Genres may have trends, but individual works may challenge such boundaries, or contain many styles at once. By taking a tacit approach, we create a new computational lens with which to look at the experience of style beyond these categorizations.

To access tacit knowledge of style, we design a crowdsourcing task to elicit human judgments of style similarity between passages of text, and create the first dataset of human comparisons of style in fiction, with ~66,000 judgments across ~21,000 comparisons. We train a machine learning model on this dataset, and operationalize this model through two interface probes. First, an "Explorer" interface plots excerpts by their style in a 2-D style space, allowing the exploration of style between texts. Second, an "Editor" interface presents a co-located visualization of style next to editable text. The Editor foregrounds style as it ebbs and flows through the text and allows instant update of the visual representation as the text is altered. These interfaces can process new texts not in the original dataset, enabling users to explore style across any written collection, even beyond our dataset. Through a user study, we highlight how these interface probes expose style, inviting new curiosity, reflection, and creativity in reading and writing.

Style is and likely always will be a literary concept resisting exact classification. This paper should not be interpreted as attempting to mechanize literature or style into an exact computational model. To the contrary, the joy we experience from engaging with literature inspires this work. We believe that even partially and selectively exposing style within texts can be inspiring, inviting genuine curiosity and the discovery of new, similar, or different styles. Our hope is that such computational literary tools can augment and accelerate the discovery, appreciation, joy, and celebration of literature.

RELATED WORK

This paper draws from techniques across several disciplines. We discuss how these disciplines have engaged with style and how we integrate elements from these fields to offer a new perspective on approaching style with computational tools.

Literary Theory and Defining Style

Literary theorists have examined the question of style for centuries; no universally accepted conclusions have been reached. However, Herrmann et al. present a potential definition for computational style research: "Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively" [11]. This definition encompasses both "complete texts or fragments of texts," and "is not limited to a given author's style." Herrmann et al. draw from the history of literary theories of style to propose a definition that creates common ground for literary scholars, digital humanists, and computational researchers. Here, we will base our exploration of style on this definition to bridge the areas of literary theory and human computer interaction. Specifically, we focus on style in "fragments of texts," at the unit of 200 word excerpts. We approach entire works as combinations of excerpts. Therfore we do not expect texts to have a uniform style throughout; rather the overall style of a text includes the variations in style within it. However, we are also influenced by other traditions of style, as we are interested in understanding people's subjective aesthetic experience, rather than adhering to a single formal definition.

Quantitative Features of Style

Since the first use of statistical methods for authorship attribution in 1960 [24], the field of stylometry has used analytical and computational methods to build quantitative models of style to classify unknown texts. Stylometry is commonly used for authorship identification [40], plagiarism detection [22], author gender identification [3], and genre classification [34], among others [19, 39]. Researchers have identified features that perform well for these uses [15], including lexical features (e.g. word frequencies), character features (e.g. n-grams), and syntactic features (e.g. parts of speech) [33], and used computational methodologies such as support vector machines, neural nets, self-organizing maps, and spanning trees [10, 26, 32]. Some quantitative style metrics have come into common use; for example, the Flesch-Kincaid readability tests compute how difficult a passage is to understand based on sentence length and the number of syllables per word [17]. Online resources such as Hemingway App [1] use highlighting to suggest ways to improve scores on these metrics. We ground our work in these methods, but instead of investigating authorship attribution or categorizations through explicit features, we create our model using a human-defined definition of style informed by tacit knowledge. A similar tacit aspect in knowledge about genres is noted by [13].

Most datasets for stylometric work are created through categorization, using metadata such as authorship or genre to define classes. Looking beyond this approach, Crosbie et al. [7] investigate the quality of "literariness" using stylometric techniques, and generate a small dataset of 10 passages rated by the general public on a Likert scale for literariness. We similarly build a dataset from human judgments, but focus on a distinctly different approach to style: comparative similarity based on tacit knowledge. We create a dataset of 800 passages combined into 21,000 comparisons. To our knowledge, no other such direct dataset of style similarity judgments exists.

Natural Language Processing

Natural language processing techniques provide approaches to semantic understanding and automatic text generation. Word embeddings [27] have been used to understand similarities between words, and improve semantic analysis of text. A conceptual extension of word embeddings, document embeddings have been used for sentiment analysis and text classification [18, 8], and can effectively cluster texts based on similarity. A technique for auto-generating text [31] can produce, to some extent, stylistically coherent content. While these techniques may be able to capture some aspects of style, they do not separate style from semantics. We demonstrate an architecture that generates a separable representation of style.

Human Computer Interaction and Writing Support Tools

We highlight here the subset of work in writing support tools most related to our approach to style. Bernstein et al. [5] integrated crowd-powered editing tools into a text editor to

handle tasks relying on human judgment; similarly, we leverage crowd knowledge to approach a problem that requires human judgment: tacit knowledge of style. Pera et al. [28] used readability and style characteristics derived from reviews, in addition to content, to recommend books for children. Vaz et al. [36] explored the integration of style analysis into recommendation systems, showing that stylometric features improve results. Vaz et al.'s prototype system [37] used stylometric comparisons to recommend similar books. We also use style to inform computational tools, but rather than utilizing the formal metrics derived from authorship identification directly, we surface the gestalt experience of style and support interactive interpretation through visualizations, rather than generating specific recommendations.

Digital Humanities and Visualization

In the literary technique of "close reading," annotating in situ and preserving the structure of the text are essential to analysis. In contrast, "distant reading" is a data-driven approach to studying texts [12, 23], in which the structure of the text is removed to provide a global view of the text or its relation to a larger corpus. Here we discuss research which combines close and distant reading to take a computational approach to text while preserving structure or detail through visualizations. Muralidharan et al. [25] created a tool for investigating patterns in text collections through visualization. Weber [38] used a word-highlighting approach where each part of speech is assigned a color to reveal contrasting visual patterns in fiction and scientific writing. Keim et al. [14] visualized texts by computing a sequence of values for individual stylometric features, creating "fingerprints" that can be compared across works. McCurdy et al. [20] visualized the sound of a poem in the context of the text. These each use explicit characteristics, directly represented. Our visualizations similarly leverage considerations of both close and distant reading, but are driven by our tacit model of style, not by explicit features.

FORMATIVE STUDY: UNDERSTANDING STYLE

Our research began with a formative study to elicit the personal concepts, perceptions, and articulations of style from people with significant knowledge of literature. We recruited 14 participants (6 men, 8 women; mean age 23, range 18-30) who self-identified as "writers" or "avid readers" from university mailing lists for creative writing, design, and computer science. We conducted semi-structured interviews around their reading and writing practice, focusing on their thoughts about style and writing support tools. Afterwards, they interacted with an early prototype of a style exploration tool. Interviews lasted an hour, and participants were compensated US$20. We analyzed the interviews with a grounded theory approach [6].

Most participants reported experience in creative writing (12 participants) and academic writing (9), as well as other types of long-form writing, with a mean of 7 years of writing experience (range 4-15). Three participants were actively studying literature, one participant wrote in a professional capacity outside of academia, and three others held volunteer editorial positions.

We found that all of our participants had a personal definition for literary style, as well as particular styles they liked and disliked. Most valued style in choosing what to read or in shaping their own writing. Only one participant considered it irrelevant to their reading and writing. As important as style was, participants did not have a clear way of talking about it:

> **F4** [I react to style], but I think it's hard to articulate what I like about it.

> **F7** I know the vibe...I don't really have a word for it.

Instead of explicit terms, interviewees relied on examples, referencing other works as touchpoints to get their meaning across:

> **F3** There are styles, but I don't know how to communicate to you, but I can tell you check out this author, see how he writes.

When asked explicitly, participants described style as a "gut feeling" **(F7)**, "an overall effect" **(F13)**, and "more of an instinct" **(F2)**.

Participants repeatedly articulated that style is learned through experience and communicated through comparisons, suggesting that style is a form of tacit knowledge. Though most computational approaches to style rely on identifying and reporting explicit quantitative features of texts, our participants experience style in a much more intuitive way. This insight motivates the design of our dataset collection, model, and applications, to capture and enhance people's tacit approach to style.

DATASET: STYLE SIMILARITY IN FICTION

To develop tools capturing tacit approaches to style, we collected a novel dataset of style judgments.[3]. Rather than asking individuals to categorize or label style, we collect judgments of stylistic similarity, using comparisons within triplets of excerpts, as shown to be effective in [2, 35]. Excerpts are drawn from contemporary fiction, as it is accessible, commonly read, and showcases diverse styles. We would expect to find a great deal of disagreement across individuals in how they judge passages, therefore we collected seven judgments per comparison. The dataset consists of:

- *Comparisons:* Crowdworkers read a set of three excerpts of text and compare the style of the first excerpt (A) to the following two (B, C), then judge which of B or C is most stylistically similar to A.
- *Explanations:* Each crowdworker provides a few words of free text to justify their decision, by describing what is similar between A and their choice of B/C.
- *Intensities:* Each crowdworker indicates on a scale of 1-5 how similar their choice of B/C is to A.

Excerpt Generation and Comparison Triplets

Each comparison used in the crowdsourcing task consists of three excerpts from contemporary fiction displayed side by side, a "triplet." We separate the data into seven sets of triplets, with between 1,050 and 6,300 triplets, created from disjoint sets of texts. This provides disjoint sets of triplets for

[3] https://github.com/style-dataset

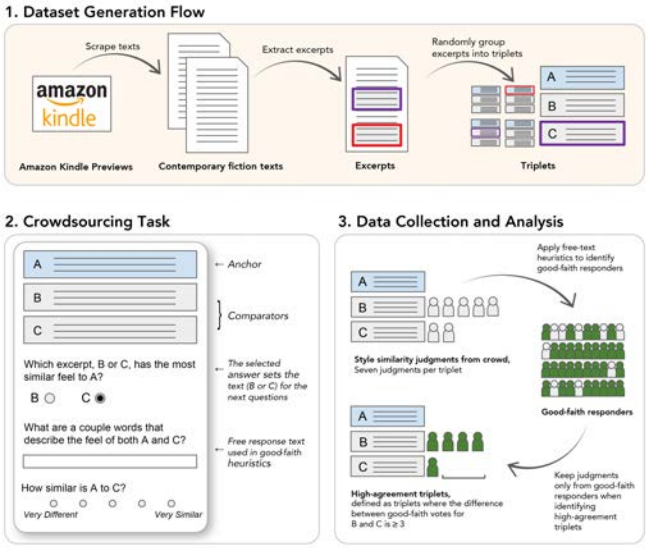

Figure 2. 1) Dataset Generation: Excerpts are extracted from texts, then combined into triplets. 2) Crowdsourcing Task: The task presents a triplet of excerpts followed by three questions about their style, or "feel." The first excerpt (A) is the anchor, to which B and C are compared. Free response text is used to identify "good-faith" respondents, i.e. those who provide reasonable free-text answers. 3) Data Collection and Analysis: Crowdsourced judgments of stylistic similarity are analyzed for reliability and agreement to identify high-agreement triplets. Here, the triplet shown is high-agreement, since 4 good-faith respondents voted for B, and only 1 voted for C, resulting in a difference of 3.

training and testing machine learning models, and varies the parameters used to select excerpts to enable different ways of looking at style (see Table 1 for a summary of parameters).

To generate the excerpts, we retrieved plain text from publicly available previews of fiction published through Amazon Kindle (Fig. 2). These books were pulled from seven genre categories as listed by Amazon: Action and Adventure, Contemporary, Historical, Horror, Humor, Literary Fiction, and World Literature. Each set includes texts from all of these genres. Amazon Kindle is used to emphasize contemporary fiction. Other sources, such as Project Gutenberg, emphasize older works in which the conventions of the era may overwhelm more subtle differences in style.

We extracted excerpts of approximately 200 words from each preview. Since the first paragraphs of a book are often quite different from the rest of the text, excerpts were extracted from the middles and ends of the previews. We rounded each to the nearest sentence end above 200 words. Choosing a style unit of 200 words allows us to analyze prose style at the paragraph level. While choosing a granular unit of comparison means we cannot look at style on the level of narrative structure, it supports investigating the local style of fragments of text (such as rhythm, sentence structure, vocabulary, etc.).

Since the number of combinations of three excerpts is prohibitively large, we generated a random subset of possible triplets for crowdsourcing. Each excerpt serves as the "anchor" in a triplet a fixed number of times; the anchor refers to excerpt A, against which B and C are compared (see Fig. 2, part 2). To avoid confounds such as shared character names, excerpts from the same text do not occur in the same triplet. Table 1 summarizes the dataset parameters; the open-source dataset provides a full characterization.

Crowdsourcing Method

To collect human judgments of style, the comparison triplets were released on a crowdsourcing platform.

Participants/platform: We recruited crowdworkers from the crowdsourcing platform *Figure Eight*[4]. This platform provides a curated workforce from around the world, with built-in quality control mechanisms, discussed below. We recruited 836 participants, from 38 countries.

Training: Participants were given an example comparison, instructions, and a brief tutorial on some concepts related to literary style. To minimize bias towards one specific interpretation of style, participants were instructed to use their intuition, rather than specific metrics. The full training instructions are provided with the dataset. After the instructions, participants completed an example task to become familiar with the task layout.

Task: After training, participants were presented with style triplets from the dataset, and answered three questions for each (Fig. 2, part 2), where the letter displayed in questions (2) and (3) depends on the answer to (1):

1. *Which text, B or C, has the most similar feel to A?*
2. *What are a couple words that describe the feel of both A and [B or C]?*
3. *How similar is A to [B or C]? (On a 5 point Likert scale from Very Different to Very Similar)*

Each triplet was presented to seven participants. Participants were paid US$0.10-0.15 per judgment.

Quality control: We used several built-in quality control mechanisms on the *Figure Eight* platform. First, participants were dropped if less time was taken than an estimate of minimum reading time for the passages. Second, participants were dropped if they failed to maintain a sufficient score on "test questions" seeded throughout the task. Test questions used the same format as the comparison triplets but consisted of two excerpts from a single text, and one from a different text, chosen to have a significantly different style.

Cleaning

To ensure that respondents took the task seriously and provided "good-faith" answers, we remove potentially "bad-faith" responses using heuristics drawing on the free-response text, such as finding nonsense words. These heuristics are provided with the dataset. 307 contributors provided good-faith judgments, with a mean of 215 good-faith judgments each (range 1 to 1715). The cleaning stage is separate from and prior to determining triplets with "high-agreement," as discussed in the next section: *Modeling Style* (Fig. 2, part 3). 45% of triplets with at least 3 good-faith judgments qualify as "high-agreement triplets."

[4] https://www.figure-eight.com/

Validation

Krippendorff's alpha is extremely low: 0.13 for all responses, and 0.15 for the cleaned responses. In crowdsourcing tasks with correct answers, low inter-rater reliability could indicate that participants lacked knowledge of the task domain or did not take care in responding. However, the interviews with experts suggest that perceptions of style are tacit, inherently subjective and vary across individuals. Due to the quality checks in place, we believe the second case holds.

We recruited three experienced writers unfamiliar with the project to perform the same task as the crowdworkers on a random sample of 30 triplets with high crowd agreement. These colleagues were recruited in-person, and completed the task remotely. All are native English speakers. The majority answer of these participants agreed with the aggregate crowd response 70% of the time. If the crowd responses were random, we would expect to see an agreement of 50%. The results show there is a perceptible style signal that aligns with overall perceptions, with individual variation.

Organization

The dataset is organized by set into comma-separated values (CSV) files. Anonymous keys link to demographic data. We provide scripts with heuristics for evaluating "good-faith" responses as discussed below, examples of how to parse the CSVs, and a full characterization of set parameters, as well as an example of how to use the data for the machine learning model described below.

This is the first dataset of tacit perceptions of style in fiction. It crosses genres and authorship boundaries, opening new directions for computational style research.

Style Similarity Dataset: Dataset Generation Parameters	
Total Texts	798
Total Excerpts	1806
Total Triplets	21,630
Excerpt Extraction Parameters	
# of excerpts per text	2-4
Words per excerpt	~200
Do excerpts include dialogue?	[None, Some, All]
Triplet Creation Parameters	
# of times an excerpt is the anchor	5-30

Table 1. We generate a set of triplets of excerpts in order to crowdsource style similarity judgments. Triplets are separated into disjoints sets to support various machine learning techniques as well as ways of looking at style. Reported totals are the sum of all sets; ranges represent parameters that vary between sets. A full characterization of all parameters and sets can be found with the open source dataset.

Style Similarity Dataset: Crowdsourcing Results	
Collected Judgments	150,720
Judgments From Good-Faith Responders	66,061
% Good-faith judgments of all judgments	44%
High-Agreement Triplets	5,162
High-Agreement triplets as a percentage of all triplets with $\geq$ 3 good-faith judgments	45%

Table 2. We crowdsource style similarity judgments for the generated triplets, and process them to select a set of good-faith, high-agreement results. Good-faith judgments refer to those left after cleaning (see subsection Dataset - Cleaning). High-agreement triplets refers to those with a preponderance of raters choosing the same answer (see subsection Modeling Style - Defining High-Agreement Triplets).

MODELING STYLE WITH THE SIMILARITY DATASET

To create computational interfaces for literary style in contemporary fiction, we need a model that reflects human experiences of style. Using the dataset presented above, we develop a model of style by training a neural net to make judgments of stylistic similarity of the form described above ("Is A more similar to B or C?"). The goal is for the model's results to align with the crowdsourced human consensus of style, instantiating the crowd's shared tacit knowledge.

Defining High-Agreement Triplets

Since style is highly subjective, no single model can reflect every individual's choices. We therefore focused on the stylistic comparisons for which there was high agreement among crowdworkers. In this way, we may develop a model that effectively captures some shared opinions about style, though it may not be effective at handling controversial cases. We define "high-agreement" triplets as those where at least three more crowdworkers chose the majority answer than the other answer (Fig. 2, Part 3). Of triplets with at least 3 good-faith judgments, 45% qualify as high-agreement (Table 2).

Training a Predictive Model

The model was trained on 916 high-agreement triplets (as 1,008 triplets had been collected at the time of the user study, and 92 were reserved for testing). We created a binary classifier trained with a binary cross entropy loss function. It takes as input an excerpt triplet, and classifies it into two categories, B or C, indicating which excerpt is most similar to A. We pre-process each excerpt in the triplet into sequences of characters, sequences of parts of speech, and sequences of word embeddings [27]. These transformations are motivated by features canonically used in stylometric work: character n-grams, syntactic features (which depend on parts of speech), and lexical features (which depend on the words themselves) [33]. The neural net then operates on the sequences independently, following [4], which explored the benefit of processing multiple input types (sequences of parts of speech, lexical features, and word n-grams) independently in the context of authorship analysis. An LSTM is used for parts of speech, and separate convolutional nets are used for characters and embeddings. After processing, the output vectors are recombined into a single vector of length 48 that represents each excerpt. A modified L2 norm of these vectors is used to calculate the distances between A and B, and A and C, which determines the final classification. See the supplemental dataset for additional details.

The model was tested against 92 high-agreement triplets. These triplets are completely disjoint from the training data, with no overlap between source texts. We achieve 67% test accuracy ((True B + True C) / All Points), and an F1-score

Figure 3. From left: Style comparisons from the collected dataset are used to train a predictive model, which learns a high-dimensional vector embedding associated with how people perceive style. The embedding is projected to 2 dimensions using principal component analysis (PCA). By mapping a color space to the 2D projection, the style of text excerpts can be associated with a color, and used to create interactive experiences (Left Interface: Explorer; Right Interface: Editor).

of .67 (precision = .61, recall = .73), with a baseline of 55%. While low in comparison to the accuracy of neural nets in domains with well-defined correct answers, accuracy is similar to that achieved by the in-person validation (70%) described above. This represents a meaningful signal in a highly subjective problem domain. Additional training data might lead to further improvement, but there may be limits to potential improvement, because people themselves disagree about style. There is no universal ground truth, so it is unlikely that any model could deliver extreme accuracy.

Visualizations

In order to predict style judgments, the model learns a 48D vector embedding of the excerpts, associated with how people perceive style. While effective for machine learning, this high-dimensional space is an intractable representation for people. To make the style information comprehensible, we downproject it to a 2D plane, which is easily represented on a screen (however, 3D or other dimensionalities could be equally valid). We call the projection the "style space."

We use principal component analysis (PCA) for the downprojection. As PCA is a common method, it provides a familiar baseline for initial explorations. To interpret the resulting axes, we identify their correlations with stylometric features in the subsection *Validating the Style Space*, below. The 15 dimensions in the embedding most important to PCA are normal or nearly normal; remaining dimension are mostly sparse.

We map colors onto the style space to help users interpret and discuss the results. Colors provide a memorable and describable representation of sub-areas. Each text excerpt can be mapped to a single color based on its position (Fig. 3, right). The color mappings also support the full-text visualizations described in the next section. We use the CIELAB color space [21], a perceptually uniform 3D representation of color, and fix it to a single lightness value to reduce the parameter space to 2D. Colors come laden with many, often contradictory, cultural associations; we do not attempt to align the style space with any prior color associations. Below we demonstrate how users successfully engaged with this visual mapping in discussing style. These representations help users interpret the style space by presenting the information in familiar ways: colors and 2D scatterplots can reveal patterns at a glance that are not apparent from numerical data.

A single color can represent an individual excerpt. An entire work, however, may consist of passages with different styles, and the work's full effect may depend on their interplay. We use the colors of the style space to create a gradient for an entire text, using a ~20 word sliding window to analyze chunks of ~200 words. Each chunk is represented as a narrow bar of its associated color; transitions are smoothed with a gradient (Fig. 4B,C). We call this visualization a "style barcode."

We leverage the considerations of both close and distant reading in designing the barcode visualizations. By retaining the vertical structure of text and aligning each color with the lines of text that produced it, a "zoomed in" view of the barcode facilitates an interaction in the manner of close reading (Fig. 4B). A "zoomed out" view can display a global perspective of multiple texts at once (Fig. 4C), facilitating high-level analysis and comparisons among texts in the manner of distant reading. In both views, the sequential nature of the data is retained, enabling users to see how style shifts within a story, and giving a visual sense for local as well as overall style.

Validating the Style Space

Since there are no universally accepted definitions of style, we validate the model and projection using several heuristic analyses.

Style Projection: The book *Exercises in Style*, by French novelist and poet Raymond Queneau, retells a single brief story (about 1 paragraph long) in 99 different styles. It has been translated into many languages, including English [30]. If the style space axes identified above (Fig. 4A) effectively separate styles, Queneau's intentionally stylistically distinct retellings should spread out across the style space. Note that we do not recalculate PCA here, rather we project the retellings onto the existing space. As expected, the retellings spread across most of the style space (Fig. 5). Some interesting clusters do arise:

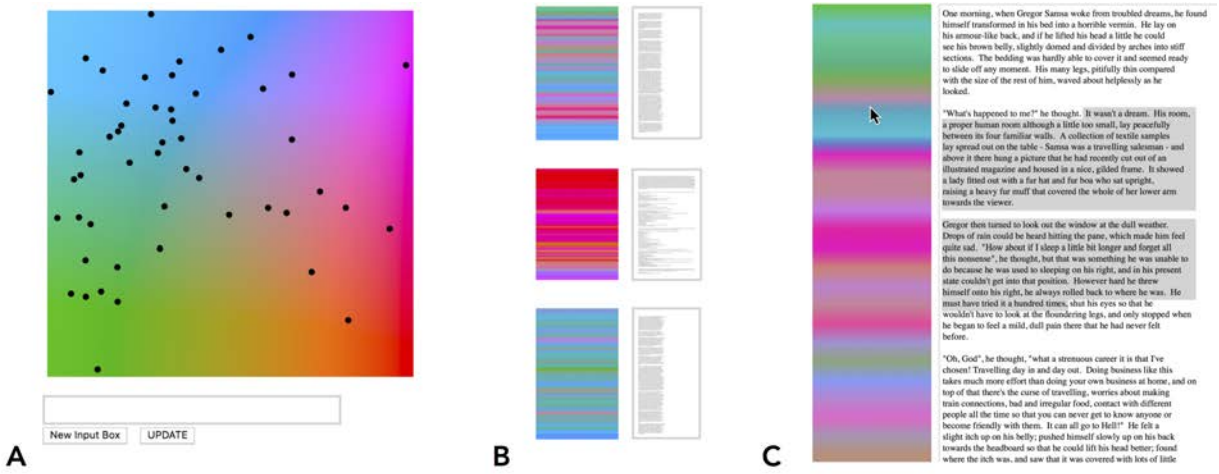

Figure 4. We implemented two style interfaces: the Explorer and the Editor. (A) Explorer: 200-word excerpts are plotted as points on a color plane. New excerpts can be added with the text box. Hovering over a point displays its text to the right of the style space. (B, C) Editor: Style is shown as a color barcode beside longer texts. (B) Three 1500-word excerpts of canonical texts are shown in the zoomed out mode for distant comparison. (Top B) Kafka's *The Metamorphosis*, (Middle B) Hemingway's *Hills Like White Elephants*, and (Bottom B) Melville's *Moby Dick*. (C) *The Metamorphosis* is zoomed in for a detailed view, demonstrating the section highlighting. Lines highlighted in grey contribute to the analysis of the blue bar directly under the cursor position. The highlighted area animates as the cursor is moved. As users edit the text, the visualization updates.

for example, the cluster along the top edge consists mostly of "nonsense" retellings, such as the *Anagrams* version in which the letters are mixed up into nonsense anagrams. These nonsense retellings take extreme values in the style space; the visualization clips their positions to fit them in the view, resulting in clustering along the top edge.

Stylometric Correlations: We would expect to see mild correlations between our style space and standard stylometric features. Perfect correlation would indicate that our approach adds little to the current understanding of style; conversely, no correlation might indicate that our model captures noise. We performed a linear regression between a selection of 23 standard stylometric features [33, 39, 40] (Table 3) and the 48D style space. The most correlated metrics are average word length ($R^2 = .25$), and the ratio of verbs to all words ($R^2 = .22$) (Table 3), indicating that the high-dimensional style space is weakly correlated with these common stylometric features. Our model captures signals associated with the stylometric approach, but cannot be fully reduced to these explicit features.

We also note correlations between the main dimension used by the second component of the PCA projection (mapped to the vertical axis of the 2D style space) and specific stylometric features associated with sentence length (e.g. average sentence length, average clause length, ratio of verbs). The vertical axis of our style space is weakly correlated with sentence length, but also represents additional nuance in the data not captured by common stylometric features.

APPLICATIONS

We instantiate our visual representations of style in two applications for searching and composing text. These applications are presented as probes into possible uses of style-aware interfaces, and explored in the user study presented below. The applications are implemented as web pages. New texts are analyzed using the model discussed in the preceding section.

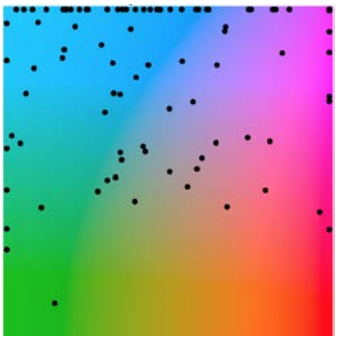

Figure 5. *Exercises in Style* by Raymond Queneau retells one scene in 99 styles. The retellings spread across much of the style space derived in Fig. 4. The cluster along the top edge arises from 'nonsense' retellings (e.g. *Anagrams*).

Metric	R^2
avg. word length	0.25
ratio verbs	0.22
ratio adverbs	0.17
ratio adjectives	0.17
ratio punctuation	0.16
avg. sentence length	0.14
function words	0.13
avg. clause length	0.11

Table 3. Linear regression of stylometric features on the 48D style space shows correlations with several common metrics. Ratios = metric / words in passage. 23 metrics were calculated; table shows those with $R^2 > 0.10$

Explorer: The explorer interface displays the style space with interactive points representing excerpts. Hovering over a point displays its associated excerpt on the screen to the right of the style space. Interactive text boxes allow the user to add new texts. Users may input their own writing or other texts. The Explorer allows users both to learn how to interpret the style space and to gain insight into the styles of works of interest to them. For the user study, we pre-loaded a subset of excerpts from the style similarity dataset (Fig. 4A).

Editor: The editor interface displays style barcodes for longer texts. Interactive text boxes allow users to input and edit text, while viewing the associated visualization. Users can view the texts at two levels of zoom: the smallest allows entire works to be seen at a glance and compared against other works (Fig. 4B), and the largest allows line-by-line style inspection and interactive editing (Fig. 4C). Hovering over a location on the

barcode highlights the lines of text that generated the selected color bar. A button updates the visualizations after text has been modified. Users can visualize the style of a work in progress, or of an existing text, in the context of both close and distant reading, and can see patterns of style change within the overall work and over time as it is edited.

The applications leverage the comparative nature of the style model. Because distance encodes similarity in style space, interpreting it requires comparing excerpts, as done via the Explorer. The Editor facilitates comparisons within or between longer works (Fig. 4A).

USER STUDY

We performed an exploratory study to investigate how users interacted with the style applications. We recruited 6 participants (1 man, 4 women, 1 not stated) from campus mailing lists for English, Literature, Computer Science, Design, and Creative Writing. All used long-form writing in their academic, personal, or professional lives. Formal training in literary analysis varied from high school or equivalent to extensive graduate training. Three are graduate students in writing-related fields. One is a professor of creative writing. All are native English speakers. Mean age was 30 (21-45), with a mean of 10 years experience in their main writing domain. Participants were compensated US$20 for a 1 hour study.

Participants began by explaining their own definition of style and discussing whether and how considerations of style featured in their reading or writing habits. They were then introduced to the Explorer interface (Fig. 4A). We demonstrated the possible interactions, and explained how the projection was created from the model, including the accuracy limitations of the model and how those errors might manifest in the projection as misplaced points. Participants were then instructed to talk aloud as they spent ten minutes investigating the excerpts projected in the style space, and to describe any patterns or contradictions they noticed. Some chose to add additional excerpts using the interactive text boxes, including academic papers, news articles, and short stories. Once they were familiar with the style space, they were introduced to the Editing interface (Fig. 4B, C). They were asked to read an excerpt of a fairy tale, and describe its style, then use the interfaces however they wished to edit that excerpt into a new style. Afterwards, they discussed their use of the tool and the style of the modified fairy tale. Finally, an open-ended interview explored their thoughts about the interfaces, whether and how they might use such tools in their everyday writing and reading, and whether their thoughts about style had changed.

RESULTS

Due to small sample size, we present qualitative results.

New Experiences of Style

Users reported that the tools prompted deeper engagement with the notion of style. Even experienced participants, who think about and teach elements of style on a daily basis, challenged their own conceptions of style:

> **P4** Seeing [excerpts] on the color map really made me try to articulate the differences. I could feel each one is different but I have not thought about [how] I could put them on different axes.

> **P5** I think that when I framed what style is in the beginning, I talked most about sentence and paragraph...Now I'm wondering...what is the smallest unit of style and what is the largest unit of style that is meaningful?

> **P6** What I first said, that style is about time, actually this is making me think that maybe style is the opposite of that, what remains constant regardless of subject matter and time span.

The authors are aware of no other writing interfaces that afford this type of critical engagement with style. Interfaces based on feature counts suppress the ambiguity and dialogue between users and texts. As Gaver et al. discuss [9], ambiguity of information can make an interface "evocative rather than didactic," and encourage self-reflection and critical engagement with the system. Our interface invites users to bring their own interpretations of style to the interaction, while encouraging them to challenge their instincts and preconceptions about style. As **P6** noted, "It's just cool to be able to play with the idea of what style is...it could be useful as much to trouble definitions of style versus fixing a definition of style. To me, making trouble is useful."

Gestalt Over Details

The idea of 'gestalt' is that of "an organized whole that is perceived as more than the sum of its parts." [5] Gestalt effects are important in writing, for instance:

> **P5** What feels right depends on how many sentences you read before and after; there is a rhythm to each paragraph that you don't get if you just go sentence by sentence.

Several participants found more value in the gestalt of the visualization than the details, appreciating the high-level view of the barcode representation:

> **P2** You can look at [the barcode] and it makes sense as both being diverse, but also unified, and different than the other text, which has its own diversity.

Viewing the barcodes as a whole gave a sense of variation and similarity that would be absent if individual passages were examined separately. The barcodes successfully surfaced both impressions of local and global style, and enabled discussion of comparisons between works. Focusing on details of exact color or position was less productive. When a detailed inspection is needed, a more explicit tool might be more appropriate, while the ambiguous visualization of style is effective for gaining an overall sense of a work and prompting engagement with style.

The Continued Ineffability of Style

Participants repeatedly confirmed that style remains best understood as a tacit, ineffable experience.

> **P6** It's just that this word is better here, I don't know why. I can make up a reason – you know the number of syllables, it's fewer syllables than the alternative and it reads faster – but a lot of times it's just that it sounds better because it sounds better.

[5] Oxford University Press, Lexico.com, 2019

Notably, while formal training influences the way people describe aspects of style decisions or analyze style, it remains tacit within their own practice. P1, a published author and professor of creative writing, said:

> **P1** The way I explain it to my students...Ideally you're absorbing [style] in class and when you're actively learning so when you do it you don't think about it. So it's like a ballet dancer who learns in class to hold in your stomach, lift your elbow, lift your chin, but then when she's on stage she's not thinking about any of those things.

These descriptions illustrate the tacit understanding of style. Explicit language and direct recommendations for changing stylistic features have value, but not during creative production, when the experience should remain tacit.

The color space supports the tacit approach through its open and flexible representation. Participants engaged with the changing colors as they changed the style of editable text, and constructed their own meanings as the colors updated:

> **P4** This section is pinker. It wasn't pink before. I imagine it's because I structured the sentences differently and took out a lot of fairy-tale style by making it super simple...I looked back at the interface, based on my estimates of different styles, what those colors meant to me, and use[d] that to figure out if this is enough of a style change.

Sometimes the ambiguity of the interface was uncomfortable,

> **P2** "It's weird to not know what do the pink and blue do, but try to talk about them."

but at the same time, it was "stimulating" **(P2)**, and inspired playful interactions with style:

> **P5** [Ernest Vincent Wright] wrote a book without using the letter 'e'...so I want to write a book that is all beige...I want to paste some of my writing and...see what kind of sunset I get.

With this kind of visual language, we could imagine interactions around learned associations: embracing a tacit approach through a shared vocabulary that does not require precise definition may encourage discussions of style; glancing at a visualization may give a sense of style instantly for text that would otherwise take hours to read and analyze.

Ambiguity Invites Personal Interpretations

As expected from the formative study, style continues to be highly personal, contextual, and fluid. There was a certain level of consistency in interpretations of the style space: for example, two participants separately arrived at the same description for the upper left section of the style space:

> **P4** [The] blue section is maybe more descriptive and not trying to mimic the way people speak.
>
> **P3** [The] ones in blue are like describing something, some situation...[the] blue ones might be about description.

A third participant described that area as "ornate, maybe introspective," **(P5)** which may correlate with "descriptive."

But no theories were universally supported; participants often encountered incongruities:

> **P2** [Rowling] has more lively style and [is] more straightforward. I am puzzled why [these two authors] are nested together.
>
> **P6** Let me look at other [points] closest to it. This would seem to go against what I just said, because this is a first person, character based story. So maybe there's not as clear of a difference as I thought.
>
> **P5** This one has longer words and denser paragraphs, so maybe that's something. Down here we have more back and forth text...well it's not universally true.

Some uncertainties may arise from flaws in the model or the projection; since it only yields 67% accuracy, and reducing style to two dimensions discards many nuances. But it may also represent the fundamental ambiguity of style: since there is no universally accepted definition, results that makes sense to one person may strike another as odd. Our tool supports this natural engagement with the fluidity of style, enabling a wide variety of interpretations. Participants spoke of looking for different metrics in the space; for example, P1 thought about how removed the reader feels from the action, P5 about word and sentence length, and P2 noticed gender in the narrative voices. No generalized representation of style will perfectly satisfy every individual's personal judgments, and should not claim to. The value lies in encouraging the interpretation.

DISCUSSION

Writers have long used libraries, references, thesauruses, and other tools to help shape their work. Word processors with grammar checking and stylometric heuristics have further shaped how people write, edit, and critique their and others' work. Our computational approach to style opens new possibilities for interactions with word processors, exploratory discovery, collaborative writing and cooperative literary spaces.

Collaborative Style and Editing

In the user study, participants envisioned uses around editing contexts, such as working on a long document like a dissertation (P4), where picking out inconsistent sections quickly is essential, or maintaining a coherent character voice (P6). Indeed, while editing this paper, the Editor application enabled the authors to see places where each had written stylistically incompatible sections. In this case, we sought a unified stylistic voice, but in other situations the visualization could also help ensure styles remain distinct when authors want separate voices (Fig. 1, far right). The style interfaces also provide access to community knowledge: the shared color representation between the Editor, which displays own style, and the Explorer, which shows examples of other writing, supports investigating how one's own style fits into the broader landscape.

Learning Style

Since knowledge of style is tacit, it can be hard to teach and learn, especially for those first encountering literary critique, learning to apply it in their own creative writing, or writing in a new language. Style interfaces may provide assistance in these contexts. Participants noted the benefits of surfacing style through a computational interface for students who are still

learning to critically engage with style: "I can imagine it being useful for students...operating with hunches, to see the breakdown and evidence of what they feel" **(P2)**. Revealing student "hunches," or tacit knowledge, through the visualizations could encourage critical reflection and further engagement. A participant in the formative study discussed the challenges of adjusting her academic writing style for an English-speaking audience:

> **F1** It's just the American way, direct sentences and simple sentences, rather than complex, long sentences. I used to write sentences that [were] like 3, 4 lines long, and that was acceptable in India, which is not how most [of the] English speaking world writes.

Style interfaces could help writers adapt to new style norms through visualizations of current or target styles.

Exploratory Discovery in Online Communities

Nontraditional corpora and cooperative literary communities may be a fruitful application area for computational style tools. In domains such as fanfiction and other free, online writing, automated recommendation systems are not common. Instead, users depend on the community and on content-based search tools to find stories. For fanfiction in particular, where writers share the same characters, settings, and plots, style might provide helpful information to users seeking stories they would enjoy. Since style is subjective, an interface that invites the user's participation and interpretations, such as visualizations, may be more appropriate than black-boxed recommendations. Visualizations could carry meaning across platforms, allowing a user to identify a style they like on a fanfiction site, then find similar styles on another site, like Amazon.com (Fig. 1, second from right). Style interfaces might even extend beyond computer screens, and into physical contexts, such as dynamic screens on book spines (Fig. 1, far left).

Enhanced Experiences

E-readers are beginning to introduce computational tools for interacting with e-books, such as the Kindle X-Ray feature, which displays the frequency of occurrences of names and key terms. Style tools offer another form of computational insight into texts. Furthermore, representations of style need not always be visual; texture could carry a similar information in a tactile manner, opening up modalities for blind or color-blind users (Fig. 1, second from left).

LIMITATIONS AND FUTURE WORK

Some of the confusions identified in the user study may be associated with the performance of the model. While a perfect model is neither possible nor desirable, pursuing higher performance may be valuable. Evaluating different text encoding methods could provide insight into the most effective modeling approach and improve performance; for instance, comparing document embeddings [8] to the encoding used here (combining sequences of parts of speech, word embeddings, and characters). Our data may be influenced by a wide range of stylistic norms, as the crowdworkers who contributed to the dataset come from many countries, and many spoke English as a second language.

Regardless of model performance, the nature of style is such that certain aspects cannot be identified by current techniques. How can we capture 'intent,' or discriminate between subversion or reinforcement of convention in an excerpt of text? These are questions of nuance that currently remain in the human realm.

Exploring other representations of style may be valuable. Nuance is lost in the downprojection from the high-dimensional representation to 2D space; while PCA is a straightforward method to project the style space, it is not the only way. One could imagine anchoring a plane on three specific works, to define a style space based on the characteristics of well-known authors, or using an interactive approach such as that described in [16] to dynamically find relevant views. The user study here focused on web-based interfaces; future work could explore user reactions to visualizations in other contexts (e.g. the envisioned applications on book spines or in online marketplaces). The interfaces could move beyond colors, exploring alternate visual representations, or beyond visual representations entirely, using textural representations or adapting to an individual's own associations with style.

Finally, it is important to note that we do not wish for style interfaces to replace close reading or direct interaction with the texts themselves. Style tools are complementary, assisting in contexts where traditional close reading is not the desired interaction.

CONCLUSION

For most people, knowledge of style is tacit. Formal analysis by literary scholars and existing computational metrics are valuable, but do not necessarily capture people's experience of style. Here we demonstrated a new way to analyze writing focused on people's tacit sense of style. Rather than using categories such as authorship or genre, we created a novel crowdsourced dataset of direct comparisons of style, leveraging the tacit knowledge of hundreds of readers, and published the dataset for others to use. We then developed a machine learning model to predict these comparisons, yielding a high-dimensional style space. Using the model, we created interactive tools for the exploration and editing of style. In a user study, we found that such interfaces afford new interactions with style and provoke creative, critical engagement. Addressing the tacit dimension of style opens up exciting new directions for computational style research and interactive style interfaces.

ACKNOWLEDGEMENTS

We thank Dr. Abigail De Kosnik for her encouragement, and our anonymous reviewers for their valuable feedback. We also want to acknowledge the generous *AI For Everyone* grant from Figure Eight that supported crowdsourcing the dataset.

REFERENCES

[1] 2017. Hemingway Editor. (29 December 2017). http://www.hemingwayapp.com/

[2] Sameer Agarwal, Josh Wills, Lawrence Cayton, Gert Lanckriet, David Kriegman, and Serge Belongie. 2007.

Generalized non-metric multidimensional scaling. In *Artificial Intelligence and Statistics*. 11–18.

[3] Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *TEXT* 23 (2003), 321–346.

[4] U. Athira and Sabu M. Thampi. 2015. Hallmarking Author Style from Short Texts by Multi-Classifier Using Enhanced Feature Set. In *Proceedings of the Third International Symposium on Women in Computing and Informatics (WCI '15)*. ACM, New York, NY, USA, 284–289. DOI: http://dx.doi.org/10.1145/2791405.2791444

[5] Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, 313–322.

[6] Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. Sage.

[7] Tess Crosbie, Tim French, and Marc Conrad. 2013. Towards a Model for Replicating Aesthetic Literary Appreciation. In *Proceedings of the Fifth Workshop on Semantic Web Information Management (SWIM '13)*. ACM, New York, NY, USA, Article 8, 4 pages. DOI: http://dx.doi.org/10.1145/2484712.2484720

[8] Andrew M. Dai, Christopher Olah, and Quoc V. Le. 2015. Document Embedding with Paragraph Vectors. *CoRR* abs/1507.07998 (2015). http://arxiv.org/abs/1507.07998

[9] William W. Gaver, Jacob Beaver, and Steve Benford. 2003. Ambiguity As a Resource for Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*. ACM, New York, NY, USA, 233–240. DOI: http://dx.doi.org/10.1145/642611.642653

[10] Ramyaa Congzhou He and Khaled Rasheed. 2004. Using Machine Learning Techniques for Stylometry. In *IC-AI* (2004-11-19), Hamid R. Arabnia and Youngsong Mun (Eds.). CSREA Press, 897–903.

[11] J Berenike Herrmann, Karina van Dalen-Oskam, and Christof Schöch. 2015. Revisiting Style, a Key Concept in Literary Studies. *Journal of Literary Theory* 9 (03 2015). DOI:http://dx.doi.org/10.1515/jlt-2015-0003

[12] Stefan Jänicke, Greta Franzini, Muhammad Faisal Cheema, and Gerik Scheuermann. 2015. On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. In *Eurographics Conference on Visualization (EuroVis) - STARs*, R. Borgo, F. Ganovelli, and I. Viola (Eds.). The Eurographics Association. DOI: http://dx.doi.org/10.2312/eurovisstar.20151113

[13] Jussi Karlgren. 2004. The whys and wherefores for studying textual genre computationally. In *AAAI Fall Symposium on Style and Meaning in Language, Art and Music*.

[14] D. A. Keim and D. Oelke. 2007. Literature Fingerprinting: A New Method for Visual Literary Analysis. In *2007 IEEE Symposium on Visual Analytics Science and Technology*. 115–122. DOI: http://dx.doi.org/10.1109/VAST.2007.4389004

[15] Foaad Khosmood and Robert Levinson. 2011. Taxonomy and Evaluation of Markers for Computational Stylistics. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*.

[16] H. Kim, J. Choo, H. Park, and A. Endert. 2016. InterAxis: Steering Scatterplot Axes via Observation-Level Interaction. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (Jan 2016), 131–140. DOI: http://dx.doi.org/10.1109/TVCG.2015.2467615

[17] J. P. Kincaid, J. A. Aagard, J. W. O'Hara, and L. K. Cottrell. 1981. Computer readability editing system. *IEEE Transactions on Professional Communication* PC-24, 1 (March 1981), 38–42. DOI: http://dx.doi.org/10.1109/TPC.1981.6447821

[18] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*. 1188–1196.

[19] Nedim Lipka and Benno Stein. 2010. Identifying Featured Articles in Wikipedia: Writing Style Matters. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*. ACM, New York, NY, USA, 1147–1148. DOI: http://dx.doi.org/10.1145/1772690.1772847

[20] Nina McCurdy, Julie Lein, Katharine Coles, and Miriah Meyer. 2016. Poemage: Visualizing the sonic topology of a poem. *IEEE transactions on visualization and computer graphics* 22, 1 (2016), 439–448.

[21] K. McLaren. 1976. XIII The Development of the CIE 1976 (L* a* b*) Uniform Colour Space and Colour-difference Formula. *Journal of the Society of Dyers and Colourists* 92, 9 (1976), 338–341. DOI: http://dx.doi.org/10.1111/j.1478-4408.1976.tb03301.x

[22] Sven Meyer zu Eissen, Benno Stein, and Marion Kulig. 2007. *Plagiarism Detection Without Reference Collections*. Springer Berlin Heidelberg, Berlin, Heidelberg, 359–366. DOI: http://dx.doi.org/10.1007/978-3-540-70981-7_40

[23] Franco Moretti. 2007. *Graphs, Maps, Trees: Abstract Models for Literary History*. Verso.

[24] Frederick Mosteller and David L. Wallace. 1963. Inference in an Authorship Problem. *J. Amer. Statist. Assoc.* 58, 302 (1963), 275–309. http://www.jstor.org/stable/2283270

[25] Aditi Muralidharan and Marti A Hearst. 2012. Supporting exploratory text analysis in literature study. *Literary and linguistic computing* 28, 2 (2012), 283–295.

[26] Antonio Neme, J.R.G. Pulido, Abril Muñoz, Sergio Hernández, and Teresa Dey. 2015. Stylistics analysis and authorship attribution algorithms based on self-organizing maps. *Neurocomputing* 147, Complete (2015), 147–159. DOI: http://dx.doi.org/10.1016/j.neucom.2014.03.064

[27] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543.

[28] Maria Soledad Pera and Yiu-Kai Ng. 2015. Analyzing Book-Related Features to Recommend Books for Emergent Readers. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media (HT '15)*. ACM, New York, NY, USA, 221–230. DOI: http://dx.doi.org/10.1145/2700171.2791037

[29] Michael Polanyi. 1967. *The Tacit Dimension*. Anchor Books.

[30] Raymond Queneau and Barbara Wright. 1958. *Exercises in style / by Raymond Queneau; translated by Barbara Wright* (first english. ed.). Gaberbocchus London. 198 pages.

[31] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019).

[32] Dmitry Shalymov, Oleg Granichin, Lev Klebanov, and Zeev Volkovich. 2016. Literary writing style recognition via a minimal spanning tree-based approach. *Expert Systems with Applications* 61, Supplement C (2016), 145 – 153. DOI: http://dx.doi.org/10.1016/j.eswa.2016.05.032

[33] Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *J. Am. Soc. Inf. Sci. Technol.* 60, 3 (March 2009), 538–556. DOI: http://dx.doi.org/10.1002/asi.v60:3

[34] Efstathios Stamatatos, George Kokkinakis, and Nikos Fakotakis. 2000. Automatic Text Categorization in Terms of Genre and Author. *Comput. Linguist.* 26, 4 (Dec. 2000), 471–495. DOI: http://dx.doi.org/10.1162/089120100750105920

[35] Omer Tamuz, Ce Liu, Serge J. Belongie, Ohad Shamir, and Adam Kalai. 2011. Adaptively Learning the Crowd Kernel. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*. 673–680. https://icml.cc/2011/papers/395_icmlpaper.pdf

[36] Paula Cristina Vaz, David Martins de Matos, and Bruno Martins. 2012. Stylometric Relevance-feedback Towards a Hybrid Book Recommendation Algorithm. In *Proceedings of the Fifth ACM Workshop on Research Advances in Large Digital Book Repositories and Complementary Media (BooksOnline '12)*. ACM, New York, NY, USA, 13–16. DOI: http://dx.doi.org/10.1145/2390116.2390125

[37] Paula Cristina Vaz, Ricardo Ribeiro, and David Martins de Matos. 2013. Book Recommender Prototype Based on Author's Writing Style. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval (OAIR '13)*. 227–228.

[38] W. Weber. 2007. Text Visualization - What Colors Tell About a Text. In *Information Visualization, 2007. IV '07. 11th International Conference*. 354–362. DOI: http://dx.doi.org/10.1109/IV.2007.108

[39] Zaihan Yang and Brian D. Davison. 2012. Writing with style: Venue classification. *Proceedings - 2012 11th International Conference on Machine Learning and Applications, ICMLA 2012* 1 (2012), 250–255. DOI: http://dx.doi.org/10.1109/ICMLA.2012.50

[40] Ying Zhao and Justin Zobel. 2007. Searching with Style: Authorship Attribution in Classic Literature. In *Proceedings of the Thirtieth Australasian Conference on Computer Science - Volume 62 (ACSC '07)*. Australian Computer Society, Inc., Darlinghurst, Australia, Australia, 59–68.