was underneath the original text, providing a blank background where new text can be written (Figure 2).

The user is then asked to correct OCR errors, which if left unchanged would lead to linguistic errors in the valenced text, via a text entry box on the smartphone. VALENTINO is queried with the corrected OCR text string, and four valenced sentences are returned, from the most positive to the most negative, and presented to the user (Figure 3) to choose from. Once we know the original text, we also send that information to the font recognition server, which needs to align known letters with the image in order to determine the font and then respond with that information.

Once the user selects one of the slanted messages, an algorithm decides how to divide the slanted text into lines, since VALENTINO typically changes the number of words in the sentence.

Then, we ask the online font service to generate a new image with the detected font and a transparent background,

[1] As much processing as possible is done directly on the mobile to avoid excessive bandwidth usage and associated costs, and to lower the needed time to complete the task. Image processing is done with the OpenCV library, while character recognition is provided by the Tesseract OCR engine, which are both open source.

[2] URL: www.myfonts.com/WhatTheFont/

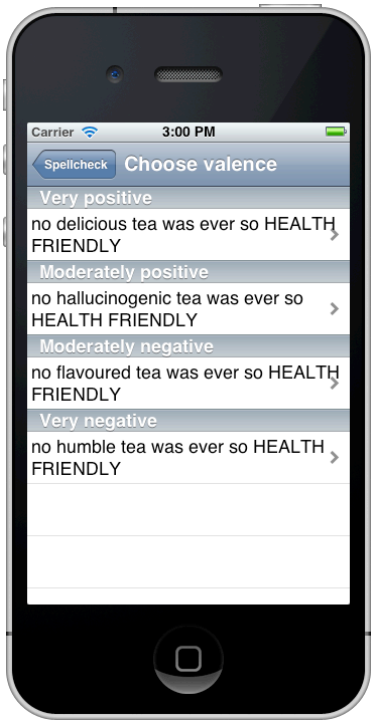


Figure 3

Figure 4

since we often may not have access to the original font due to limitations of the smartphone, and the image containing the new text is downloaded to the phone.

We next identify the original text color by looking at the first line of text in the original image, and treat that as the color for the whole text (even if the ad is written in multiple colors), to save processing time. This identification is performed by clustering the colors of the pixels in two groups with k-means, then considering the two means as colors of text and background.

Finally the new message is copied inside the bounding boxes of the original text. The image is shown onscreen (Figure 4) and the user can save it to the image library or share it by mail or MMS.

SUBVERTISER is currently implemented on the iPhone. An Android version is under development. The only external resources needed are the font recognition service and VALENTINO's server.

Further work

From the technical, NLP processing point of view, the basic capabilities of VALENTINO are currently being expanded. First, we are now starting to take into account the sentence structure, so that it can be better focused. An analysis of rhetorical aspects of a given sentence will also benefit the quality of the intervention. As for metrics that help decide the best quality for lexical choice, we have converged so far on the best mutual information measure (see above), although we think in different situations different measures should be applied.

More generally we can note that an extended VALENTINO could be parameterized to achieve different goals, including: a) generic valence shifting; b) focused biased language to influence the audience's view on one element (e.g., a human or a thing) in the sentence; c) "cleansing" of biases present in the original expression; or d) special effects, such as ironic or hyperbolic reconstructions.

As for point b, specifically for evaluative expressions, subjective evaluations can be along different dimensions: the ethical aspect (related to moral values), the epistemological aspect (related to truth), aesthetics (related to beauty or pleasure), and the utilitarian (related to utility, resources, results), and can be especially reflected in the lexical choice.

The aim is also to link a set of preferences and information about the context to the system. For instance audience preferences can shift its behavior in line with the user's attitude; also independent preferences (e.g. originating from a social institution) might produce expressions that could influence the audience towards a specific direction.

As for the overall mobile application we have described, some technical improvements, like automatic spell checking of the OCR output, would enhance the app's usability.

We would like to mention that additional uses can be envisaged, apart from defense against unwanted advertising expressions. In a sophisticated but not unusual twist of fate, the same technology can be used by the advertisers themselves. A new form of promotion could be based on an active role on the part of their target, which, by adding a

creative variation, contributes to the reinforcement of the basic advertising goal.

Another prospect is in an artistic direction. For instance the mobile application could be monitored on a large display by a crowd at an exhibition, where the audience could see posters in the city being continuously and automatically changed by different individuals walking around with their smartphones, so as to counter the messages on the walls and introducing a collective sense of liberation.

Another setting is with mobile games, where the user may interact with existing linguistic expressions to produce anagrams, wordplay and so on.

Conclusions

Creativity is widely used in advertising, which must appeal to people of all walks of life in every imaginable situation. But advertising also tries to change people's actions, beliefs and behavior, which they rightfully resist as an invasion of their time and attention. The resulting conflict leads to increasingly pervasive, aggressive and frequent advertisements on one hand, while on the other to a conscious refusal to pay attention to those ads or a profaning transformation of the message. Inspired by the latter, a system, even if just based on some degree of combinational creativity (Boden 2009) can aid people in defending themselves against elements in their environment.

We thus built SUBVERTISER, a creative subvertising system, which assists consumers in proactively "taking back" their daily outdoor routine. SUBVERTISER allows consumers to use the power of satire and virtual profaning to push back at advertisers. By combining the utility and pervasiveness of smartphones with the capability of the VALENTINO affective valencing system, consumers can take a picture of an advertisement in public that offends them, select the wording they want to change, use VALENTINO to supply them with language variations that subverts the intended message, and then modify the advertisement with their chosen variations to look just like the original. By sending the new version to their friends, they can join in a collective release of tension from the perpetual barrage of advertisements.

References

Boden, M. 2009. Computer Models of Creativity. *AI Magazine*, 30(3):23-34.

Brants, T., and Franz, A. 2006. Web 1T 5-Gram Version 1. *Linguistic Data Consortium*, Philadelphia.

Burke, M., Gorman, N., Nilsen, E., and Hornof, A. 2004. Banner Ads Hinder Visual Search and Are Forgotten. In *Proceedings of CHI 2004*, 1139-1142.

Esuli, A., and Sebastiani, F. 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of LREC 2006*, 417-422.

Freud S. 1905. *Der Witz und Seine Beziehung zum Unbewussten*. Deutike, Vienna.

Giora, R. 2003. *On Our Mind: Salience, Context and Figurative Language.* Oxford University Press, New York.

Guerini, M., Strapparava, C., and Stock, O. 2012. Ecological Evaluation of Persuasive Messages Using Google AdWords. In *Proceedings of ACL 2012*.

Guerini, M., Strapparava, C., and Stock, O. 2011. Slanting Existing Text with Valentino. In *Proceedings of IUI 2011*, 439-440.

Guerini, M., Strapparava, C., and Stock, O. 2008a. CORPS: A Corpus of Tagged Political Speeches for Persuasive Communication Processing. *Journal of Information Technology & Politics*, 5(1):19-32.

Guerini M., Strapparava C., and Stock O. 2008b. Valentino: A tool for Valence Shifting of Natural Language Texts, In *Proceedings of LREC 2008*, 243-246.

Mateas, M., Vanouse, P., and Domike, S. 2000. Generation of Ideologically-Biased Historical Documentaries. In *Proceedings of AAAI 2000*, 36-42.

Müller, J., Alt, F., and Michelis, D. (eds.). 2011. *Pervasive Advertising*. HCI Series - Springer.

Pagendarm, M., and Schaumburg, H. 2001. Why are users banner-blind? The impact of navigation style on the perception of web banners. *Journal of Digital Information*, 2(1).

Pianta, E., Girardi, C., and Zanoli, R. 2008. The TextPro tool suite, In *Proceedings of LREC 2008*, 2603-2607.

Stock, O., and Strapparava, C. 2003. Getting Serious about the Development of Computational Humour. In *Proceedings of IJCAI 2003*, 59–64.

Veale T. 2011. Creative Language Retrieval: A Robust Hybrid of Information Retrieval and Linguistic Creativity. In *Proceedings of ACL 2011*, 278-287.

Van der Sluis, I., and Mellish, C. 2009. Towards empirical evaluation of affective tactical NLG, in *Proceedings of ENLG 2009*, 146-153.

Whitehead, S., and Cavedon, L. 2010. Generating Shifting Sentiment for a Conversational Agent, in *Proceedings of NAACL HLT 2010*, 89-97.

Weaving creativity into the Semantic Web: a language-processing approach

Anna Jordanous
Centre for e-Research
Department of Digital Humanities
King's College London, UK
anna . jordanous at kcl . ac . uk

Bill Keller
Department of Informatics
University of Sussex
Brighton, UK
billk at sussex . ac . uk

Abstract

This paper describes a novel language processing approach to the analysis of creativity and the development of a machine-readable ontology of creativity. The ontology provides a conceptualisation of creativity in terms of a set of fourteen key components or *building blocks* and has application to research into the nature of creativity in general and to the evaluation of creative practice, in particular. We further argue that the provision of a machine readable conceptualisation of creativity provides a small, but important step towards addressing the problem of automated evaluation, 'the Achilles' heel of AI research on creativity' (Boden 1999).

Introduction

Creativity is a complex, multi-faceted concept encompassing many related aspects, abilities, properties and behaviours. This complexity makes the production of a comprehensive and generally applicable account of creativity problematic. Existing definitions of creativity are often too superficial for use by the research community and may be subject to discipline or domain bias, limiting their application. The need for a comprehensive, multi-dimensional account has been widely recognised (Rhodes 1961; Torrance 1967; Plucker, Beghetto, and Dow 2004; Kaufman 2009). Such an account would assist our understanding of creativity, highlighting areas of common ground and avoiding the pitfalls of disciplinary bias (Hennessey and Amabile 2010; Plucker and Beghetto 2004).

Words associated with academic debate about the nature of creativity are strongly linked to our understanding of its meaning and attributes. Analysis of this language provides a sound basis for constructing a sufficiently detailed and comprehensive account of the concept. In the present work, statistical language processing techniques are used to identify words significantly associated with creativity in a corpus of academic papers on the topic. A measure of lexical similarity provides a basis for clustering words and identifying key themes or components of creativity. The set of components yields information about the nature of creativity, based on what we emphasise when we discuss the concept.

Within the field of computational creativity, the problem of automatic evaluation remains a significant issue: 'the Achilles' heel of AI research on creativity' (Boden 1999). Recently, the Semantic Web has emerged as a way to address the troublesome but important issue (Boden 1999) of articulating values, concepts and information in an open and *machine-readable* format. Linked Data is the term used in the Semantic Web community to describe published data that is machine-readable and connected together using semantically typed links. We take the step of encoding our components in RDF, the current W3C standard for implementing Linked Data.[1] The resulting ontology is available to the wider research community as a resource in the Semantic Web, under the permanent URI `http://purl.org/creativity/ontology`, a form familiar to Semantic Web researchers and also accessible through browsers such as Marbles.[2]

Currently, most content on the Semantic Web is in the form of ontologies of 'things': semantically structured collections of factual or objective data on topics as diverse as people, places, narratives, or music.[3] To date, little work has been done on specifically defining subjective concepts in an ontology. However, current work on lexical resources such as WordNet has laid foundations for more definitionally troublesome concepts to be considered in detail; the time is ripe for development of ontologies of subjective concepts such as creativity.

Components of creativity

We identify a core lexicon consisting of just those words that appear to be highly associated with discussions of creativity in a corpus of academic papers on the topic. Our approach substantially develops and refines work described in Jordanous (2010). A key innovation is the use of a measure of lexical similarity, which allows the words to be clustered automatically to reveal a number of common themes or factors of creativity. Further analysis results in a set of fourteen

[1] `http://www.w3.org/TR/rdf-syntax-grammar`, last accessed 27th January 2012.

[2] `http://www.w3.org/2001/sw/wiki/Marbles`, last accessed 27th January 2012.

[3] Example ontologies are available at `http://www.foaf-project.org`, `http://www.geonames.org/ontology`, `http://www.contextus.net/ontomedia` and `http://musicontology.com` respectively, all last accessed 27th January 2012.

key components.

Corpus data

A 'creativity corpus' was assembled from a sample of 30 academic papers examining creativity from a variety of stand-points (Jordanous 2010). The selected papers cover a wide range of years (1950-2009) and academic disciplines, from psychological studies to computational models. Academic papers were used due to ease of location (e.g. through targeted literature search), accessibility (electronic publication for download), format (ease of conversion to text allows for computational analysis) and availability of citation data (used as a criterion for inclusion of a paper).[4]

In Jordanous (2010), language use in the creativity corpus was compared to general language use as represented by the British National Corpus (BNC) (Leech 1992). This had the undesired effect of highlighting words that were predominant in academic papers but not necessarily specific to creativity literature, e.g. *et', al'*. In the present study, a further corpus of 60 academic papers on topics unrelated to creativity was assembled (a 'non-creativity corpus'). For each paper in the creativity corpus, we retrieved the two most-cited papers in the same academic discipline[5] and with the same year of publication, that did not contain any words with the prefix *creat* (i.e. *creativity*, *creative*, *creation*, etc.).

Each corpus was processed using the RASP natural language processing toolkit (Briscoe, Carroll, and Watson 2006) to perform lemmatisation and part-of-speech (POS) tagging. Lemmatisation allows us to ignore morphological variation so that, e.g., *processed* and *processing* are both recognised as forms of *process*. POS tagging allows us to distinguish between different grammatical usages of the same orthographical form: e.g. *process* as a noun or as a verb. Two lists of frequency counts were produced: one for all words occurring in the creativity corpus and one for all words in the non-creativity corpus. Only 'content-bearing' words (i.e. nouns, verbs, adjectives and adverbs) were considered to be of interest. Any 'function words' or other minor categories (pronouns, articles, prepositions etc.), were ignored as they have little or no independent semantic content and are therefore of limited interest for the present study.

Finding words associated with creativity

A standard, statistical measure of association was used to identify words salient to discussions of creativity. The log-likelihood ratio (or G-squared statistic) is a measure of how well observed frequency data fit a model or expected frequency distribution. The statistic is an alternative to Pearson's chi-squared (χ^2) test that has been advocated as a more appropriate measure for corpus analysis as it does not rely on the (unjustifiable) assumption of normality in word distribution (Dunning 1993). This is a particular issue when analysing relatively small corpora as in the present case.[6] The log likelihood ratio is more accurate than χ^2 in its treatment of infrequent words in the data, which often hold useful information.

Our use of the log-likelihood ratio follows that of Rayson and Garside (2000). Given two corpora (in our case, 'creativity corpus' and 'non-creativity corpus') the log-likelihood score for a given word is calculated as:

$$LL = 2 \sum_{i \in \{1,2\}} O_i \ln(\frac{O_i}{E_i}) \quad (1)$$

where O_i is the observed frequency of the given word in corpus i and E_i is its expected frequency in corpus i. The expected frequency E_i is given by:

$$E_i = \frac{N_i \times (O_1 + O_2)}{N_1 + N_2} \quad (2)$$

where N_i denotes the total number of words in corpus i.

Following standard statistical practice, any word occurring fewer than five times was excluded. This ensures that the statistics are robust. To identify significant results, we also removed words with a log-likelihood score less than 10.83, representing a chi-squared significance value for p=0.001 (one degree of freedom). To identify words strongly associated with discussion of creativity it was necessary to select just those words with observed counts higher than than expected in the creativity corpus. This resulted in a total of 694 distinctive *creativity words*: a collection of 389 nouns, 205 adjectives, 72 verbs and 28 adverbs that occurred significantly more often than expected in the creativity corpus. The 20 such words with the highest log-likelihood ratio scores are listed in Table 1.

It is important to note that our objective is to identify key themes in the lexical data, not to induce a comprehensive terminology of creativity. Despite the relatively small size of the available corpora, the resulting set of 694 creativity words is sufficiently rich for this purpose.

Identifying components of creativity

In Jordanous (2010) an attempt was made to identify key components by clustering creativity words by inspection of the raw data. In practice, this proved laborious and made it impossible systematically to consider all of the identified words. It also raised issues of subjectivity and experimenter bias. Here we address these problems, at least in part, by first clustering all the words automatically according to a statistical measure of *distributional similarity* (Lin 1998). The more manageable collection of clusters are then inspected manually to identify key components.

Intuitively, words that tend to occur in similar linguistic contexts will tend to be similar in meaning (Harris 1968). For example, evidence that the words *concept* (LLR=189.90) and *idea* (LLR=475.74) are similar in meaning might be provided by occurrences such as the following:

[4] Note that some papers have been published in very recent years and therefore have few citations. In this case selection was based on subjective judgement of influence.

[5] As categorised by the literature database *Scopus* `http://www.scopus.com/`), last accessed 27th January 2012.

[6] At around 300K and 700K words respectively, the creativity and non-creativity corpora are very small compared to the British National Corpus ($\approx$ 100M words) and tiny in comparison to recent, web-derived text collections of billions of words.

Word (and part of speech tag)	LLR
thinking (N)	834.55
process (N)	612.05
innovation (N)	546.20
idea (N)	475.74
program (N)	474.41
domain (N)	436.58
cognitive (J)	393.79
divergent (J)	355.11
openness (N)	328.57
discovery (N)	327.38
primary (J)	326.65
originality (N)	315.60
criterion (N)	312.61
intelligence (N)	309.31
ability (N)	299.27
knowledge (N)	290.48
create (V)	280.06
experiment (N)	253.32
plan (N)	246.29
agent (N)	246.24

Table 1: The top 20 results of the log-likelihood ratio (LLR) calculations. A significant LLR score at p=0.001 is 10.83.

1. the *concept/idea* involves (subject of verb 'involve')
2. applied the *concept/idea* (object of verb 'apply')
3. the basic *concept/idea* (modified by adjective 'basic')

Word occurrence data of this kind was obtained from an analysis of the written portion of the BNC, which had previously been processed using the RASP toolkit to extract grammatical dependency relations (*subj-of*, *obj-of*, *modified-by*). Each word in the creativity corpus was then associated with a list of all of the grammatical relations in which it participated, together with corresponding counts of occurrence.

Distributional similarity of two words is measured in terms of the similarity of their associated lists of grammatical relations. The present work adopts an information-theoretic measure devised by Lin (1998), which has been widely used in language processing applications and shown to perform well against other similarity measures as a means of identifying near-synonyms (Weeds and Weir 2003) . Similarity scores were obtained separately for pairs of nouns, pairs of verbs and so on. For a given set of words, the similarity data is conveniently visualised as a graph or network, where nodes correspond to words and edges are weighted by similarity scores, as in Figure 1.

A possible problem with obtaining word similarity data this way would arise if the majority of the creativity words were used with distinctive or technical senses within the creativity corpus. This is unlikely, however: whilst some narrowly specialised usage may be present in our creativity lexicon, most words retain general senses reflected in the wider BNC data set.

The graph clustering software *Chinese Whispers* (Biemann 2006) was used to automatically identify word clusters in the dataset. This algorithm uses an iterative process to group together graph nodes that are located 'close' to each other. By grouping words with similar meanings, the number of data items was effectively reduced and themes in the data could be identified more readily by inspection. Themes discovered through clustering were further analysed in terms of the *Four Ps* of creativity (Rhodes 1961; Mooney 1963; MacKinnon 1970) to identify alternative perspectives and reveal subtler (but still important) aspects of creativity. From the analysis it was possible to extract a set of fourteen key components of creativity.

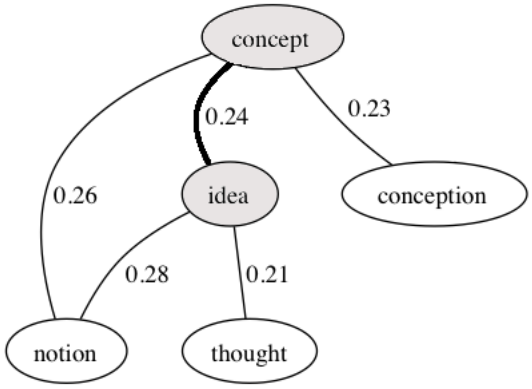


Figure 1: Graph representation of the similarity of the nouns *concept* and *idea* and related words. Words are drawn as nodes linked by weighted edges representing word similarity (maximum similarity is 1.0).

Implementing an ontology of creativity

The fourteen components provide a clear account of the constituent parts of the concept of creativity. Our remaining contribution is to express these components in a machine-readable form. We also want to use Linked Data principles (Heath and Bizer 2011) to connect the individual components to other data sources within the Semantic Web, so that creativity is defined in terms of concepts that have already been defined. To achieve this, we used SKOS (Simple Knowledge Organisation System),[7] a W3C standard which provides a model for representing ontological data within the Semantic Web. We also made use of WordNet (Reed and Lenat 2002), a large lexical database of English in which words are grouped by sense and interlinked by lexical and conceptual relations. WordNet has recently been made available as a Semantic Web ontology.[8]

The SKOS ontology incorporates three main classes: *skos:Concept* (anything we may want to record information about), *skos:ConceptScheme* (a set that collectively defines a skos:Concept) and *skos:Collection* (a collection of semantically-related information).

We created an instance of *skos:ConceptScheme* called

[7] `http://www.w3.org/TR/skos-reference`, last accessed 27th January 2012.

[8] `http://wordnet.rkbexplorer.com/`

CreativityComponents to represent the set of components that defines the *skos:Concept* of *Creativity*. Each component is represented as an individual *skos:Concept*. As RDF is a graph-based model, the resulting encoding can be visualised as in Figure 2. The graph has also been published in serialised format as an RDF/XML text file and made available as `http://purl.org/creativity/ontology`. The *skos:Concept* labelled *Creativity* has the unique URI `purl.org/creativity/ontology#Creativity` and any Linked Data that needs to refer to the concept can use this identifier.

The distributed nature of Semantic Web research means that the enormous task of defining concepts in a machine-readable form is divided across the research field, rather than being the sole responsibility of one particular research group. This work practice acts as a form of peer review, as ontologies are developed, critiqued, and ultimately judged by the extent to which they are adopted and re-used as points of reference by other researchers.

Upper ontologies allow us to link the concepts in our ontology to related ontological work on creativity in the future (even if these future researchers are not aware of our ontological contribution). An upper ontology defines higher-level vocabularies and concepts necessary to implement ontologies themselves, providing the meta-vocabulary to link specific ontologies to more general concepts. The implementation of the Wordnet dataset and structure as an ontology provides WordNet as an upper ontology for us to use, linking a lexical string (e.g. "creativity") to various concepts associated with that string, such as its sense, hyponyms, type, 'gloss' (brief definition) and other related lexical information.

Each component in our ontology is comprised of a cluster of keywords. It makes sense, therefore, to link each component back to the appropriate keywords, using the WordNet ontology at `http://wordnet.rkbexplorer.com/`. In this way, our components are linked into the Semantic Web through the WordNet ontology. This linkage also provides further semantic information on each component via the lexical relations and other information represented in the WordNet hierarchy. Finally, following Linked Data principles, we also link our interpretation of creativity as an extension of the representation of the concept in WordNet. In this way, machines (and people) can see the relationship between this general concept of creativity and our more detailed ontological analysis.

Discussion and Implications

The current work is part of a wider project engaged with the question of the evaluation of computational creativity (Jordanous 2011). The components of creativity have already been applied, both for in-depth expert evaluation and in forming snapshot judgements of the creativeness of a given system. The resulting component-based evaluation yields detailed information about creative strengths and weaknesses. Crucially, the evaluation highlights those components where a system performs poorly, providing insight into areas where improvement in performance is needed.

By publishing the ontology in the Semantic Web we ensure that it is freely available to the research community. This has a number of implications. First, it may be freely referred to, extended or amended. Refinement is clearly possible, for example in providing more fine-grained analysis of the components or in articulating the relationships between them. Second, it facilitates the development of creativity-aware applications to support manual evaluation of creativity based on the components. It also represents a step towards the development of methods of automated evaluation. One intriguing possibility is to further exploit language processing techniques to provide automated evaluation by proxy based on textual reviews or descriptions of system performance. This is analogous to the way that sentiment analysis techniques are now used to automatically evaluate attitude and opinion based on reviews of products or services (Pang and Lee 2008).

The current work illuminates the sorts of issues that arise in formal modelling of subjective or 'soft' concepts such as creativity. For example, some of our components appear logically inconsistent with others in the set: e.g. the need for autonomous, independent behaviour (*Independence and Freedom*) versus the requirement for social interaction (*Social Interaction and Communication*). Also, creativity clearly manifests itself in different ways across different domains (Plucker and Beghetto 2004) and components will vary in importance, according to the requirements of a particular domain. For example, creative behaviour in mathematical reasoning has more focus on finding a correct solution to a problem than is the case for creative behaviour in, say, musical improvisation (Colton 2008). Questions remain about how such dialectical and fluid aspects might be modelled. We present the set of components as a rather loose collection of dimensions – attributes, abilities and behaviours, etc. – which contribute to our overall understanding of creativity, rather than a unified definition.

Concluding remarks

This paper has described the development of an ontology of creativity using corpus-based, language processing techniques and its publication as machine-readable, Linked Data in the Semantic Web. The resulting ontology provides a multi-perspective analysis of creativity in terms of a set of fourteen key components and has application to the study and evaluation of computational creativity. Weaving the ontology into the Semantic Web has implications for future work on modelling subjective concepts and suggests some interesting directions for future research into the problem of automated evaluation of creativity.

References

Biemann, C. 2006. Chinese Whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing*, 73–80. Morristown, NJ: Association for Computational Linguistics.

Boden, M. A. 1999. Introduction [summary of Boden's keynote address to AISB'99]. In *AISB Quarterly - Special issue on AISB99: Creativity in the arts and sciences*, volume 102, 11.

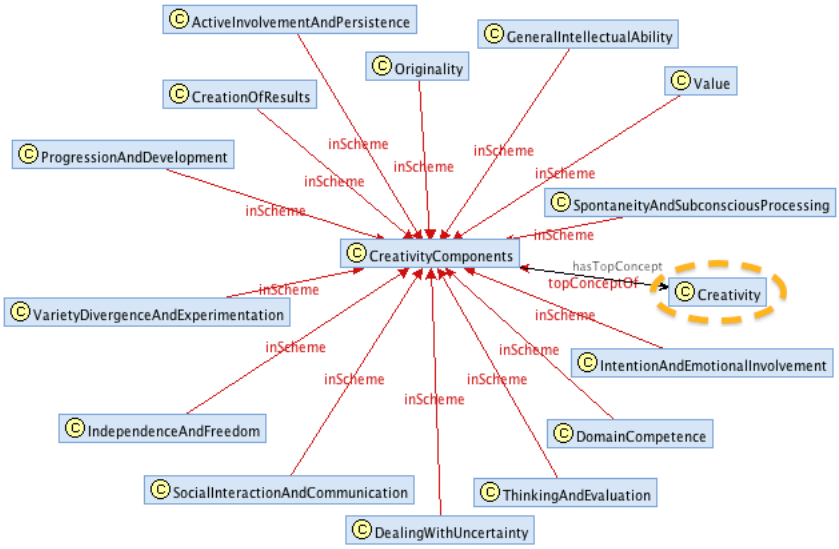


Figure 2: The RDF ontology of Creativity, in graph form.

Briscoe, E.; Carroll, J.; and Watson, R. 2006. The second release of the RASP system. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *Proceedings of AAAI Symposium on Creative Systems*, 14–20.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1):61–74.

Harris, Z. 1968. *Mathematical Structures of Language*. New York: Wiley.

Heath, T., and Bizer, C. 2011. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool.

Hennessey, B. A., and Amabile, T. M. 2010. Creativity. *Annual Review of Psychology* 61:569–598.

Jordanous, A. 2010. Defining creativity: Finding keywords for creativity using corpus linguistics techniques. In *Proceedings of the International Conference on Computational Creativity*, 278–287.

Jordanous, A. 2011. Evaluating evaluation: Assessing progress in computational creativity research. In *Proceedings of the Second International Conference on Computational Creativity (ICCC-11)*.

Kaufman, J. C. 2009. *Creativity 101*. The Psych 101 series. New York: Springer.

Leech, G. 1992. 100 million words of English: the British National Corpus (BNC). *Language Research* 28(1):1–13.

Lin, D. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, 296–304.

MacKinnon, D. W. 1970. Creativity: a multi-faceted phenomenon. In Roslansky, J. D., ed., *Creativity: A Discussion at the Nobel Conference*. Amsterdam, The Netherlands: North-Holland Publishing Company. 17–32.

Mooney, R. L. 1963. A conceptual model for integrating four approaches to the identification of creative talent. In Taylor, C. W., and Barron, F., eds., *Scientific Creativity: Its Recognition and Development*. New York: John Wiley & Sons. chapter 27, 331–340.

Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. *Foundation and Trends in Information Retrieval* 2(1-2):1–135.

Plucker, J. A., and Beghetto, R. A. 2004. Why creativity is domain general, why it looks domain specific, and why the distinction doesn't matter. In Sternberg, R. J.; Grigorenko, E. L.; and Singer, J. L., eds., *Creativity: From Potential to Realization*. Washington, DC: American Psychological Association. chapter 9, 153–167.

Plucker, J. A.; Beghetto, R. A.; and Dow, G. T. 2004. Why isn't creativity more important to educational psychologists? Potentials, pitfalls, and future directions in creativity research. *Educational Psychologist* 39(2):83–96.

Rayson, P., and Garside, R. 2000. Comparing corpora using frequency profiling. In *Proceedings of ACL Workshop on Comparing Corpora*.

Reed, S. L., and Lenat, D. B. 2002. Mapping ontologies into Cyc. In *Proceedings of AAAI'02 workshop on Ontologies and the Semantic Web*.

Rhodes, M. 1961. An analysis of creativity. *Phi Delta Kappan* 42(7):305–310.

Torrance, E. P. 1967. Scientific views of creativity and factors affecting its growth. In Kagan, J., ed., *Creativity and Learning*. Boston: Beacon Press. 73–91.

Weeds, J., and Weir, D. 2003. Finding and evaluating nearest neighbours. In *Proceedings of the 2nd Conference of Corpus Linguistics*.

Coming Together: Composition by Negotiation by Autonomous Multi-Agents

Arne Eigenfeldt
School for the Contemporary Arts
Simon Fraser University
Vancouver, Canada
arne_e@sfu.ca

Philippe Pasquier
School of Interactive Arts and Technology
Simon Fraser University
Surrey, Canada
pasquier@sfu.ca

ABSTRACT

Coming Together is a series of computational creative systems based upon the premise of composition by negotiation – within a controlled musical environment, autonomous multi-agents attempt to converge their data, resulting in a self-organised, dynamic, and musically meaningful performance.

All the *Coming Together* systems involve some aspect of *a priori* structure around which the negotiation by the agents is centered. In the versions demonstrated, the structure presupposes several discrete movements that together form a complete composition of a predetermined length. Characteristics of each movement – density, time signature, tempo – are generated using a fuzzy-logic method of avoiding similarity between succeeding movements.

Two versions of *Coming Together* are described, used in two different musical compositions. The first, for the composition *And One More*, involves agents interacting in real-time, their output being sent via MIDI to a mechanical percussion instrument. This version has nine different agents performing on eighteen different percussion instruments, and includes a live percussionist whose performance is encoded and considered an additional agent.

The second version, for the composition *More Than Four*, involves four agents, whose output is eventually translated into musical notation using *MaxScore*[1], for performance by four instrumentalists. Agent interaction is transcribed to disk prior to performance; at the onset of the performance, a curatorial agent selects previous movements from the database, and chooses from those to create a musically unified composition.

[1] www.computermusicnotation.com

Continuous Improvisation and Trading with Impro-Visor

Robert M. Keller
Computer Science Department
Harvey Mudd College
Claremont, CA 91711 USA
keller@cs.hmc.edu

Demonstration

Impro-Visor is a free open-source program designed to help musicians learn to improvise. Its main purpose is to help its user become a better improviser. It can exhibit creativity by improvising continuously on its own in a variety of soloist styles. We demonstrate that, in principle, Impro-Visor can continue creating indefinitely, without repeating the same sequence of musical ideas. We also demonstrate how Impro-Visor can alternate ("trade") phrases with the soloist, again continuously, as well as recording what the soloist plays on a MIDI device. Related aspects that can be shown are learning an improvisational style through grammar acquisition and using "roadmaps" as a basis for trading. The figure shows a screen shot of Impro-Visor creating phrases in real-time and capturing the soloist's input in real-time from a MIDI device.

Acknowledgements

The author thanks the NSF (CNS REU #0753306), Impro-Visor co-developers, and Harvey Mudd College for their generous support.

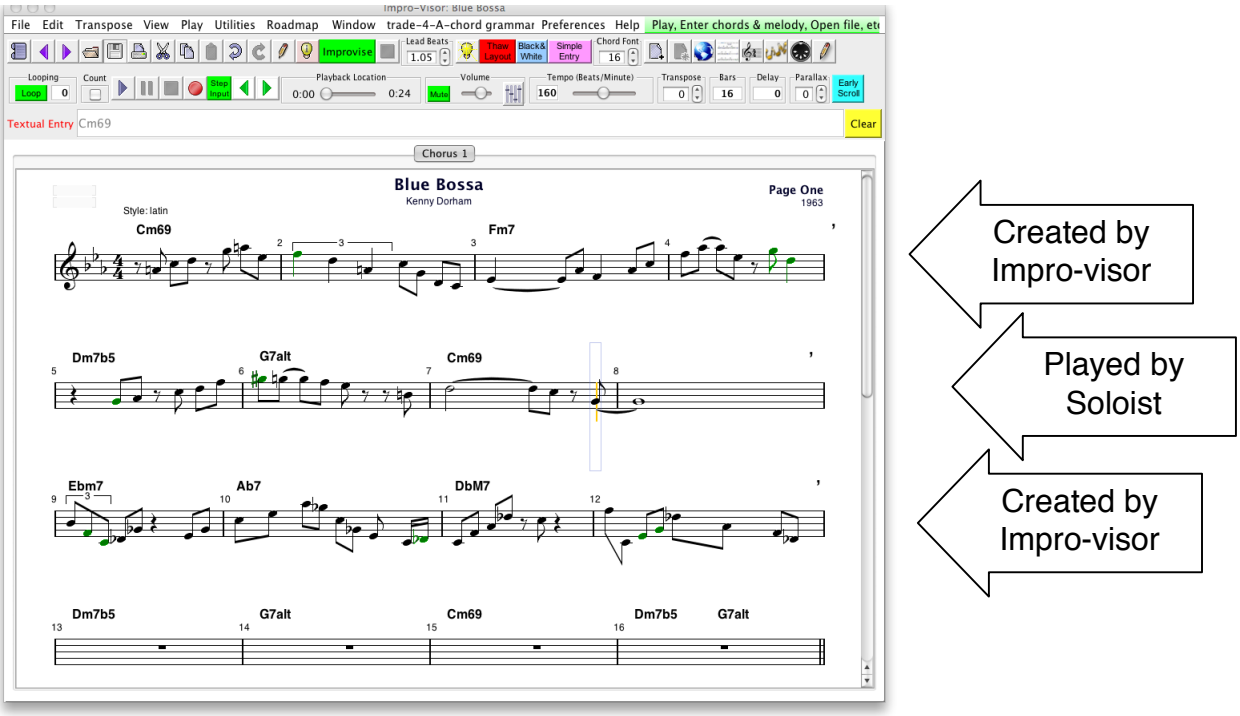
References

Elliott, J. 2009. *Insights in Jazz: An Inside View of Jazz Standard Chord Progressions*. http://www.dropback.co.uk/

Gillick, J.; Tang, K.; Keller, R. 2010. Machine learning of jazz grammars. *Computer Music Journal*, September 2010.

Impro-Visor. 2012. http://www.impro-visor.com/

Keller, R., Toman-Yih, A., Schofield, A., and Merritt, Z., A creative improvisational companion based on idiomatic harmonic bricks. *Proc. 3rd ICCC 2012, Dublin*.



Exploring Everyday Creative Responses to Social Discrimination with the Mimesis System

D. Fox Harrell[†+*], Chong-U Lim[*], Sonny Sidhu[†], Jia Zhang[†], Ayse Gursoy[†], Christine Yu[+]

Comparative Media Studies Program[†] | Program in Writing and Humanistic Studies[+]
Computer Science and Artificial Intelligence Laboratory[*]
Massachusetts Institute of Technology
{fox.harrell, culim, sidhu, zhangjia, agursoy, czyu}@mit.edu

Introduction

We have created an interactive narrative system called *Mimesis*, which explores the social discrimination phenomena through gaming and social networking. *Mimesis* places players in control of a mimic octopus in its marine habitat that encounters subtle discrimination from other sea creatures. Relevant to computational creativity, Mimesis explores:

1) Collective creativity by constructing game characters algorithmically from collective musical preferences on a social networking site.
2) Everyday creativity by modeling the diverse creative ways people respond to covert acts of discrimination.

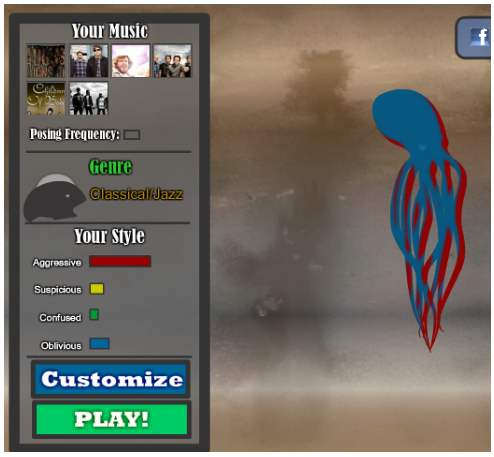


Figure 1: The player character is customized based on the player's musical preferences on Facebook.

Collective Creativity

Building on previous work [2], *Mimesis* requests access to information from the player's Facebook profile, using music preferences in the player's social network as a stand-in for qualities of individual and social identity. *Mimesis* generates corresponding moods for each musical artist. By associating the player character with artists' moods such *oblivious, confused, suspicious,* or *aggressive*, players can impart these qualities onto the player character (see **Figure 1**). Within gameplay, moods are mapped to strategies of conversationally responding to microaggressions.

Everyday Creativity

The player character encounters other sea creatures who utter sentences like: "Where are you from?" and "You don't seem like the typical creature around here." This is shown in **Figure 2**. While such questions may seem benign, they can also covertly imply the theme: "You are an alien in your own land" (such might be encountered by an Asian American in the United States). The player responds by using gestural input such as pinching out for an open/oblivious attitude or pinching in for a closed/aggressive attitude. Each encounter plays out according to a conversational narrative schema based on sociolinguistic studies of narratives of personal experience.

Figure 2: The screen shows the player's character (left) in a microinvalidation encounter with an NPC (right).

These encounters convey aspects of the experience of microaggressions, which are covert acts of discrimination. Researchers Sue et al. identify "microinvalidations" as communications that exclude, negate, or nullify the experiential reality of others). The "alien in your own land" theme is an example of microinvalidation. Microaggressions have been clinically found to have strong cumulative effects on health and happiness, restrict understandings between groups. [1]

We hope the system is an effective tool for increasing awareness of this subtle form of social discrimination.

References

[1] Sue, D., Capodilup, C., Torina, G., Bucceri, J., Holder B., Nadal, K., Esquilin, M. Racial Microagressions in Everyday Life, *American Psychologist* 62 (2007).

[2] Harrell, D.F., Vargas, G., Perry, R. "Steps Toward the AIR Toolkit: An Approach to Modeling Social Identity Phenomena in Computational Media," *Proceedings of the 2nd International Conference on Computational Creativity*, 2011.

Functional Representations of Music

James McDermott*, University College Dublin.

April 30, 2012

Music is an interesting domain for the study of computational creativity. Some generative formalisms for musical composition (e.g. Markov chains) achieve plausible music over short time-scales (a few notes) but appear to be "meandering" over longer time-scales. Imposing a sense of teleology or purpose is an important goal in creating valuable music.

In the field of evolutionary computation (EC), researchers draw inspiration from Darwinian evolution to address computational problems. EC can be applied to aesthetic and creative domains. Although EC is commonly used to generate music, key open issues remain. Formal measurement of the quality of a piece of music in a computational *fitness function* is an obvious obstacle. A naive *representation* for music, such as a list of integer values each corresponding directly to a note, will tend to produce disorganised music.

In previous work, Hoover et al. [1, and later] showed that a *functional* representation could impose organisation and a sense of purpose. In this system, music is represented as a function of time. A fixed piece of pre-existing music is used as a "scaffold": the system then evolves functions, i.e. mappings from the scaffold to new accompanying material. Time-series of numerical "control" variables are also proposed as a means of imposing structure on the music. Fitness is judged interactively.

The XG project is partly inspired by this work. It discards the "scaffold", but relies on the time-series of control variables (see Figure 1).

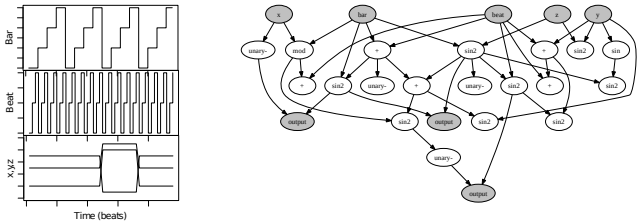


Figure 1: Time-series of control variables (left) impose a bar/beat structure and an overall AABA structure. The evolved function (right) maps these variables to numerical outputs, once per time-step. The outputs are interpreted as music.

It also differs in its internal representation for the mappings (a simple language of arithmetic functions, with special *accumulator* functions at the outputs to control volume), and its use of a computational (non-interactive) fitness function. Surprisingly good results arise using this representation in combination with a simple fitness function which rewards *variety* in the output music. Neither the functional representation nor the fitness function is alone capable of producting good results. More details are available in a full paper [2] and online[1].

A longer-term goal of the XG project is to create large-scale musical works as mappings from pre-existing time series arising in nature and human affairs, and from non-musical artforms such as film or still images with a sequential aspect.

[1] Amy K. Hoover, M. P. Rosario, and Kenneth O. Stanley. Scaffolding for interactively evolving novel drum tracks for existing songs. In *Proceedings of EvoWorkshops*, volume 4974 of *LNCS*, page 412. Springer, 2008.

[2] James McDermott and Una-May O'Reilly. An executable graph representation for evolutionary generative music. In *GECCO '11*, Dublin, 2011.

*Funded by IRCSET/Marie Curie; jamesmichaelmcdermott@gmail.com

[1] http://www.skynet.ie/~jmmcd/xg

MaestroGenesis: Computer-Assisted Musical Accompaniment Generation

Paul A. Szerlip, Amy K. Hoover, and Kenneth O. Stanley
Department of Electrical Engineering and Computer Science
University of Central Florida
Orlando, FL 32816-2362 USA
{paul.szerlip@gmail.com, ahoover@eecs.ucf.edu, kstanley@eecs.ucf.edu}

Abstract

This demonstration presents an implementation of a computer-assisted approach to music generation called functional scaffolding for musical composition (FSMC) whose representation facilitates creative combination, exploration, and transformation of musical ideas and spaces. The approach is demonstrated through a program called MaestroGenesis with a convenient GUI that makes it accessible to even non- musicians. Music in FSMC is represented as a functional relationship between an existing human composition, or *scaffold*, and a generated accompaniment. This relationship is represented by a type of artificial neural network called a compositional pattern producing network (CPPN). A human user without any musical expertise can then explore how accompaniment can relate to the scaffold through an interactive evolutionary process akin to animal breeding.

Composing with MaestroGenesis

MaestroGenesis is a program that helps users create complete polyphonic pieces with only the musical expertise necessary to compose a simple, monophonic melody. Users begin creating accompaniments by establishing a *scaffold*, or melody that will provide the initial rhythmic and harmonic seed for the accompaniment. The accompaniment is then represented as a functional transformation of this original scaffold through a method called functional scaffolding for musical composition (FSMC) (Hoover et al. 2012). FSMC explots the structure already present in the human-composed scaffold by computing a *function* that transforms its structure into the accompaniment.

These FSMC-accompaniments are then bred like animals might be bred. Once the scaffold is chosen, a population of ten accompaniments is displayed. Each is rated as good or bad by pressing the "thumbs-up" button (figure 1). By ratings accompaniments with favorable qualities higher than those without, the next generation of accompaniments tends to possess similar qualities to the well-liked parents. Through interactively evolving these accompaniments, they grow to reflect the personal inclinations of the user.

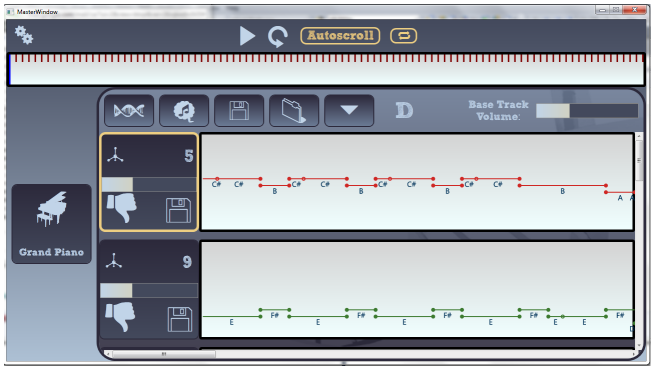


Figure 1: **MaestroGenesis Candidate Accompaniments.** Accompaniments in MaestroGenesis are evolved through a process similar to animal breeding. Candidate accompaniments are evolved ten at a time in an interative process in which each subsequent generation inherits traits from the previous population.

Conclusion

MaestroGenesis is a program that facilitates creativity in music composition through functional scaffolding for musical composition (FSMC) (Hoover et al. 2012). Accompaniments are evolved through a process similar to animal breeding. The program is availble for download at `http://maestrogenesis.org`.

Acknowledgements

This work was supported in part by the National Science Foundation under grant no. IIS-1002507 and also by a NSF Graduate Research Fellowship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

References

Hoover, A. K.; Szerlip, P. A.; Norton, M. E.; Brindle, T. A.; Merritt, Z.; and Stanley, K. O. 2012. Generating a complete multipart musical composition from a single monophonic melody with functional scaffolding. In *Proc. of the 3rd Intl. Conf. on Computational Creativity (ICCC-2012)*.

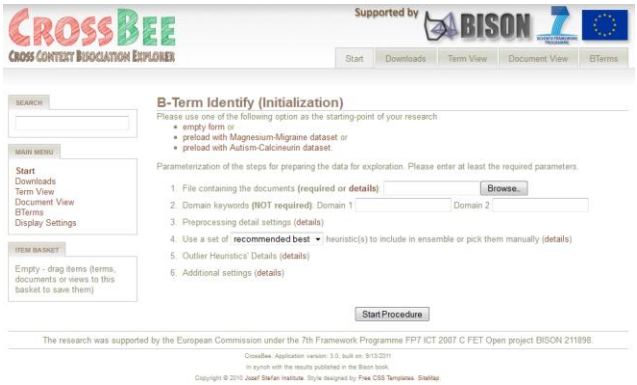
CrossBee: Cross-Context Bisociation Explorer

Matjaž Juršič[1,2], Bojan Cestnik[3,1], Tanja Urbančič[4,1], Nada Lavrač[1,4]
[1] Jožef Stefan Institute, Ljubljana, Slovenia
[2] International Postgraduate School Jožef Stefan, Ljubljana, Slovenia
[3] Temida d.o.o., Ljubljana, Slovenia
[4] University of Nova Gorica, Nova Gorica, Slovenia
{matjaz.jursic, bojan.cestnik, tanja.urbancic, nada.lavrac}@ijs.si

CrossBee is an exploration engine for text mining and cross-context link discovery, implemented as a Web application with a user-friendly interface. The system supports the expert in advanced document exploration supporting document retrieval, analysis and visualization. It enables document retrieval from public databases like PubMed, as well as by querying the Web, followed by document cleaning and filtering through several filtering criteria. Document analysis includes document presentation in terms of statistical and similarity-based properties, topic ontology construction through document clustering. A distinguishing feature of CrossBee is its powerful cross-context and cross-domain document exploration facility and bisociative (Koestler 1964) term discovery aimed at finding potential cross-domain linking terms/concepts. Term ranking based on an ensemble heuristic (Juršič et al. 2012) enables the expert to focus on cross-context links with high potential for cross-context link discovery. CrossBee's document visualization and user interface customization additionally support the expert in finding relevant documents and terms through similarity graph visualization, a color-based domain separation scheme and highlighted top-ranked bisociative terms.

A typical user scenario starts by inputting two sets of documents of interest and by regulating the parameters of the system. The required input is a file with documents from two domains. Each line of the file contains exactly three tab-separated entries: (a) document identification number, (b) domain acronym, and (c) the document text. The other options available to the user include specifying the exact preprocessing options, specifying the base heuristics to be used in the ensemble, specifying outlier documents identified by external outlier detection software, defining the already known bisociative terms (*b* terms), and others. Next, CrossBee starts a computationally very intensive step in which it prepares all the data needed for the fast subsequent exploration phase. During this step the actual text preprocessing, base heuristics, ensemble, bisociation scores and rankings are computed in the way presented in the previous section. This step does not require any user intervention. After this computation, the user is presented with a ranked list of *b* term candidates. The list provides the user with some additional information including the ensemble's individual base heuristics votes and term's domain occurrence statistics in both domains. The user then browses through the list and chooses the term(s) he believes to be promising *b* terms, i.e. terms for finding meaningful connections between the two domains. At this point, the user can start inspecting the actual appearances of the selected term in both domains, using the efficient side-by-side document inspection. In this way, he can verify whether his rationale behind selecting this term.

CrossBee is available at website: http://crossbee.ijs.si/. The system's home page is shown below.



References

Juršič, M.; Sluban, B.; Cestnik, B.; Grčar, M.; and Lavrač, N. 2012. Bridging concept identification for constructing information networks from text documents. In: Berthold, M.R. ed., *Bisociative Knowledge Discovery*. Springer LNAI 7250 (in press).

Koestler, A. 1964. *The Act of Creation*. New York: MacMillan.

Computer Software for Measuring Creative Search

Kyle E. Jennings
Department of Psychology
University of California, Davis
Davis, CA 95616 USA
kejennings@ucdavis.edu

The creative process can be thought of as the search through a space of possible solutions for one that best satisfies the problem criteria. To better understand this search process, two face valid creative tasks have been created, both of which track the intermediate configurations that creators explore. These data—called search trajectories—may yield valuable insights into the creative process. This demonstration allows visitors to try both tasks and to see the sorts of data that are produced.

The first task is a computerized version of Amabile's (1982) popular collage task, wherein participants make themed collages using colored shapes (see Figure 1). The software allows the shapes to be moved, rotated, flipped, and stacked using an intuitive mouse-based interface. The creator's moves can be characterized according to the extent that the set of shape movements actually performed exceeds the minimal set of movements needed to produce the final collage.

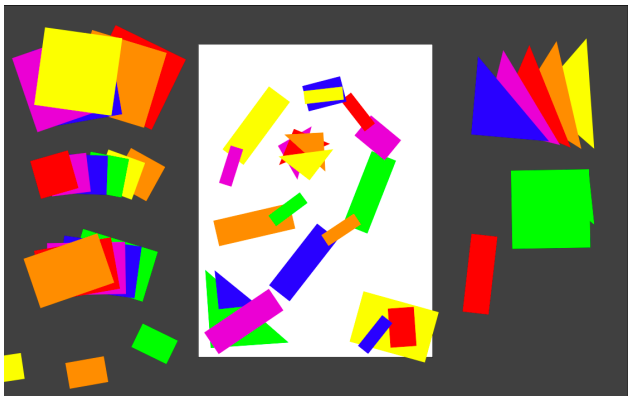


Figure 1: Intermediate screen of the collage task. The white area represents a piece of paper and the gray area is a work area. Initially all of the shapes are stacked in the work area, similar to the triangles in the upper-right corner.

The second task, called the orbital composition task (Jennings 2010; Jennings, Simonton, and Palmer 2011), involves arranging a camera and light that lie in fixed circular orbits around a set of objects. The configuration space has only three dimensions—camera angle, camera zoom, and light angle—but the scene is constructed in a way that permits many interesting and varied images (see Figure 2). While less face valid than the collage task, the orbital task's simplicity permits more consistent analyses and makes it possible to collect ratings from a standardized subset of the space, thereby providing a sense of the search landscape topology that can help overcome some of the ambiguities inherent in analyzing search trajectories alone (see Jennings 2012).

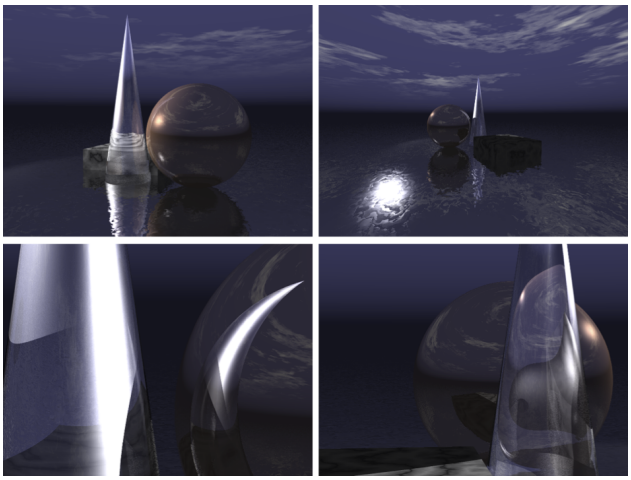


Figure 2: Final images from the orbital composition task as selected by four different research participants.

References

Amabile, T. 1982. Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology* 43(5):997–1013.

Jennings, K. E.; Simonton, D. K.; and Palmer, S. E. 2011. Understanding exploratory creativity in a visual domain. In *Proceedings of the Eighth ACM Conference on Creativity and Cognition*.

Jennings, K. E. 2010. Search strategies and the creative process. In *Proceedings of the First International Conference on Computational Creativity*, 130–139.

Jennings, K. E. 2012. Creative search trajectories and their implications. In *Proceedings of the Third International Conference on Computational Creativity*.

ANGELINA - Coevolution in Automated Game Design

Michael Cook and Simon Colton
Computational Creativity Group, Imperial College, London (ccg.doc.ic.ac.uk)

Figure 1: Screenshot from a game about a murdered aid worker from Scotland. The background image is of the Scottish landscape, and a red ribbon image has been selected to represent the aid charity.

ANGELINA

ANGELINA is a co-operative co-evolutionary system for automatically creating simple videogames. It has previously been used to design both simple arcade-style games and two-dimensional platformers. In the past, ANGELINA's efforts have been focused on mechanical aspects of design, such as level creation, rule selection and enemy design. We are now in the process of expanding ANGELINA's remit to cover other aspects of videogame design, including aesthetic problems such as art direction and the selection and use of external media to evoke emotion or communicate meaning.

ANGELINA has produced several new games for this demonstration, exemplifying the new abilities the system now has. Its co-operative co-evolutionary system for platform games is composed of four modules: (i) a level designer that places solid blocks and locked doors to shape the progress of the player (ii) a layout designer that places and designs the enemies the player faces, as well as the start and end of the level (iii) a powerup designer that defines what bonus items the player can acquire during gameplay and (iv) a creative direction module that arranges a set of media resources in the level for the player to discover during gameplay. This latter module is the newest addition to the system, and takes advantage of many new capabilities built into ANGELINA for retrieving content from the web dynamically for use in themed videogames.

Design Task

Inspired by the collage-creation problem described in (Cook and Colton 2011) ANGELINA obtains current affairs articles by accessing the website of the British newspaper The Guardian. It selects a news story, and attempts to design a short platform game whose theme is inspired by the news article selected. Currently, this allows ANGELINA to demonstrate simple abilities such as the appropriate selection of media from a wide variety of sources, and arrangement in a potentially nonlinear level space.

Figure 2: Media retrieved for a game inspired by an inquiry into a newspaper. Left: an image retrieved using the phrase 'newspapers and magazines'. On the right is Rebekah Brooks, one of the journalists in the investigation.

ANGELINA uses online knowledge sources such as Wikipedia to extract additional information about data retrieved from the news articles - it can, for instance, identify when a country is the subject of a news article, allowing the system to search photography websites such as Flickr for photographs of that country to use as a backdrop to the game. Keyword-based searches can also be augmented with emotional keywords to alter the results they return, based on techniques described in (Cook and Colton 2011). By reading live Twitter search results about a named person in the news article, ANGELINA can use search augmentation appropriate to the opinions it finds to retrieve media that reflect perceived public opinion of a particular topic. Although a simple technique, it is a first step towards the system dealing with opinion and bias through the work it produces.

Games

The games produced are simple platform games, loosely following the design tenets of the *Metroidvania* subgenre. The player must navigate the level space to reach the exit, but in order to gain access to later level sections, it is necessary to seek out and obtain items that add to the player's capabilities (for example: unlocking doors or changing the player's jumping abilities). As the player explores further they will encounter enemies, as well as images and sound content that is appropriate to the game's theme.

ANGELINA is implemented in Java, but the games the system produces are Flash-based. When ANGELINA has evolved a game design, it modifies an existing ActionScript game template to include the generated design content, as well as incorporating the media downloaded and selected from the internet. All of the games available in the demonstration, as well as others developed by the system, are available on the project website: www.gamesbyangelina.org

References

Cook, M., and Colton, S. 2011. Automated collage generation - with more intent. In *Proceedings of the Second International Conference on Computational Creativity*.

SynAPP: An online web application for exploring creativity

Alberto Gael Abadín Martínez, Universidade de Vigo (Spain) / AGH-UST (Cracow, Poland)
Bipin Indurkhya, AGH University of Science and Technology (Cracow, Poland)
Juan Carlos Burguillo Rial, Universidade de Vigo (Spain)

DESCRIPTION

SynAPP is a web application currently hosted at AGH-UST (http://149.156.205.250:15180) designed to stimulate users' creative skills through image association tasks and a rating feedback system. In SynAPP users perform two tasks related to image-image associations:

-Associating two images using a word or short phrase. The two images can be presented simultaneously, left and right, or sequentially with a five seconds delay in between. The user is allowed to make only one association per couple.

-Evaluating associations generated by other users according to two criteria: originality (0, 0.5 or 1 points) and intelligibility (0, 0.5 or 1 points).

The set of image pairs for these two tasks are mutually disjoint, so if a user generates association for an image pair, then she or he does not evaluate associations generated by other users for the same pair; and vice versa.

All the responses are recorded with their respective time stamps, and the time taken to perform each association is also recorded. A user can see how her or his associations were rated (with respect to their originality and intelligibility) by other users, and also how this evaluation evolved over time. This information is shown in an intuitive way using tables and graphs.

Users perform three standard tests of creativity before and after using SynAPP:

-Will Shortz & Morgan Worthy's word equation (ditloid) puzzles like "24 = H. in O. D." ("24 = Hours in One Day"). Different equations are used for before-SynAPP and after-SynAPP tests.

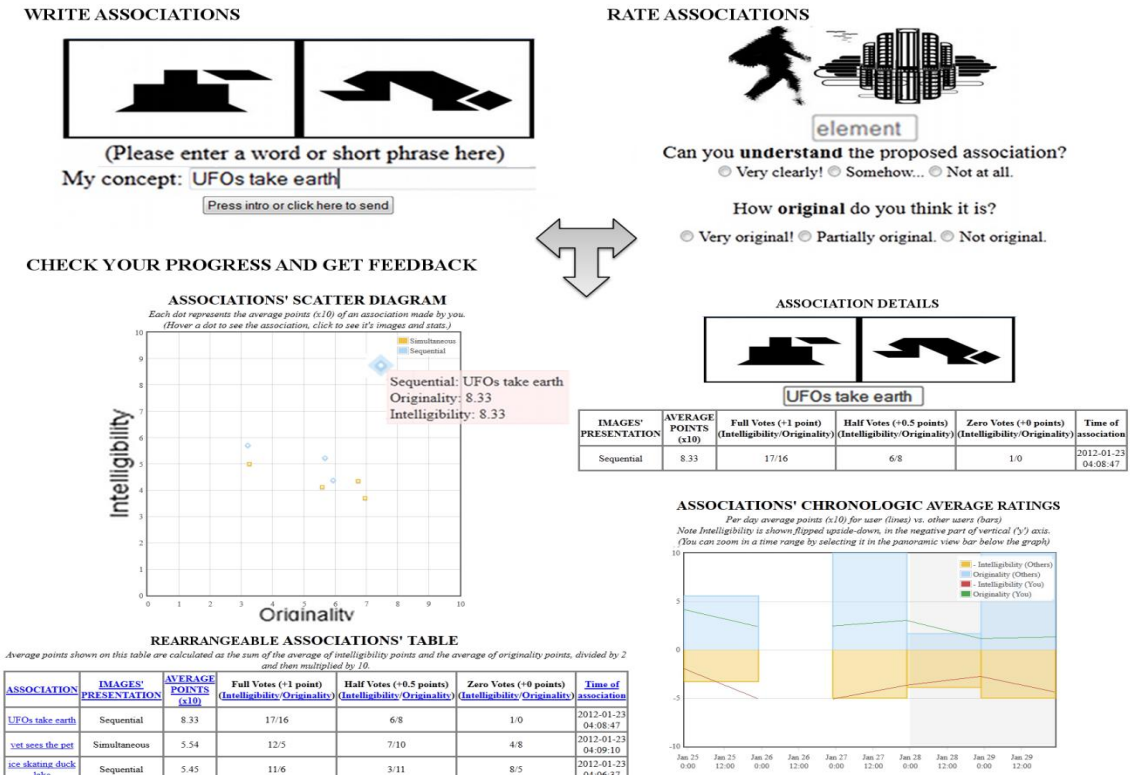
-Guilford's alternative uses task: the user is asked to give as many uses as possible of a common item. Different objects are used for before-SynAPP and after-SynAPP tests.

-Wallace & Kogan's assessment of creativity: A test similar to Guilford's, but the user is asked to find objects with a common property instead.

The answers given by each user are evaluated by the other users, similar to the image associations, and statistics on these evaluations is also displayed graphically to the user.

We hypothesize that SynAPP helps users to improve their creative, out-of-the-box divergent thinking cognitive abilities, and our goal is to properly evaluate this hypothesis based on the analysis of the data collected from the association tasks and the creativity tests.

APPLICATION WORKFLOW



Co-creating Game Content using an Adaptive Model of User Taste

Antonios Liapis, Georgios N. Yannakakis, and Julian Togelius
Center for Computer Games Research
IT University of Copenhagen
Rued Langgaards Vej 7, 2300 Copenhagen, Denmark
{anli, yannakakis, juto}@itu.dk

Mixed-initiative procedural content generation can augment and assist human creativity by allowing the algorithm to take care of the mechanisable parts of content creation, such as consistency and playability checking. But it can also enhance human creativity by suggesting new directions and structures, which the designer can choose to adopt or not.

The proposed framework generates spaceship hulls and their weapon and thruster topologies in order to match a user's visual taste as well as conform to a number of constraints aimed for playability and game balance. The 2D shapes representing the spaceship hulls are encoded as pattern-producing networks (CPPNs) and evolved in two populations using the feasible-infeasible 2-population approach (FI-2pop). One population contains spaceships which fail ad-hoc constraints pertaining to rendering, physics simulation and game balance, and individuals in this population are optimized towards minimizing their distance to feasibility. The second population contains feasible spaceships, which are optimized according to ten fitness dimensions pertaining to common attributes of visual taste such as symmetry, weight distribution, simplicity and size. These fitness dimensions are aggregated into a weighted sum which is used as the feasible population's fitness function — the weights in this quality approximation are adjusted according to a user's selection among a set of presented spaceships. This adaptive aesthetic model aims to enhance the visual patterns behind the user's selection and minimize visual patterns of unselected content, thus generating a completely new set of spaceships which more accurately match the user's tastes. A small number of user selections allows the system to recognize their preference, minimizing user fatigue.

The proposed two-step adaptation system, where (1) the user implicitly adjusts their preference model through content selection and (2) the preference model affects the patterns of generated content, should demonstrate the potential of a flexible tool both for personalizing game content to an end-user's visual taste but also for inspiring a designer's creative task with content guaranteed to be playable, novel and yet conforming to the intended visual style.

Related Work

A. Liapis, G. N. Yannakakis, and J. Togelius, "Adapting Models of Visual Aesthetics for Personalized Content Creation," *IEEE Transactions on Computational Intelligence and AI in Games, Special Issue on Computational Aesthetics in Games,* 2012, (to appear).

A. Liapis, G. N. Yannakakis, and J. Togelius, "Optimizing Visual Properties of Game Content Through Neuroevolution," in *Artificial Intelligence for Interactive Digital Entertainment Conference,* 2011.

A. Liapis, G. N. Yannakakis, and J. Togelius, "Neuroevolutionary Constrained Optimization for Content Creation," in *Computational Intelligence and Games (CIG), 2011 IEEE Conference on,* 2011, pp. 71–78.

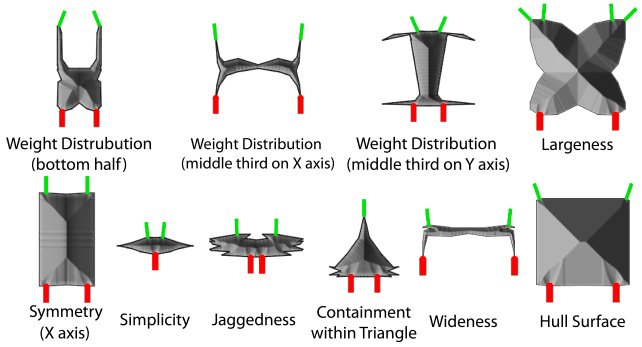


Figure 1: The fitness dimensions used to evaluate spaceships' visual properties and sample spaceships optimized for each fitness dimension. Weapons are displayed in green and thrusters in red.

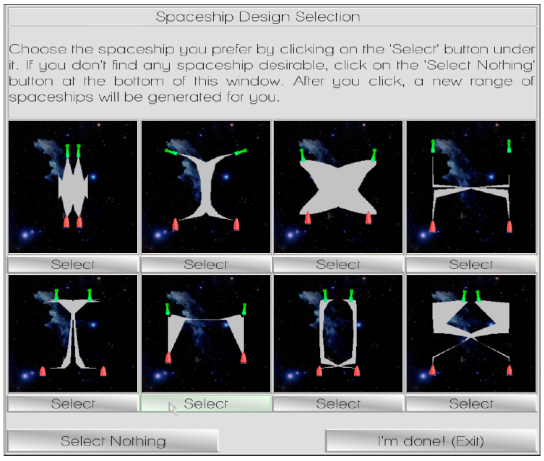


Figure 2: The graphic user interface for spaceship selection.