Evaluating Mixed-Initiative Creative Interfaces via Expressive Range Coverage Analysis

Max Kreminski[1], Isaac Karth[1], Michael Mateas[1] and Noah Wardrip-Fruin[1]

[1]*University of California, Santa Cruz, 1156 High St, Santa Cruz, CA 95064, USA*

Abstract

We introduce *expressive range coverage analysis* (ERaCA): a technique for evaluating mixed-initiative creative interfaces (MICIs) in which creative responsibility is shared between a human user and a generative model. ERaCA revolves around the examination of a small number of human-created artifacts in the context of a visualization of the broader *expressive range* from which these artifacts were sampled. As a pilot study of our approach, we apply ERaCA to the evaluation of Redactionist—a MICI for erasure poetry creation—and find that ERaCA allows us to visually answer questions about how thoroughly users explore the underlying model's expressive range; whether users produce artifacts that are typical or unusual from the underlying model's perspective; whether different users of a single MICI tend to produce similar or different artifacts; whether a MICI tends to promote divergent or convergent thinking; and how a single user's artifacts evolve as they continue to use a MICI over time.

Keywords

expressive range analysis, mixed-initiative co-creativity, creativity support tools, evaluation methods

1. Introduction

Mixed-initiative creative interfaces [1, 2], or MICIs, are a genre of creativity support tools [3] in which creative responsibility is shared between a human user and an artificially intelligent system. Many MICIs consist of two layers: an underlying generative model that defines a *possibility space* of artifacts—sometimes learned from a corpus of training data [4], sometimes defined by a set of rules or constraints [5]—and a supervening mechanism for navigating this space to locate artifacts that match a user's prompt or intent. In MICIs that take this approach, an artifact's *creation* is synonymous with its *discovery and selection* by a user.

Evaluating the effectiveness of these systems can be difficult, in part because neither the user nor the generative model is solely responsible for the artifacts produced [6]. In particular, a skilled user may be able to coax compelling artifacts from even the most unwieldy MICI, making it difficult to characterize how effectively a MICI supports its users in realizing their creative goals. Additionally, insofar as these tools often lead users to create artifacts that they would not have thought to create before, it is difficult to compare a MICI-plus-user system with the unassisted user in terms of creative capabilities, because the user's original creative intent can be substantially shaped or modified by their interaction with the tool. As a result, assessments of MICIs often focus on evaluating the subjective perception of creativity support from the user's perspective [6]. The artifacts that users produce are comparatively rarely evaluated, and even when they are evaluated, discerning what role the MICI played in shaping these artifacts may still be out of reach.

Though evaluating creativity is difficult in general [7], researchers have developed a number of effective approaches to the evaluation of *computationally creative* systems [8, 9] in which creative responsibility is attributed primarily or solely to the machine [10]. In particular, a technique known as *expressive range analysis* (ERA) [11] can be used to characterize the behavior of a generative model by visualizing its possibility space. This makes it easy to visually compare the *expressive range* of different generative models that produce the same kind of artifact—and to describe a generative model in terms of its *grain*, or the characteristics of the artifacts that it tends to produce [12].

However, because ERA relies on the rapid generation and characterization of a very large number of artifacts [13], this method of evaluation cannot straightforwardly be applied to mixed-initiative creative collaborations. When a human user must be involved in the production of every artifact, it becomes prohibitively time-consuming to produce the hundreds or thousands of artifacts that ERA demands. As a result, although ERA is frequently applied to the evaluation of end-to-end computationally creative systems, including the generative models underlying some MICIs [14], its application to understanding the influence of MICI design on user behavior and user experience has remained limited.

In this paper, we propose a new technique for evaluating MICIs—*expressive range coverage analysis* (ERaCA)—

Joint Proceedings of the ACM IUI Workshops 2022, March 2022, Helsinki, Finland

mkremins@ucsc.edu (M. Kreminski); ikarth@ucsc.edu (I. Karth); mmateas@ucsc.edu (M. Mateas); nwardrip@ucsc.edu (N. Wardrip-Fruin)

https://mkremins.github.io (M. Kreminski)

CEUR Workshop Proceedings (CEUR-WS.org)

that extends ERA to a co-creative context by visualizing a small number of co-created artifacts in the context of the broader expressive range from which these artifacts were sampled. ERaCA applies a set of quantitative artifact evaluation metrics to the simultaneous assessment of many model-created artifacts and a handful of co-created artifacts, then produces a visualization of the results, allowing us to visually answer such questions as:

- Does a MICI allow its users to access the entirety of the underlying generative model's expressive range, or only a limited subset?
- How typical or unusual are the artifacts created by a user in the context of the broader expressive range?
- Are all of a MICI's users drawn toward the same parts of its expressive range, or do different users typically explore different regions of the possibility space?
- As users continue to interact with a MICI, do the artifacts they produce tend to get closer together or further apart within the expressive range? In other words, does the MICI tend to promote convergent or divergent thinking?
- More generally, as users continue to interact with a MICI, what trends appear in a single user's artifacts over time?

We demonstrate ERaCA via a pilot study in which we apply the ERaCA method to the evaluation of Redactionist, a MICI for erasure poetry creation. The resulting visualizations provide preliminary answers to several of the above questions based on data collected from a small number of users. Altogether, the argument for our approach can be summed up as follows: **we learn more about a MICI from inspecting co-created artifacts *in the context of the underlying expressive range* than we do from inspecting both co-created artifacts and the underlying expressive range individually.**

2. Background

Expressive range analysis (ERA) [11] is a visualization-based approach to understanding and evaluating the effectiveness of generative models. Application of ERA follows a four-step approach:

1. **Determine appropriate quantitative metrics** for the kinds of artifacts that the generative model will produce. Ideally these metrics are computationally inexpensive to evaluate, so that they can be efficiently applied to a large number of individual artifacts.
2. **Generate a large number of artifacts** using the generative model to collect a representative sample of the model's output, using the metrics defined in step 1 to evaluate each artifact.
3. **Visualize the results of evaluation**, typically as a set of two-dimensional histograms in which pairs of metrics are plotted against one another to showcase artifact density in different "slices" of the overall expressive range.
4. **Analyze the impact of parameters** passed to the generative model on the resulting expressive range, allowing for the visual determination of how different parameters influence the artifacts that the model produces.

Although ERA has been integrated into tools for human creators of generative models [15], extended in various ways [13, 16], and applied to domains as wide-ranging as emergent narrative [17] and road network generation [18], it has several important limitations. In particular, conventional ERA is data-hungry and poorly suited to the evaluation of small numbers of artifacts, which has prevented its application to creative contexts in which artifacts are individually time-consuming or costly to generate [13]—as is often the case when human users are involved in the creative process.

However, the ideas captured by ERA remain important to the evaluation of tools for human-AI creative collaboration. Among nine potential pitfalls for co-creative systems discussed by Buschek et al. [19], at least five ("Invisible AI boundaries", "Lack of expressive interaction", "Agony of choice", "Time waster", and "AI bias") can be viewed as stemming from either an insufficiently wide expressive range; an expressive range that does not overlap well with user desires; or a flawed user interface for accessing the available expressive range. Evaluations based exclusively on self-reported subjective user experience can produce misleading results [20], leading some to suggest that inspection of co-created artifacts is also needed to arrive at a holistic picture of a co-creative system's success or failure [21, 22, 23, 24]—but even these hybrid evaluations cannot clearly diagnose whether a MICI's weaknesses are due to the underlying generative model or the interface through which the model is accessed. And some studies of user behavior in MICIs have suggested that some users are motivated by a drive to explore the extremes of a MICI's expressive range [25], necessitating the comparison of co-created artifacts against the expressive range to verify these findings. In sum, these difficulties all point to a common unmet need: an evaluation method for MICIs that can illuminate the relationship between individual co-created artifacts and the MICI's overall expressive range.

3. Expressive Range Coverage Analysis

Expressive range coverage analysis (ERaCA) is a new evaluation technique for mixed-initiative creative interfaces (MICIs) in which a human user and a generative model share creative responsibility for the discovery and selection of artifacts from a large possibility space. ERaCA builds on ERA, but also extends the evaluation process by incorporating the solicitation and examination of a small number of co-created artifacts (i.e., artifacts made or discovered by human study participants through their interaction with the MICI) in the context of the generative model's expressive range.

ERaCA as a process consists of seven steps:

1. Determine appropriate quantitative metrics for the kinds of artifacts that the generative model will produce.
2. Generate a large number of artifacts using the generative model, and evaluate each artifact using the metrics defined in step 1.
3. Visualize the results of evaluation.
4. **Solicit the co-creation of a small number of artifacts** by human study participants, ideally drawn from among the MICI's target user base, and evaluate these artifacts using the same metrics that are used to evaluate the purely machine-created ones.
5. **Visualize the location of co-created artifacts** within the context of the larger possibility space, for instance as a set of scatterplots drawn directly on top of the two-dimensional histograms created in step 3.
6. (Optional) **Construct per-user visualizations of the user's trajectory within the possibility space**, using a color gradient to indicate the order in which artifacts were created on the plot. We discuss this visualization approach in greater detail in section 5.4, and an example can be seen in Figure 6.
7. **Visually analyze the results** to make determinations about users' coverage of and trajectory within the generative model's possibility space.

Steps 1-3 of this process are the same as for ERA, while steps 4-7 (which rely on incorporation of co-created artifacts into the evaluation process) are unique to ERaCA.

4. Pilot Study Procedure

In preparation for a larger-scale user study to be conducted in the future, we ran a small-scale pilot study to test and illustrate our approach. Our pilot study used the ERaCA method to evaluate the mixed-initiative erasure poetry creation tool Redactionist on the basis of artifacts created by four participants (all coauthors of this paper) from a single fixed paragraph of source text.

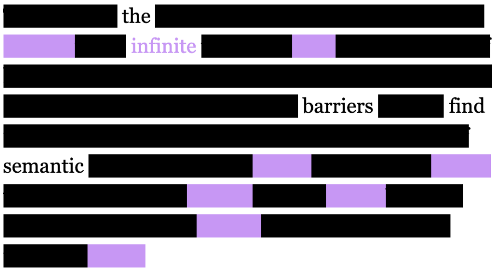

Figure 1: A screenshot of the Redactionist user interface (taken from [26]), showing a partly constructed erasure poem. Words that cannot be used together with the currently selected words are blacked out; words that can be selected are covered by purple boxes, but revealed when the user hovers over them; and selected words are displayed as uncovered text. The user can click on any selected word to unselect it, or on any purple-highlighted word to select it and add it to the poem.

4.1. Redactionist

Redactionist [26], previously known as Blackout [27], is a browser-based[1] casual-creator [28] MICI that helps users create English-language *erasure poetry* by interactively removing most of the words from a user-provided source text. Once given a source text, Redactionist uses a rules-based generative model (adapted from an earlier model created by Liza Daly [29]) to generate a large number of potential erasure poems that could be created from the text. Then it provides the user with an interface for navigating this space of potential poems by toggling whether specific words should be present in the final poem. A screenshot of Redactionist's interface, showing a half-constructed poem, can be seen in Figure 1.

Given a source text, Redactionist's rules look for poems that take the form of several short and grammatically correct declarative sentences—one sentence per paragraph of input text. For instance, one of Redactionist's rules—the grammatical pattern `ARTICLE NOUN VERB ARTICLE ADJECTIVE NOUN`—would find and match sequences of words such as "the poem conceals an elusive metaphor" within a paragraph of source text, with any other words in the source text paragraph being erased. The words in each matched sentence might be separated by any number of other words, as long as they occur in the correct sequence within a single paragraph of the source text. The version of Redactionist used here contains 136 rules, each of which matches sentences of a particular form.

[1] https://mkremins.github.io/blackout/interactive

4.2. Data Collection

Due to logistical constraints (described further in section 6.1), the four coauthors of this paper served as our pilot study participants. Each participant was instructed to use Redactionist with a fixed source text (a one-paragraph excerpt from a transcript of a talk by Allison Parrish on computational poetry [30]) to create a sequence of ten short erasure poems. To ensure that participants were not composing their poems with a particular metric or evaluation criterion in mind, we avoided deciding what metrics would be used to evaluate the poems until after the data had been collected, and we did not confer with one another about our aesthetic intentions for the poems we had made.

In addition to these 40 co-created poems, we also gathered and analyzed the complete set of 57,195 potential poems that the Redactionist generative model considers to be possible erasures of the fixed input text. This larger set of poems, which we call the "full poemspace", forms the backdrop for our analysis: by comparing the 40 co-created poems to the full poemspace, we can identify the co-created poems as typical or atypical in various ways and analyze the extent to which the co-created poems cover (or fail to cover) the full poemspace. For some generative models, it may be easier to instead establish a backdrop set of artifacts by uniformly sampling many (but not all) possible artifacts for the given user input; the details of this sampling vary depending on how the generative model is implemented.

4.3. Artifact Evaluation Metrics

ERaCA, like ERA, uses several domain-specific quantitative metrics to characterize each of the artifacts produced by a generative model or creative collaboration. Erasure poetry is an unusual form of poetry that has not been investigated much in the scholarly literature [31, 32, 33], and the short, single-declarative-sentence poems produced by Redactionist given a single paragraph of input text do not contain many of the features (such as end rhyme) that are most widely studied in the analysis of poetry. Consequently, rather than drawing directly on metrics that have been defined for more conventional forms of poetry [34], we instead defined several preliminary but easy-to-implement metrics of our own that attempt to capture key aesthetic features of erasure poetry as a form. These metrics include:

Average word position within the source text. Erasure poems are characterized partly by the visual spacing of the non-erased words within the source text. Since Redactionist represents poems internally as a set of numerical indexes into the source text pointing to the user-selected words, averaging these indexes together can give a simple approximation of whether a poem mostly contains words taken from near the start, middle, or end of the source text.

Distance between the poem's first and last words within the source text. This metric can be used to differentiate poems that draw exclusively from one narrow region within the source text from poems that draw from a larger span. It is especially useful when applied alongside the previous metric to identify where in the source text the user focused their attention when selecting words to retain.

Poem length in characters. This metric counts the total number of characters in the selected words that comprise the poem. Many erasure poems attempt to visually overwhelm the reader with the sheer amount of text that is erased [33]; counting non-erased characters relative to a fixed source text length works as a loose proxy for the proportion of the source text that is erased.

Average English-corpus word frequency of the words selected for inclusion in the poem. This metric attempts to quantify how unusual a given poem's word choices are *in the context of the English language as a whole*, under the logic that retained words in erasure poems are often chosen with the intent to surprise the reader. For English word frequency data, we used the SUBTLEX$_{\text{US}}$ dataset of film and television subtitles [36]—specifically the word frequency per 1,000,000 words measure (SUBTL$_{\text{WF}}$), as given by the file that contains word frequency data for all 74,286 distinct words that appear within the dataset.

Average within-poemspace word frequency of the words selected for inclusion in the poem. This metric attempts to quantify how unusual a given poem's word choices are *in the context of the complete poemspace*, with each word's frequency determined by counting how often it appears in the complete set of poems that the generative model is able to create from this source text. Because the meaning of an erasure poem is partly defined in relation to the meaning of its source text [31], including the alternative erasures of the same source text that might have been performed, it makes sense to consider the individual poem's relationship to the full poemspace as a potential aesthetic measure.

Average word pair probability within the poemspace across all word pairs in the poem. The probability of a word pair $\langle a, b\rangle$ is the probability that, given word a is present in a poem, word b is also present within that same poem. Like the word frequency metrics, this metric attempts to capture the surprising quality of word choices in many human-created erasure poems; here, it is particularly useful for identifying poems that contain pairs of words that the generative model would not often use together when unguided by a human user.

Letter repetition score. This metric counts all of the *unique* letters in a poem and divides this count by the *total* number of letters in the poem. Poems receive a low

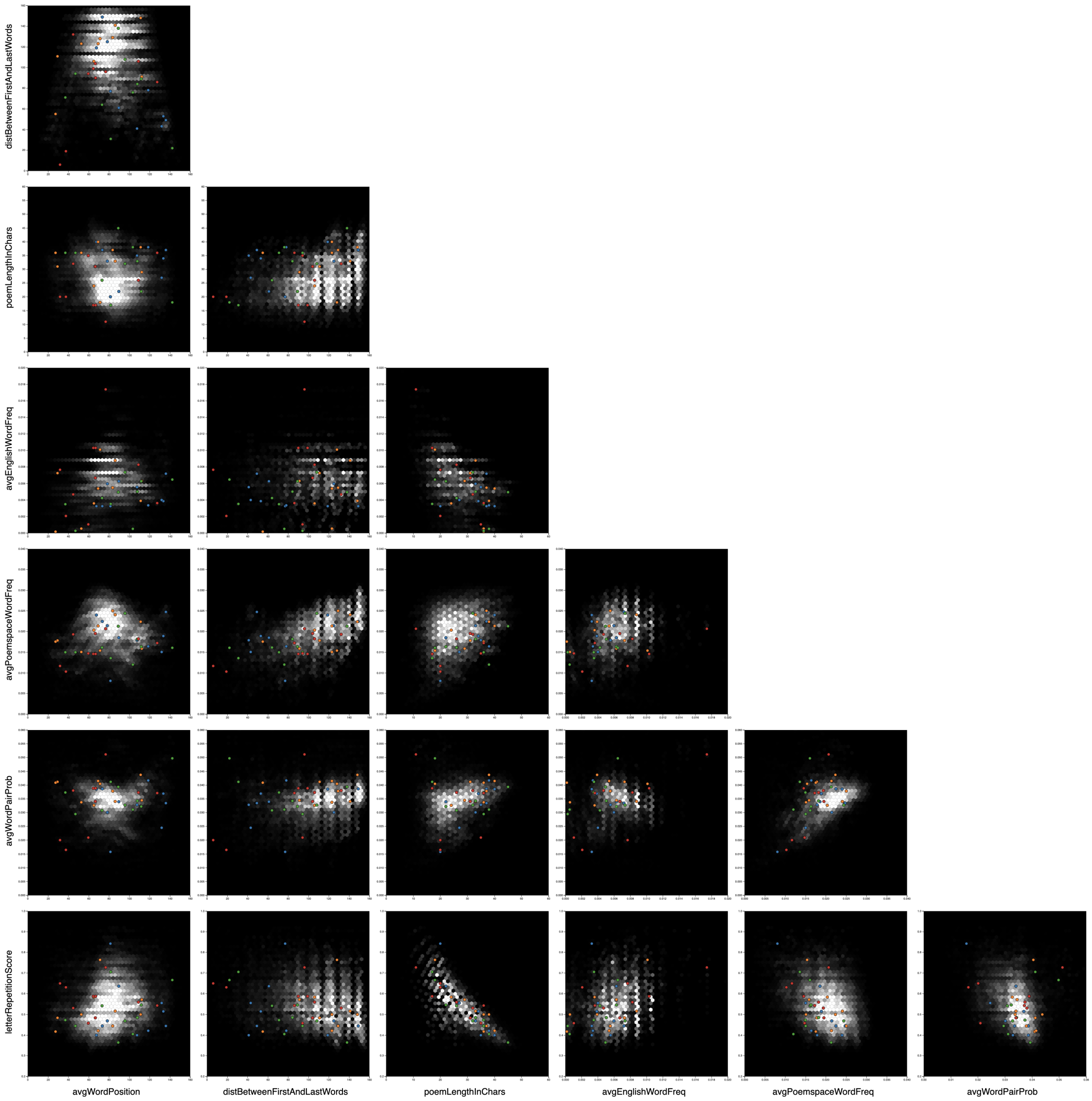

Figure 2: A corner plot [35, 13] summarizing all of the metric pairs we used to evaluate erasure poems. A black-and-white histogram showing the shape and density of the generative model's expressive range forms the background of each individual plot, while co-created artifacts are drawn as a scatterplot layer on top of this background. Each participant's artifacts are indicated in this and subsequent scatterplots by a participant-specific color: P1, P2, P3, P4.

score if they reuse the same letter many times, and a high score if they reuse letters infrequently. This score is intended as a loose proxy for *sound reuse*, an aesthetic quality of poems related to how similar the words in the poem sound to one another when pronounced. Sound devices [34] such as assonance, consonance, alliteration, and rhyme are all varieties of sound reuse. Low letter repetition scores may indicate intentional selection of words that sound similar to one another, while very high letter repetition scores may indicate intentional selection of words that phonetically clash.

We also defined minimum and maximum variants of each metric that reports an average value—for instance, metrics that report the probability score of the most and least likely word pairs in each poem, to accompany the metric that reports the average probability of all of a poem's word pairs. However, for reasons of space, we do not report results related to these metrics here.

4.4. Data and Code Availability

All data for this study (including the participant-created poems and the full poemspace), as well as the code that we used to run the analysis and generate our visualizations, is available online: https://github.com/mkremins/redactionist-eraca.

5. Results and Discussion

Examination of the visualizations we created allows us to characterize Redactionist's effects on users in terms of the artifacts they tend to create. Below, we briefly discuss some of the key findings from our pilot study.

5.1. Users collectively explore most of the model's expressive range

At a high level, inspection of the metric pair visualizations in the corner plot (Figure 2) shows that the four participants collectively created artifacts that cover the generative model's expressive range well. Although the densest clusters of co-created artifacts within the possibility space mostly do not align with the densest clusters of possible machine-generated artifacts, the placement of co-created artifacts across the possibility space suggests that users are capable of creating poems that occupy any point within the generative model's expressive range as defined by these metrics. This provides evidence that the Redactionist interface is successful at exposing the full possibility space of the underlying generative model to its users: no regions of the possibility space are inaccessible to users due to interface limitations.

A particularly good example of expressive range coverage can be seen in the visualization of the `poemLengthInChars` and `avgEnglishWordFreq` metric pair (Figure 3). Although co-created poems largely fall outside of the densest parts of the possibility space, and although some co-created poems stand out as extreme outliers relative to the possibility space as a whole, the overall distribution of co-created artifacts shows that users can access the entirety of the possibility space.

5.2. Co-created artifacts are disproportionately unusual

Further inspection of the corner plot (Figure 2) shows that co-created artifacts rarely occupy the densest parts of the generative model's expressive range, and that they are unusually likely to be outliers in comparison to most possible model-created poems. This is backed up by closer examination of individual metric pairs: for instance, Figure 4 shows that co-created artifacts are much more likely than model-created artifacts to contain unusual individual words and word pairs (from the model's perspective).

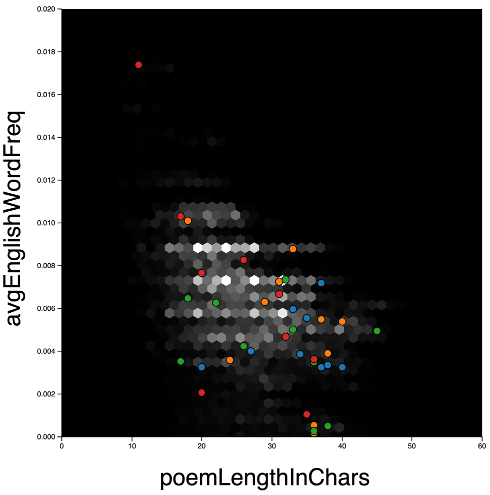

Figure 3: A single expressive range coverage visualization highlighting the artifacts created by four participants on top of a histogram showing the complete expressive range for this metric pair. Participants covered the expressive range well, but the density of co-created artifacts does not follow the density of the underlying possibility space.

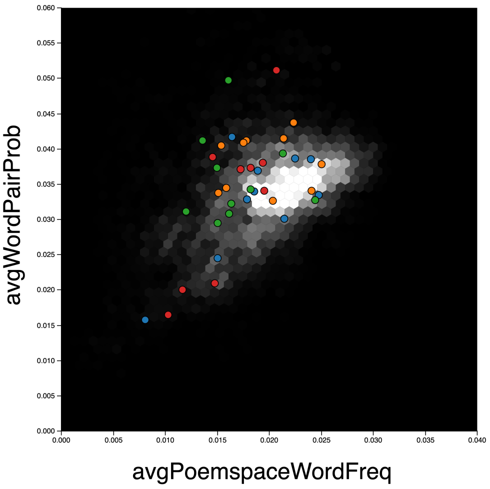

Figure 4: A visualization showing that, for one metric pair, co-created artifacts do not seem to follow the "center of mass" for the generative model's expressive range as a whole: in fact, they seem to avoid it.

This may suggest that the generative model's expressive range contains many poems that human users would tend to reject as unsuitable, leading to a focusing of human attention on poems that are considered to be outliers.

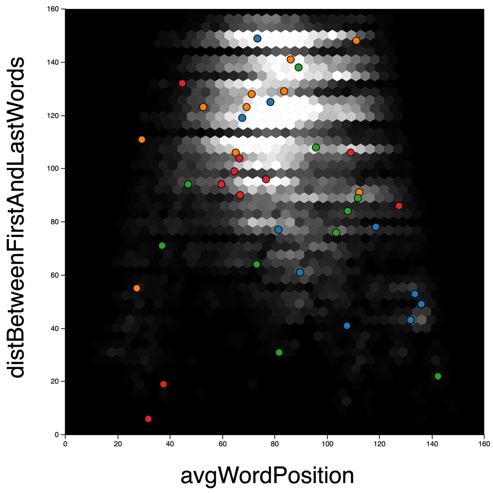

Figure 5: A visualization illustrating differences in user behavior. One participant (P4) primarily selected words from near the start of the source text; one (P1) primarily selected words from near the end; one (P2) primarily created poems that spanned a very large proportion of the source text; and one (P3) created poems using words from throughout the source text, but maintaining a relatively small "window of attention" within the source text for each poem.

5.3. Different users explore different portions of the expressive range

We can also see from the corner plot (Figure 2) that different users tend to explore different portions of the expressive range. Each participant's co-created artifacts tend to cluster together, allowing for the visual determination of each participant's "style" in terms of the metrics we defined. Figure 5 shows this especially well: the visual clustering of poems created by each participant is highly evident here, suggesting that each participant tended to behave differently when deciding where in the text they should select words from. Participants P1, P3, and P4 all tended to pick a relatively narrow "window" within the source text and construct poems from several close-together words, but P4 tended to draw from near the start of the source text; P1 tended to draw from near the end; and P3 moved throughout the source text while still selecting mostly close-together words for each individual poem. Meanwhile, participant P2 tended to create poems that drew words from all throughout the source text, resulting in unusually high `distBetweenFirstAndLastWords` scores relative to the other participants.

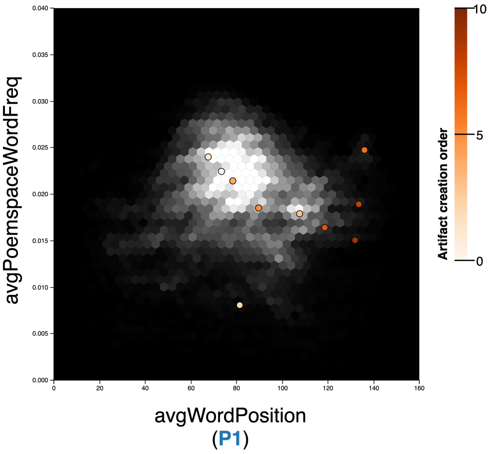

Figure 6: A trajectory visualization illustrating one participant (P1)'s gradual convergence on selecting words from a particular part of the source text (toward the end) and on selecting words that rarely appear in model-produced poems.

5.4. Redactionist tends to promote convergent thinking over divergent

One question that it would be useful to answer about MICIs involves the tendency of the MICI's design to promote divergent or convergent thinking within a single user: do users tend to jump around between very different regions of the possibility space, or do users tend to select a single region of the possibility space and then "mine it out" by creating several artifacts all drawn from that same region? This question can be answered to some extent with a standard scatterplot overlay, but coloring the points representing a single user's artifacts in the order that these artifacts were created (according to a color gradient) can further enable us to discern whether artifacts drawn from a particular region of the possibility space were created contiguously or noncontiguously. We call these augmented scatterplots "trajectory visualizations", because they attempt to illuminate a single user's trajectory through the possibility space over time; an example trajectory visualization can be seen in Figure 6.

Side-by-side per-user trajectory visualizations for the `avgWordPosition` and `distBetweenFirstAndLastWords` metrics (Figure 7) shows that Redactionist users tend to converge on a specific approach to selecting words from the source text for inclusion in poems, essentially choosing a "home region" within the source text that they repeatedly revisit for multiple poems over the course of a single session. Specifically, by examining the order in which poems were created alongside their positioning within the expressive range, we can see that all four participants created at least three poems that fall within a visually distinct region of the expressive

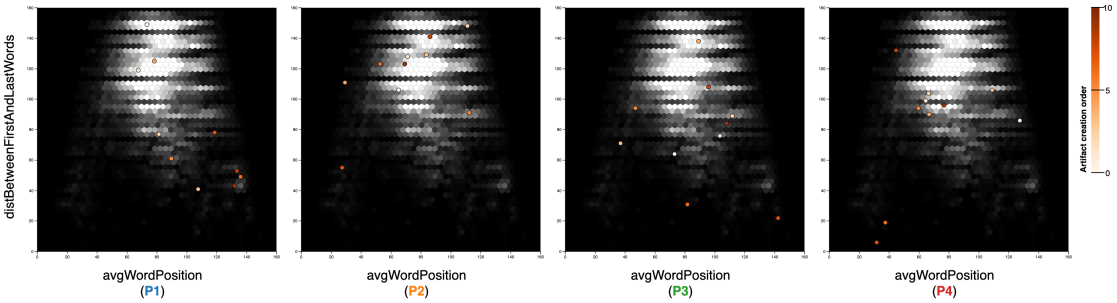

Figure 7: A set of four trajectory visualizations (one for each participant) illustrating the tendency of participants to converge on a single "home" region within the source text and repeatedly return to this region for multiple poems within the course of a single session. For instance, P2 repeatedly revisits a cluster in the upper center of the distribution throughout their session, alternating between exploration of this region and farther-flung alternatives.

range from a source text location perspective; that two of these participants (P2 and P4) created an even larger number of poems sampled largely from similar locations within the source text; and that these poems were not created in immediate sequence with one another, indicating that the user's preference for a particular "home location" endures over the course of a session rather than disappearing after a few successive poems are sampled from the same region.

The tendency of Redactionist users to work convergently may be partly attributable to interface design. In Redactionist, once you have locked in a large number of words to finish a poem, it is easier to change only a few of these selections than to change a large number of them at once. Additionally, the actual word attached to a span of selectable text is not made visible to users until they hover over this span. Consequently, users often take small, incremental steps within the possibility space and less frequently make the large jumps needed to switch from one region of the space to another—and even when they do make larger jumps, they tend to anchor their jumps on potentially selectable words that they had used in poems previously. Insofar as these behaviors are attributable to the user's inadvertent *fixation* on a narrow region of the expressive range rather than intentional *commitment* to certain design choices [37], this analysis suggests the possibility of user interface features that deliberately encourage users to work divergently: for instance, an option to randomly select a new set of words containing none of the words that are currently selected, or a process that randomly highlights a nearby selectable word that a user has not yet used in any poems.

5.5. Users experiment with highly unusual word choices before regressing to the mean

We hypothesized that, as users are exposed to more of the generative model's choices and explore a wider variety of the words available to them, they might be driven toward selecting more unusual words over time—both from the perspective of the Redactionist poemspace (i.e., avoiding words that tend to be used very frequently in generated poems) and from the perspective of the English language as a whole (i.e., preferring words that occur less frequently in a corpus of general English usage). Examination of trajectory visualizations for the `avgPoemspaceWordFreq` and `avgEnglishWordFreq` metric pair, however, does not show this expected trend—see Figure 8. Instead, we observe that all four participants at some point during their session *experimented* with the selection of highly unlikely words, but that no participant *remained* consistently focused on the selection of highly unlikely words afterward.

In particular, in the bottom left-hand corner of their respective trajectory visualizations, we can see that three of four participants (P2, P3 and P4) all discovered a region of poemspace in which the poems contain words that are highly unlikely from both a poemspace word frequency and English word frequency perspective. Each of these participants created two poems within this region of poemspace; for P2 and P4 one of these poems was created shortly after the other, while for P3 these poems were separated in time by several others. However, none of these participants' penultimate or final poems fall within this region, suggesting that none of these participants were primarily attempting to optimize for surprising word choice over the course of their session.

This may be an instance of the curiosity-driven behavior previously observed in some MICI users [25]: deliberate probing of the MICI in an effort to discover the edges

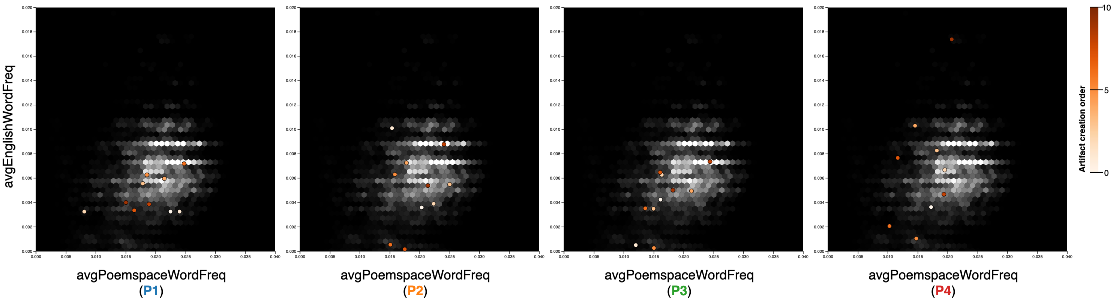

Figure 8: A set of four trajectory visualizations (one for each participant) illustrating the tendency of participants to create a few poems containing highly unusual words at some point during their session, before eventually regressing to the mean.

of the possibility space. This explanation may also help to explain why P4's final poem in particular is visibly an extreme outlier on the `avgPoemspaceWordFreq` metric, containing much more common English words on average than any other co-created poem: all of the participants were driven by curiosity to some extent, but P4 was especially successful in probing the extreme corners of the possibility space.

6. Limitations

6.1. Pilot Study Limitations

Our four participants for the ERaCA pilot study presented here were all members of this paper's authorship team. We took this unusual approach because obtaining IRB approval for collection of user data at scale was not possible prior to the workshop submission deadline, due partly to the late-breaking nature of this work and partly to ongoing pandemic-related IRB reviewing backlogs. The small number of participants limits generalizability of the study's results, and there was obviously an incentive for authors to try to "behave interestingly" while using the MICI so that publishable results would emerge. We tried to mitigate this potential source of bias (in particular by avoiding selection of poem evaluation metrics until after the data collection was complete), but this attempt at establishing a firewall between data collection and analysis is clearly imperfect. In the near future, we plan to run a larger user study (with a larger number of non-coauthor participants) to validate and expand on our findings. In the meantime, however, because the primary goal of this paper is to introduce the idea of expressive range coverage analysis and present a minimal case study of its application, we believe that our pilot study results are sufficient to illustrate the methodology.

6.2. Visualization Limitations

The visualizations that we presented here use only color to indicate which user created each artifact (in multi-user visualizations) and the order in which artifacts were created (in single-user trajectory visualizations). This limits the accessibility of these visualizations to users who have difficulty perceiving color [38]. Future work should explore the use of shape, pattern, or another redundant visual channel alongside color in the co-created artifacts visualization layer. Particularly for trajectory visualizations, we suspect there may be value in shaping each data point as a small arrowhead pointing in the direction of the next data point in sequence, so that the order in which a user created their artifacts can be visually analyzed more easily.

6.3. Limitations of ERaCA as a Method

Like ERA, ERaCA is a qualitative and visual evaluation technique. It is not capable of producing a single summary value that tells you how good a MICI is—but it does illuminate the MICI's influence on users and co-created artifacts in useful ways, especially when the information that ERaCA provides is considered in terms of the MICI's overall goals. It may be the case that ERaCA is best employed alongside other user-centered evaluation methods, such as the think-aloud method [39] and interviews [20], to provide an additional channel of information. For instance, there may be potential value in showing ERaCA plots to study participants in a debriefing interview after a conventional user study session, using the plots as prompts or visual aids to elicit remarks or insights from participants about specific aspects of their experience.

Also like ERA, ERaCA relies on domain-specific artifact evaluation metrics to characterize artifacts in a particular creative domain. A few standard metrics [40] are widely used to evaluate 2D platformer game levels, and metrics for several other domains [18, 41, 17] have also been defined. However, there are many domains for which appropriate metrics have not yet been developed, necessitating additional work before ERaCA can be applied to these domains.

Finally, ERaCA can only be applied to MICIs where the underlying generative model is capable of producing complete artifacts without human input. Fortunately, many recently developed MICIs for a wide variety of creative domains—including sketching [42], creature design [43], prose-level creative writing [44, 45], plot-level storytelling [46], poetry [47], instrumental music [48], songwriting [49], game design [50, 51], and level design [52]—follow this architectural pattern. However, ERaCA may not be as readily applicable to the evaluation of MICIs for domains such as physical crafts, in which the generative models employed by MICIs often cannot produce complete artifacts on their own due to the need for human involvement in the physicalization of generated designs [53, 54].

7. Conclusion

Expressive range coverage analysis (ERaCA) is a potentially powerful new methodology for the evaluation of mixed-initiative creative interfaces (MICIs). However, it still needs to be evaluated at a greater scale; visually polished to improve visualization legibility; integrated with other approaches to MICI evaluation, including conventional user studies; and extended to many new creative domains. We are excited to undertake many of these efforts in the future and intend to adopt ERaCA in the evaluation of our own co-creative systems going forward.

Acknowledgements

This paper was partly inspired by Gillian Smith's questions about evaluation during Max Kreminski's advancement to candidacy. We hope that ERaCA represents a step toward a method of evaluating co-creative systems that better reflects what we value about co-creativity.

References

[1] S. Deterding, J. Hook, R. Fiebrink, M. Gillies, J. Gow, M. Akten, G. Smith, A. Liapis, K. Compton, Mixed-initiative creative interfaces, in: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, 2017, pp. 628–635.

[2] A. Liapis, G. N. Yannakakis, C. Alexopoulos, P. Lopes, Can computers foster human users' creativity? Theory and praxis of mixed-initiative co-creativity, Digital Culture & Education (DCE) 8 (2016) 136–152.

[3] B. Shneiderman, Creativity support tools: Accelerating discovery and innovation, Communications of the ACM 50 (2007) 20–32.

[4] A. Summerville, S. Snodgrass, M. Guzdial, C. Holmgård, A. K. Hoover, A. Isaksen, A. Nealen, J. Togelius, Procedural content generation via machine learning (PCGML), IEEE Transactions on Games 10 (2018) 257–270.

[5] A. M. Smith, M. Mateas, Answer set programming for procedural content generation: A design space approach, IEEE Transactions on Computational Intelligence and AI in Games 3 (2011) 187–200.

[6] P. Karimi, K. Grace, M. L. Maher, N. Davis, Evaluating creativity in computational co-creative systems, in: Proceedings of the 9th International Conference on Computational Creativity, 2018, pp. 104–111.

[7] E. A. Carroll, C. Latulipe, R. Fung, M. Terry, Creativity factor evaluation: towards a standardized survey metric for creativity support, in: Proceedings of the Seventh ACM Conference on Creativity and Cognition, 2009, pp. 127–136.

[8] C. Lamb, D. G. Brown, C. L. Clarke, Evaluating computational creativity: An interdisciplinary tutorial, ACM Computing Surveys (CSUR) 51 (2018).

[9] A. Jordanous, Evaluating evaluation: Assessing progress and practices in computational creativity research, in: Computational Creativity, Springer, 2019, pp. 211–236.

[10] S. Colton, G. A. Wiggins, Computational creativity: The final frontier?, in: ECAI 2012 - 20th European Conference on Artificial Intelligence, IOS Press, 2012, pp. 21–26.

[11] G. Smith, J. Whitehead, Analyzing the expressive range of a level generator, in: Proceedings of the 2010 Workshop on Procedural Content Generation in Games, 2010.

[12] M. Kreminski, M. Mateas, Toward narrative instruments, in: International Conference on Interactive Digital Storytelling, Springer, 2021, pp. 499–508.

[13] A. Summerville, Expanding expressive range: Evaluation methodologies for procedural content generation, in: Fourteenth Artificial Intelligence and Interactive Digital Entertainment Conference, 2018.

[14] G. Smith, J. Whitehead, M. Mateas, Tanagra: Reactive planning and constraint solving for mixed-initiative level design, IEEE Transactions on Computational Intelligence and AI in Games 3 (2011) 201–215.

[15] M. Cook, J. Gow, G. Smith, S. Colton, Danesh: Interactive tools for understanding procedural content generators, IEEE Transactions on Games (2021).

[16] S. Snodgrass, A. Summerville, S. Ontañón, Studying the effects of training data on machine learning-based procedural content generation, in: Thirteenth Artificial Intelligence and Interactive Digital

Entertainment Conference, 2017.

[17] Q. Kybartas, C. Verbrugge, J. Lessard, Tension space analysis for emergent narrative, IEEE Transactions on Games 13 (2020) 146–159.

[18] E. Teng, R. Bidarra, A semantic approach to patch-based procedural generation of urban road networks, in: Proceedings of the 12th International Conference on the Foundations of Digital Games, 2017.

[19] D. Buschek, L. Mecke, F. Lehmann, H. Dang, Nine potential pitfalls when designing human-AI co-creative systems, in: Joint Proceedings of the ACM IUI 2021 Workshops, 2021.

[20] A. Adams, P. Lunt, P. Cairns, A qualititative approach to HCI research, in: Research Methods for Human-Computer Interaction, Cambridge University Press, 2008, pp. 138–157.

[21] A. Kantosalo, Human-Computer Co-Creativity: Designing, Evaluating and Modelling Computational Collaborators for Poetry Writing, Ph.D. thesis, University of Helsinki, 2019.

[22] J. Kim, M. L. Maher, S. Siddiqui, Studying the impact of AI-based inspiration on human ideation in a co-creative design system, in: Joint Proceedings of the ACM IUI 2021 Workshops, 2021.

[23] M. Kreminski, B. Samuel, E. Melcer, N. Wardrip-Fruin, Evaluating AI-based games through retellings, in: Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, volume 15, 2019, pp. 45–51.

[24] M. P. Eladhari, Re-tellings: the fourth layer of narrative as an instrument for critique, in: International Conference on Interactive Digital Storytelling, Springer, 2018, pp. 65–78.

[25] M. J. Nelson, S. E. Gaudl, S. Colton, S. Deterding, Curious users of casual creators, in: Proceedings of the 13th International Conference on the Foundations of Digital Games, 2018.

[26] M. Kreminski, M. Mateas, Reflective creators, in: International Conference on Computational Creativity, 2021.

[27] M. Kreminski, I. Karth, N. Wardrip-Fruin, Generators that read, in: Proceedings of the 14th International Conference on the Foundations of Digital Games, 2019.

[28] K. Compton, M. Mateas, Casual creators, in: International Conference on Computational Creativity, 2015, pp. 228–235.

[29] L. Daly, The days left forebodings and water, https://lizadaly.com/pages/blackout, 2016.

[30] A. Parrish, Exploring (semantic) space with (literal) robots, http://opentranscripts.org/transcript/semantic-space-literal-robots, 2015.

[31] T. Macdonald, A brief history of erasure poetics, Jacket Magazine 38 (2009).

[32] B. McHale, Poetry under erasure, in: Theory into Poetry: New Approaches to the Lyric, Rodopi Amsterdam, 2005, pp. 277–301.

[33] B. C. Cooney, "Nothing is left out": Kenneth Goldsmith's *Sports* and erasure poetry, jml: Journal of Modern Literature 37 (2014) 16–33.

[34] J. Kao, D. Jurafsky, A computational analysis of style, affect, and imagery in contemporary poetry, in: Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature, 2012, pp. 8–17.

[35] D. Foreman-Mackey, corner.py: Scatterplot matrices in Python, The Journal of Open Source Software 1 (2016) 24.

[36] M. Brysbaert, B. New, Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english, Behavior Research Methods 41 (2009) 977–990.

[37] J. S. Gero, Fixation and commitment while designing and its measurement, The Journal of Creative Behavior 45 (2011) 108–115.

[38] W3C Web Content Accessibility Guidelines Working Group, Use of color: Understanding sc 1.4.1, https://www.w3.org/TR/UNDERSTANDING-WCAG20/visual-audio-contrast-without-color.html, 2016.

[39] K. A. Ericsson, H. A. Simon, Protocol Analysis: Verbal Reports as Data, MIT Press, 1984.

[40] A. Canossa, G. Smith, Towards a procedural evaluation technique: Metrics for level design, in: The 10th International Conference on the Foundations of Digital Games, 2015.

[41] A. Liapis, G. N. Yannakakis, J. Togelius, Sentient Sketchbook: computer-assisted game level authoring, in: Proceedings of the 8th International Conference on the Foundations of Digital Games, 2013.

[42] J. E. Fan, M. Dinculescu, D. Ha, collabdraw: an environment for collaborative sketching with an artificial agent, in: Proceedings of the 2019 Conference on Creativity and Cognition, 2019, pp. 556–561.

[43] Z. Epstein, O. Boulais, S. Gordon, M. Groh, Interpolating GANs to scaffold autotelic creativity, in: Joint Workshops of the International Conference on Computational Creativity, 2020.

[44] A. Calderwood, V. Qiu, K. I. Gero, L. B. Chilton, How novelists use generative language models: An exploratory user study, in: Joint Proceedings of the Workshops on Human-AI Co-Creation with Generative Models and User-Aware Conversational Agents, 2020.

[45] M. Roemmele, A. S. Gordon, Automated assistance for creative writing with an RNN language model, in: Proceedings of the 23rd International

Conference on Intelligent User Interfaces Companion, 2018.

[46] M. Kreminski, M. Dickinson, M. Mateas, N. Wardrip-Fruin, Why Are We Like This?: The AI architecture of a co-creative storytelling game, in: International Conference on the Foundations of Digital Games, 2020.

[47] H. G. Oliveira, T. Mendes, A. Boavida, A. Nakamura, M. Ackerman, Co-PoeTryMe: interactive poetry generation, Cognitive Systems Research 54 (2019) 199–216.

[48] R. Louie, A. Coenen, C. Z. Huang, M. Terry, C. J. Cai, Novice-AI music co-creation via AI-steering tools for deep generative models, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020.

[49] M. Ackerman, D. Loker, Algorithmic songwriting with ALYSIA, in: International Conference on Evolutionary and Biologically Inspired Music and Art, 2017.

[50] M. Nelson, S. Colton, E. Powley, S. Gaudl, P. Ivey, R. Saunders, B. Perez Ferrer, M. Cook, Mixed-initiative approaches to on-device mobile game design, in: Proceedings of the CHI'17 Workshop on Mixed-Initiative Creative Interfaces, 2017.

[51] M. Kreminski, M. Dickinson, J. Osborn, A. Summerville, M. Mateas, N. Wardrip-Fruin, Germinate: A mixed-initiative casual creator for rhetorical games, in: Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, volume 16, 2020, pp. 102–108.

[52] M. Guzdial, N. Liao, J. Chen, S.-Y. Chen, S. Shah, V. Shah, J. Reno, G. Smith, M. O. Riedl, Friend, collaborator, student, manager: How design of an AI-driven game level editor affects creators, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019.

[53] L. Albaugh, S. E. Hudson, L. Yao, L. Devendorf, Investigating underdetermination through interactive computational handweaving, in: Conference on Designing Interactive Systems, 2020, pp. 1033–1046.

[54] A. Sullivan, Embroidered Ephemera: Crafting qualitative data physicalization designs from twitter data, in: Joint Workshops of the International Conference on Computational Creativity, 2020.