Process Based Evaluation of Computer Generated Poetry

Conference Paper · January 2016

DOI: 10.18653/v1/W16-5508

Stephen Mcgregor
Queen Mary, University of London

Matthew Purver
Queen Mary, University of London

Geraint Anthony Wiggins
Queen Mary, University of London

Process Based Evaluation of Computer Generated Poetry

Stephen McGregor and **Matthew Purver** and **Geraint Wiggins**
Queen Mary University of London
School of Electronic Engineering and Computer Science
s.e.mcgregor@qmul.ac.uk m.purver@qmul.ac.uk
geraint.wiggins@qmul.ac.uk

Abstract

This paper presents and evaluates a novel system for computer generated poetry. Framed within contemporary theoretical trends in the evaluation of computational creativity, we investigate how evidence of generative process influences readers' opinions of computer generated textual output. In addition to a technical description of our system, we present results from a study asking respondents to evaluate short computer generated poems prefaced with different types of descriptions, in some cases objectively presenting the poem as the product of a statistical analysis of corpora and in some cases subjectively presenting the computer as a self-aware agent.

1 Introduction

The trope of the poet as inscrutable genius figures large in our collective cultural appreciation of poetry. Coleridge emerging from a drunken stupor with the lines to "Kubla Kahn" fully formed in his mind, Blake hallucinating trees full of angels, the drunken, stoned verse of Rimbaud and Verlaine, the psychic divinations of Breton and Soupault: regardless of the legitimacy of these legends, we as readers are seduced by the idea of the poet as a transmitter of the ineffable, tapped into an mental space inaccessible and unknowable to most of us.

Of course when it comes to computers, we are not willing to give them this kind of credit, nor should we be. When we encounter a machine that produces exemplary poetry, we suspect there might be an element of human interference lurking in the mechanism. Such output, without any explanation of the generative procedure employed by the system, including its engagement with a corpus of relevant extant cultural artefacts, is subject to suspicions of pastiche or even plagiarism. The burden of creative justification is on the system itself: it is reasonable to expect a creative computational agent to justify its output in terms of the way in which it was generated, and in particular to demonstrate the way in which its procedures can be ostensibly construed as instances of autonomous engagement with an existing *inspiring set* (Ritchie, 2012). The judgment of discerning observers of computational output will ultimately be influenced by the effectiveness of this presentation of process.

In this study, we systematically test the difference between how human readers react to poems generated by computers when the computational process is, on the one hand, framed as a procedure of statistical analysis, and, on the other hand, as a creative endeavour undertaken by an autonomous and ostensibly self-aware agent. In both cases, we are exploring the ways in which humans react to creative artefacts which have been openly generated by computers; in this work, we are not concerned with exploring the ability of human observers to distinguish between work created by other humans versus output covertly generated by computational processes. The appreciation of creative work is always a moving target, in that popular opinions of what qualifies as innovation is perpetually evolving, and our stance is that public consideration of art is moving towards a point where the idea of creative machines is becoming increasingly palatable. In this regard in particular, we feel that poetry generated by processes transparently

Proceedings of the INLG 2016 Workshop on Computational Creativity and Natural Language Generation, pages 51–60,
Edinburgh, September 2016. ©2016 Association for Computational Linguistics

grounded in the machine learning paradigm will be judged favourably.

In this paper, we'll begin with a overview of computational creativity, focusing in particular on ideas about the evaluation of not only creative artefacts but also creative processes. The nature of our study is motivated by an examination of some specific examples of computational poetry generating systems. In Section 3, we'll outline the technical details of a novel system for generating short, simple poems, grounded in statistical analyses of large sets of textual data. In Section 4, we'll present a study collecting evaluations of our systems output influenced by different ways of presenting the computational process behind the generation of the output. In our final analysis, we will discover that the procedural presentation does not, in fact, influence the ratings returned by readers, at least to a statistically significant degree, and at least for the type of highly autonomous output produced by the system which we'll describe here.

2 Computers, Creativity, and Poetry

This paper is in particular concerned with the evaluation of computer generated poetry. With this in mind, however, an overview of recent and ongoing general trends in the field of computational creativity seems an appropriate starting point. In particular here we're concerned with presenting some thoughts on the question of the evaluation of creative work undertaken by computational agents, and in particular the issue of the assessment of computational process as a critical element in this kind of evaluation.

2.1 Evaluating Computational Creativity

While the concept of the Turing Test – the behaviouralist assessment of a symbol manipulating system sheerly on the basis of its output – has captured the popular imagination, the field of Computational Creativity has probably since its inception been concerned not only with the evaluation of artefacts produced by machines but also with the perception of the machine itself as a producer. (Boden, 1990), for instance, is generally concerned with the importance of self-evaluation in the creative process, and in particular considers the way in which the "computer's performance" (p. 159) contributes to the perception of aesthetic value in the case of computer generated art. More recently, (Colton and Wiggins, 2012) have advocated "assessing the behaviour of software via process rather than product" (p. 24), by way of creative systems "*framing* their creative acts with information that adds value" (p. 25, emphasis in original).

The work presented here has been undertaken very much in this spirit of offering the computational process behind the generation of our system's output as an element of the artefact itself. In fact, in line with (Jordanous, 2015), we feel that much of what counts as creativity exists not merely within the creative agent, but also in the dynamic between agent, audience, and environment. In the specific case of the new system for poetry generation which will be described throughout Section 3, the computational agent engages with the world through a set of statistical analyses with large scale, highly public corpora, spanning the canonical and the encyclopedic. Our hope is that, on multiple levels, this kind of engagement with data-in-the-world or, alternatively, world-as-data offers a perspective on an agent which is situated in an accessible and even familiar environment.

Notably, this idea of statistical analysis as environmental grounding has likewise been adopted by the field of cognitive science, where, for instance, (Barsalou, 2008) has proposed the integration of statistical information about words and linguistic structures as part of a model of cognition as grounded in dynamic environmental processes. The upshot of this kind of theory is that there is some hope of understanding the seam between words and ideas in terms of the data that is available in large scale corpora, that cultural level of linguistic phenomena between the evolutionary and the developmental which has been described by (Smith et al., 2003) as *glossogenetic*. For this reason, we think that the machine learning paradigm in particular, which takes as its basis corpora on the comprehensive scale of large cultural repositories such as an exhaustive encyclopedia or a literary canon, is an appropriate setting for exploring computer generated poetry as a creative process.

2.2 Computer Generated Poetry

The prevalent trend in computer generated poetry to date has involved a combination of rule-based manipulations of symbols and clever heuristic data mining designed to populate templates affording varying degrees of freedom. The WASP system (Gervás, 2000), for instance, uses a battery of "judges" to evaluate an unfolding "draft" of a poem along a series of criteria such as rhyme, scansion, line length, and so forth. The resulting poem is a product of the interaction of these various weighted constraints, coupled with n-gram driven text generation based on an analysis of a corpus of canonical Spanish poetry. Similarly, PoeTryMe (Oliveira, 2012) employs a network of information processing nodes that interact to generate grammatical, metrical verse.

Moving into a more statistical mode of production, (Toivanen et al., 2012) describe a poetry generating system which discovers semantic relationships based on word co-occurrence statistics in a large scale corpus. In addition to this statistical technique for modelling semantics, this system imposes additional syntactic and phonological constraints on its output, and in this regard is comparable with the system described in this paper. Also within the general family of statistical, corpus based models, Haiku generation in particular has been a target for vector space model approaches to computational poetry. Gaiku (Netzer et al., 2009), for instance, uses a combination of human generated word association norms and sequences of syntax derived from a statistical analysis of a corpus of existing haiku to generate new haiku which are designed to be as meaningful, grammatical, and poetic as possible. The First Sally system for Haiku generation (Droog-Hayes and Wiggins, 2015) uses a distributional semantic model, based on an analysis of word co-occurrences in a large scale textual corpus, to generate sets of conceptually related words, and in this regard is closely related to the semantic element of the new system described in Section 3.1.

Of particular interest to the study presented here is the system for poetry generation based on the FACE model for assessing computational creativity (Colton et al., 2012). This model focuses on the evaluation of creativity associated not just with assessment of the artefact generated by the system, but also with observation of the process which the system undertakes to produce its output. With regard to poetry in particular, the model architects are concerned with the critical conveyance of "communicative purpose" (p. 96) which is essential to the understanding of linguistic expression: as consumers of poetry, we rely on the belief that something more than just a random or cleverly constrained but decontextualised process lies at the other end of the poem itself. In short, we count on meaning being anchored in intent.

In the case of the poetic implementation of the FACE model, this has meant that poems are coupled with expository statements regarding data analysis that has served as a situation specific motivation for the generation of each poem. The system itself operates by way of template completion, inserting into prefigured lines of verse similes mined from the web using a pattern fitting heuristic to determine viable word combinations. In order to convey a sense of intent to its output, the system weights the phrases it extracts from the web based on a sentiment analysis, seeking to choose similes which correspond in sentiment with a similarly analysed selection of text from a current newspaper. The idea here is that, by rooting the poem in the mood of a currently or recently unfolding event, the system's output becomes tied to something happening in the world, and the reader becomes more committed to the idea that the computer is an agent creating an artefact in reaction to a situation. In particular, the system frames its explanation as a first-person exposition involving an analysis of the mood of the news on a given day, with a degree of justification for this analysis: the system presents itself as a willful actor knowingly engaged in a creative, interpretive process.

Our hypothesis is that the system based on the FACE model, when it comes to the evaluation of computer generated poetry, has got it at least half right, in that the perception of a creative procedure underlying the computational generation of poetry is a crucial factor in the creative quality of the poetic artefact. And one way to convey a creative procedure is to couch the operation of the computer in a narrative of the machine having a self-reflective sense of goal-directedness, a kind of transparent fiction of agency exploiting the human tendency to read intentions and beliefs into all sorts of situations in the world where we know there actually are none (Carruthers, 2011). We believe, though, that readers of poetry are at this stage in the history of technology and art prepared to engage with computer produced verse in a more frank way, acknowledging the statistical character of the underlying operation without losing regard for the inherent degree of creativity, and in fact possibly taking the output more seriously when the generative procedure is presented in a straightforward, objective manner.

3 Autonomous and Contextual Poetry

In order to evaluate human assessment of both creative process and output, we have designed a relatively straightforward system for generating short, loosely constrained poems. This system has been designed with three critical principles in mind:

- The system uses a machine learning technique for the unsupervised generation of semantic relationships.
- The semantic relationships which serve as one of the constraints on the systems output are context sensitive, and in this way can be associated with *ad hoc* input allowing the poems to be *about* something topical.
- The system uses a statistical technique to constrain the phonology of the poem, and so is designed to produce text that sounds poetic.

Over the course of this section, we'll lay out a series of models which are algoritmically concatenated into a system which seeks to fulfill these requirement.

3.1 A Semantic Model

At its root this system is based on a statistical model of word meaning constructed within the distributional semantic paradigm, construing words as vectors within a space of dimensions representing co-occurrences with other words over the course of a large-scale textual corpus (Turney and Patel, 2010; Clark, 2015). The key feature of this particular model is its context sensitivity: it dynamically generates new semantic spaces based on an analysis of the conceptual relationships between a set of input terms (McGregor et al., 2015; Agres et al., 2015). The objective of this component of the poetry generating system is to generate spaces in which the conceptual relationships between words

The motivation for using this particular model is twofold. For one thing, the model derives its features from an unsupervised traversal of a corpus, so the semantic relationships which it captures are discovered without the human dictated assignment of symbol manipulating rules. This property ostensibly gives the poetry generating system at least a semblance of agency. And, on the other hand, the dynamic, contextual component of the model enables it to engage with *ad hoc* input, allowing the model to generate output topically related to other textual artefacts. This means there is some hope of conveying a sense of intentionality or aboutness to an observer of the system's process.

This semantic model is based on a very high-dimensional (approximately 7.5 million), very sparse space of word-vectors generated from a traversal of the English language component of Wikipedia. The dimensions of this space correspond to terms that co-occur in sentences with words from the model's 200,000 word vocabulary. The value for each dimension is based on a pointwise mutual information metric derived as follows, where $n_{w,c}$ corresponds to the frequency with which vocabulary word w co-occurs with context word c, W is the overall count of vocabulary word tokens, n_w and n_c are the respective independent frequencies of w and c, and a is a smoothing constant:

$$M_{w,c} = \log_2 \left(\frac{n_{w,c} \times W}{n_w \times (n_c + a)} + 1 \right) \qquad (1)$$

The sparse space generated through this process can be reduced to a context dependent, conceptually oriented subspace through an analsyis of a set of input terms. So, for instance, in a 200 dimensional subspace based on co-occurrence dimensions salient to the words `cat`, `dog`, and `goldfish`, `cat` is closest to the words like `rabbit`, `hamster`, and `pet`. If, on the other hand, we build a subspace based on the input words `cat`, `lion`, and `tiger`, `cat` becomes proximate to words like `leopard`, `hyena`, and `wild`. Technical details for generating subspaces are laid out in detail in (McGregor et al., 2015).

3.2 A Phonological Model

This process of building a space of potential subspaces is coupled with a phonological model which similarly uses an information theoretic metric to try to capture the way in which word-sounds are expected to co-occur in poetry. This model is also constructed from a statistical model of a corpus, in this instance a corpus containing about 1500 English language sonnets.[1] These sonnets are rendered into a format containing both phonemic and syllabic information, based on a syllabified version of the CMU Arpabet (Bartlett et al., 2009). Frequencies of phonemic co-occurrence $C_i(p_a, p_b)$ are then tabulated, where the count C is the total number of times phoneme p_b occurs i syllables in front of phoneme p_b in a line of a poem. Once all frequencies for all lines in all poems in the corpus are compiled, these statistics are converted into mutual information measures, formulated here with $C_i(T)$ representing the total number of phonemes occurring i syllables apart and $C_i(p_a)$ and $C_-i(p_b)$ standing for the independent frequencies at which phonemes p_a and p_b occur i or $-i$ syllables away respectively from any other syllable:

$$P_i(p_a, p_b) = \log_2 \left(\frac{C_i(p_a, p_b)C_i(T)}{C_i(p_a)C_-i(p_b)} + 1 \right) \quad (2)$$

From this matrix of phoneme-distance relationships, a score can be generated for the phonological strength of any two given candidate syllables s_1 and s_2 potentially occurring in a line of poetry generated by the system, where l_1 and l_2 are the respective lengths of s_1 and s_2, and p_1 and p_2 are corresponding constituent phonemes:

$$S_i(s_1, s_2) = \frac{1}{l_1 l_2} \times \sum_{p_1=1}^{l_1} \sum_{p_2=1}^{l_2} P_i(p_1, p_2) \quad (3)$$

This phonological model is incorporated into our poetry generating system in order to impose a sense of prosody on the output. As with the semantic model, there are no phonetic or metric constraints hand-coded by human designers, and so we can claim that, to the degree that prosodic features do emerge in the system's output, these elements are discovered by the system itself as statistical properties inherent to the underlying corpus of sonnets.

3.3 A Syntactic Model

The third constraint placed on our poetry generating system consists of an n-gram model for stringing together parts of speech in ostensibly syntactic ways. Statistics are once again harvested from the corpus of about 1,500 English language sonnets, in this case with each word tagged with a part of speech label using the Python Natural Language Toolkit word tokeniser.[2] Once these tagged renditions of the corpus are generated, a probabilistic model for predicting the syntactic continuation of a string of parts of speech is built, describe here with $n_{t,q}$ representing the frequency with which part of speech t follows the sequence of parts of speech q in a line of poetry, and n_q signifying the total number of times the sequence q is observed in any line:

$$G(t|q) = \frac{n_{t,q}}{n_q} \quad (4)$$

If, in the course of generating a line verse, the system generates a sequence q that has no observed extension, is will remove the first element in q to produce sequence q' and will then generate element t with probability $G(t|q')$. The purpose of this operation is to impose an arguably superficial element of grammaticality on the system's output. Anecdotally, but also significantly, professional poets who have interacted with the system have actually suggested that this component of the process over-constrains the output to the detriment of the interesting conceptual and phonological relationships generated by the other models.[3] Nonetheless, for the purpose of the comparative study presented here, this component of the system is maintained. Also of note is that this syntactic model is the only component of the system that simulates a non-deterministic process.

3.4 Sentiment Analysis

The final aspect of the poetry generating system is a model for analysing the sentiment of a document

[1] www.sonnets.org.

[2] www.nltk.org/_modules/nltk/tokenize.html

[3] In the course of the Globe Road Poetry Festival at Queen Mary in November 2015 and the Portrait of the Machine as a Young Artist event at the British Library in February 2016.

within a corpus. In the case of the poems used for the study here, the corpus in question is the Penn Treebank Switchboard corpus, consisting of 1,126 transcribed telephone conversations.[4] A straightforward term frequency-inverse document frequency technique is employed in order to create a topic model for each conversation within the corpus (Salton and McGill, 1983). Specifically, for a given document (conversation) d, the words representing the salient topics of the conversation are ranked according to this equation, where w is a word that occurs within the document, w_d is the number of times w appears in d, and w_c is the number of times w occurs in the entire corpus of conversations:

$$T(d, w) = \frac{w_d}{w_c} \tag{5}$$

For each conversation in the corpus, the top four topical terms based on the above equation are selected, and the sentiment of these terms is rated along a negative-positive spectrum. The rating for a given word is derived from the SentiWordNet database of word sentiment scores, which assigns negative and positive ratings to senses of a large number of words.[5] In order to rank the sentiment of each conversation, each word is assigned the mean score of its various sentences, and then the average scores of the four most salient terms is taken to give each conversation an overall ranking.

The purpose of analysing the sentiment of a transcribed conversation is to give the poetry generating system a topic as a topical handle, allowing the poem to be about something specific and intersubjective. The idea, following on from (Colton et al., 2012), is that a poem that is endowed with intentionality is more likely to be deemed as creative by an observer.

3.5 Assembling a Poem

Finally, the various modular components described above are linked together to algorithmically generate poems for subsequent analysis by human readers according to the following procedure:

1. The 17 most negative and 16 most positive conversations, ranked as described in Section 3.4, are selected as topics for poems.

2. The four topics of each conversation are fed into the semantic model describe in Section 3.1. A subspace of conceptually related words is generated, with the salient region of this space considered that which is closest to the mean of the topical input terms. The words in this subspace are tagged with their most likely part of speech.

3. A syntactic string is probabilistically generated based on Equation 4, and a line of poetry of no more than 11 syllables is correspondingly composed.

4. At each step in the generative process, the word that is closest to the salient region of the space described in Step 2, aside from the input terms themselves, that matches the next part of speech prescribed by Step 3 is choosen as a continuation of the line being composed. A base poem of four lines is generated.

5. Each word in each line is given a score of phonological appropriateness based on the average score of each of its syllables compared with all other syllables in the line, including the other syllables in the word itself. This phonological score is then multiplied by $1 - (z \times sent(w))$, where $sent(w)$ is the sentimental rating of word w according to the SentiWordNet database, while the value of z is -1 if the overall sentiment of the input terms is negative and +1 if the prevalent sentiment is positive.

6. The least appropriate word in the poem is removed and replaced with the most appropriate word, selected from a vocabulary defined in terms of the 1,000 most conceptually salient words as established in the subspace derived in Step 2. Steps 5 and 6 are repeated until the poem converges to a maximally scored state.

The final product of this process is intended to be a poem which is conceptually relevant to the conversation serving as the basis for the input terms while exhibiting poetic phonology, sentiment appropriate to the input topic, and a modicum of grammaticality.

[4] www.cis.upenn.edu/˜ treebank/

[5] http://sentiwordnet.isti.cnr.it/

4 Evaluation of Process and Product

Based on the generative process described throughout Section 3, 33 poems have been randomly generated, each associated with a conversation summarised by the four words derived through the technique described in Section 3.4. We have subsequently generated three different versions of these poems, one prefaced with a brief objective description of the generative process, one prefaced with a brief subjective description framing the system as a self-aware agent, and one with no preface at all. An example of the objective preface is as follows:

> *This poem is based on a sentimental and conceptual analysis of a conversation containing words like 'sickening', 'shitty', 'novice', and 'hack'. The sentimental component of the analysis determined the conversation was negative. The poem emerged as a pathway in a space of word points derived from this statistical analysis, with an additional criterion for selecting poetic sounding combinations of words.*

And the subjective description of the same poem reads as follows:

> *I listened to a conversation containing words like 'sickening', 'shitty', 'novice', and 'hack'. I considered this to be a negative conversation. I decided to write a poem about this conversation, and have tried to capture some of the negative sentiment while also focusing on how the poem sounds.*

Finally, the poem that accompanies these descriptions reads like this:

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

We constructed a survey consisting of a total of 99 poems: each of the 33 poems our system generated, with each of the three versions of the explanatory preface (or lack thereof). Each survey participant was first presented with a introduction page laying out the survey, informing them that they would be reading a poem generated by a computer and then asked to evaluate the poem. On the next page, the poem itself was presented, preceded by either one of the two types of procedural descriptions illustrated above or by no description at all. On the same page, subjects were asked to evaluate the poem they had just read based on three different criteria: *creativity*, *meaningfulness*, and *quality*, in each case giving the poem a rating along a seven point scale ranging from "low" to "high". Finally the subjects were presented with a third page where they were asked to provide optional information about their age and their self-assessed proficiency or knowledge in the English language, poetry, and computer science, again in each case rating themselves along a seven-point scale, in these instances ranging from "novice" to "expert".

	creativity	meaningful	quality
obj	3.14 (*1.88*)	1.67 (*0.78*)	2.05 (*1.05*)
subj	2.93 (*1.63*)	2.00 (*1.32*)	2.07 (*0.96*)
none	2.93 (*1.60*)	1.54 (*0.63*)	2.14 (*1.33*)

Table 1: Mean scores along a seven-point scale (with standard deviations in parentheses) for human subject evaluations of creativity, meaningfulness, and quality of computer generated poems prefaced with an objective description of the generative process, a subjective description, or no description at all.

We received responses from 79 participants, with each participant evaluating a unique preface-poem combination. Reported ages ranged from 20 to 72, with a mean of 40. The mean value for proficiency in English was 6.26, with standard deviation of 0.92; for knowledge of poetry, the mean was 4.12 with stdv of 1.46; for knowledge of computer science, the mean was 4.66, with stdv of 1.96. The mean responses, along with standard deviations, are presented in Table 4.

The overall picture these results paint is that, in the case of the type of poetry being generated by our system, the mode of presentation has a marginal effect on the evaluation of content. The higher value of creativity typically accorded to poems presented with an objective description correlates with our hypothesis that readers would react favourably to this transparent presentation of process, by the difference between this value and the mean creativity assigned to poems subjectively framed is not statistically significant: a two-tailed Student's t-test on the results gives a p-value of 0.68 and a t-value of

0.42. The relatively similar mean scores, combined with high degrees of standard deviation, indicate that these results, at least in terms of a comparison between the data for each type of presentation, aren't distinguishable from what we would expect if subjects randomly assigned values to poems.

Also of not is the relatively high scores given to the subjectively presented poems in terms of the meaningfulness of the poems. Statistical significance is slightly higher here, with a p-value of 0.31 and a t-value of 1.02, but still hardly noteworthy. The one thing that does perhaps bear further consideration here is the way that subjects seem relatively comfortable ascribing creativity to poems presented as products of statistical processes versus the meaningfulness attributed to poems framed as subjective experiences of information in and about the world. Perhaps the appropriate interpretation here is that readers appreciate the insight into the productive mechanism afforded by the objective presentation, and associate this with creativity, whereas meaningfulness is more closely connected to the impression of agency and individuation conveyed by the subjective presentation.

Finally, it is also worth mentioning that the poems presented with no procedural description at all do just about as well as the lesser of the two explained poems in terms of creativity and meaningfulness, and actually do slightly better than the other two types in terms of quality. Quality is arguably a somewhat vague category, and was intended to cover a range of properties such as poeticness and composition. On the whole, though, the story here seems to be that, at least in terms of this type of poetry, with the relatively cursory kind of procedural description we were able to offer in the course of a survey that was, by design, quite brief, the way that the poems are presented doesn't make a big difference in terms of how humans rate this type of output.

5 Conclusion

Further to the brief analysis offered above, another point of interest with this study relates to the relatively high degree of standard deviation evident across all the results. The story here would seem to be that there is a wide range of opinion on how exactly computer generated poetry should be evaluated in the first place. Anecdotally, responses in most categories for most types of presentation ranged from one to seven, despite all of the 33 poems being of a generally similar quality. There seems to be a lack of consensus regarding how to consider computers as poets.

This analysis aligns with the feedback received in the course of the the events involving engagement between human poets and computational systems for poetry generation mentioned in Section 3.3. Specifically, a self-selecting group of technologically receptive poets found much value in engaging with the system described here, which they saw as a mechanism for discovering interesting, novel, and potentially productive conceptual concordances within a corpus which were obscure to a human reader but nonetheless poetically valuable. This approach to poetry as an artefact of a dynamic engagement between poets, readers, corpora, society, and the environment is conducive to the type of poetry generated by our system—but this particular aesthetic stance is hardly universal in the world of poetry readers.

Compared to the output of the Full FACE system, the output of our system is, more or less objectively, more garbled and less structured. On the other hand, the FACE system resorts to heuristic simile mining and template filling, where our system maintains a somewhat higher degree of autonomy in its analysis of a corpus and dynamic projection of conceptually loaded semantic subspaces. Whether readers provided with more comprehensive descriptions of the differences between these approaches would consider one system more creative than the other remains to be seen, and is beside the point of the study presented here, which has been a first attempt at assessing whether or not the way that the creative process involved in the computational production of poetry is framed has a significant impact on evaluation of output.

Returning to our earlier discussion of creativity as a phenomenon dynamically distributed across a society and an environment, we ultimately expect evaluations of creativity to take into account various factors integrating the overall situation of an artefact. So, for instance, in the case of poetry, we would predict that the relationship between a poem, its mode of production, and the milieu in which the poem is produced should all contribute to the assessment of the inherent creativity, quality, and meaningfulness of both the poem itself and the poetic act. Similar insight seems to have motivated the implementation of the FACE model that has been discussed here, which seeks to act as an agent of both generation and interpretation. While the study described in this paper has focused on the effect of procedural description on poetic evaluation, we might conjecture that the dynamically context-sensitive model that provides the conceptual component of our system is the right kind of computational process to offer a compelling platform for environmental situatedness.

The criteria for evaluating creativity discussed here, construed in terms of three values and presented to study participants without any further explanation, admittedly offer a relatively blunt approach to judging the merit of computational output, let alone to assessing the more general creative process and the relationship between this process and its situation in the world. In the future, in addition to improvements to the system itself, further advances to this work would involve the construction of an evaluative mechanism which incorporates a more complete description of the system's operation and environmental situation, as well as a more nuanced range of questions to enrich the evaluative process. For now, the outcome seems to be that there is still too varied an attitude towards what it means for a computer to claim creative autonomy for there to be a meaningful consensus on the merit of poetic output based on a relatively straightforward encounter with a poetry generating computer.

Acknowledgments

This research has been supported by EPSRC grant EP/L50483X/1. The project ConCreTe acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 611733.

References

Kat Agres, Stephen McGregor, Matthew Purver, and Geraint Wiggins. 2015. Conceptualising creativity: From distributional semantics to conceptual spaces. In *Proceedings of the 6th International Conference on Computational Creativity*, Park City, UT.

Lawrence W. Barsalou. 2008. Grounded cognition. *Annual Review of Psychology*, 59:617–645.

Susan Bartlett, Grzegorz Kondrak, and Colin Cherry. 2009. On the syllabification of phonemes. *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 308–316.

Margaret A. Boden. 1990. *The Creative Mind: Myths and Mechanisms*. Weidenfeld and Nicolson, London.

Peter Carruthers. 2011. *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford University Press.

Stephen Clark. 2015. Vector space models of lexical meaning. In Shalom Lappin and Chris Fox, editors, *The Handbook of Contemporary Semantic Theory*. Wiley-Blackwell.

Simon Colton and Geraint Wiggins. 2012. Computational creativity: The final frontier? In *Proceedings of the 21st European Conference on Artificial Intelligence*.

Simon Colton, Jacob Goodwin, and Tony Veale. 2012. Full-FACE poetry generation. *Proceedings of the Third International Conference on Computational Creativity*, pages 95–102.

Max Droog-Hayes and Geraint A. Wiggins. 2015. Improved meaning in poetry using statistical methods. In *Proceedings of the Sixth International Conference on Computational Creativity*.

Pablo Gervás. 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*, pages 93–100.

Anna Jordanous. 2015. Four pppperspectives on computational creativity. In *Proceedings of AISB 2015s Second International Symposium on Computational Creativity*.

Stephen McGregor, Kat Agres, Matthew Purver, and Geraint Wiggins. 2015. From distributional semantics to conceptual spaces: A novel computational method for concept creation. *Journal of Artificial General Intelligence*.

Yael Netzer, David Gabay, Yoav Goldberg, and Michael Elhadad. 2009. Gaiku : Generating haiku with word associations norms. In *Proceedings of the NAACL HLT Workshop on Computational Approaches to Linguistic Creativity*, pages 32–39.

Hugo Gonçalo Oliveira. 2012. PoeTryMe: A versatile platform for poetry generation. In *Proceedings of the ECAI 2012 Workshop on Computational Creativity, Concept Invention, and General Intelligence*.

Graeme Ritchie. 2012. A closer look at creativity as search. In *Proceedings of the 2012 International Conference on Computational Creativity*.

Gerard Salton and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company, New York, NY.

Kenny Smith, Henry Brighton, and Simon Kirby. 2003. Complex systems in language evolution: The cultural emergence of compositional structure. *Advances in Complex Systems*, 06(04):537–558.

Jukka Toivanen, Hannu Toivonen, Alessandro Valitutti, and Oskar Gross. 2012. Corpus-based generation of content and form in poetry. In *Proceedings of the Third International Conference on Computational Creativity*, pages 175–179.

Peter D. Turney and Patrick Patel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.