Turbulent Flow: A Computational Model of World Literature

Hoyt Long and Richard Jean So

Abstract This article uses computational modeling and large-scale pattern detection to develop a theory of global textual transmission as a process of turbulent flow. Specifically, it models stream-of-consciousness narration as a discrete set of linguistic features and rhetorical elements and uses this model to track the movement of this modernist technique across generic boundaries (from anglophone modernism to more popular genres) and linguistic ones (from English to Japanese). Oscillating between statistical models and moments of close reading, the article shows how a quantitatively scaled-up approach, rather than reinforcing an image of global textual flows as singular and monolithic, illuminates world literature as a system constituted by patterns of divergence in structure and of difference in sameness.

Keywords world literature, computational modeling, modernism, Japan, stream of consciousness

"It is no surprise," writes Franco Moretti (1996: 124), "that the stream of consciousness should be the most famous technique of the twentieth century: in view of what it has done, it fully deserves to be." Volumes of criticism on stream of consciousness (SOC) attest that it has indeed done many things, not the least of which is travel the world. Moretti himself counts "two basic 'strains,' and . . . a dozen variants," many generated from "the morphological explosion of the twenties" that spread across Europe and America (177, 178). By the late 1920s it had reached Japan; by the 1950s, Latin America. SOC is an ideal object for studies of world literature. Yet how do we analyze its variety of forms and their geographic reach? What can we say about its diffusion?

An initial hypothesis: SOC functioned as a high-prestige international form that moved through an integrated world literary system in a wavelike pattern. The hypothesis comes from scholars like Moretti and Pascale Casanova, who have pushed for theoretical models that picture

Modern Language Quarterly 77:3 (September 2016)
DOI 10.1215/00267929-3570656

world literary space after the late nineteenth century as a unified system in which "foreign form[s] and local materials" are combined according to the laws of a hierarchical market of literary exchange (Moretti 2000: 60; Casanova 2004). In this article we test this hypothesis using a computational approach. First, we develop an abstract model of SOC so as to identify it algorithmically in large bodies of texts in different languages via specific linguistic features and rhetorical elements. We then show that the patterns of SOC thus discovered can refine a systems-based theory of world literature by clarifying its assumptions about universal guiding principles and frictionless textual flows. These patterns expose degrees of variance and deviation in much larger structures of diffusion, leading us to a theory of SOC as a form that traveled beyond modernism into more popular genres—and beyond English into other languages—as a set of features held together in varying but measurable combinations. In short, we propose a theory of stylistic transmission as *turbulent flow*.

The apparent variability and definitiveness of SOC as a literary object invite such an approach. No one can agree on what it is, yet everyone seems to know it when they see it. The critical scholarship on SOC over the past fifty years adds to Moretti's two strains and a dozen variants a long list of categorical descriptions and identifying features: it is a novel, a genre, a method, a technique, or a style (Humphrey 1954: 1). Stylistically, it is defined by interior monologue, soliloquy, free indirect discourse, free association, imagery and symbols, irregular punctuation, fragmented sentences, ellipses, paratactical paragraphs, discontinuous syntax, onomatopoeia, sensory perception, lexical opaqueness, or lexical bombardment. These features, we are told, are used in varying degrees and combinations across the whole population of writers who experimented with the form. Wyndham Lewis underscored this cacophony in his description of SOC as a "jellyfish structure . . . without any articulation of any sort" (quoted in Edel 1964: 91). However, even as the varieties of SOC have proliferated, there remains a tendency to see it as a definite, empirical thing—"the most famous technique of the twentieth century."

The question of whether the form is infinitely varying or inherently measurable is not unique to SOC. We confront it whenever we address aesthetic objects that seem to persist over time and space even as they are recomposed by local formal compromises. Computation offers a way to

reason across this difference. It allows us to reconstruct shared structural patterns while identifying degrees of difference in sameness that are constitutive of the structure itself. Ultimately, computation generates a theoretical space where apprehensions of texts as constitutive of larger structures or as radically singular can coexist, illuminating world literature as both structure and divergence, diffusion and variance.

Modeling SOC and Its Anglophone Diffusion

We begin with the assertion that the literary form known as stream of consciousness can be modeled. By *model* we mean a quantitative representation of a more complex underlying reality: here, a literary form that contains countless aspects and manifestations depending on the particular orientation of the writer and/or reader. A model will never capture the full complexity of SOC, but it can seek to identify the aspects that account for the most substantial portions of its many articulations in different contexts. Through this reduction such a model can be used to compare a large number of texts computationally using a limited set of quantifiable dimensions in a systematic way. Modeling makes it possible to compare not one text with others across infinitely many dimensions but groups of texts across a finite set of dimensions aggregated and averaged over each group.

While scholars do not agree on a single definition of SOC, they do concur on a number of underlying features. Some of these features represent abstract concepts, such as "internal analysis" (Bowling 1950: 344), while others appear highly technical and empirical, such as "truncated syntax" (Chatman 1978: 188). Scholars will also characterize SOC as constituted by explicit literary techniques, such as "interior monologue," which are difficult to reliably identify or measure (Humphrey 1954: 23). There exists a wide variety of possible features. Yet the persistent claim that SOC arises through a discrete set of these features suggests that it can be modeled and detected across multiple instances.

To develop our model, we began by collecting a corpus of SOC passages. We selected sixty novels identified by scholars as having elements of SOC and then isolated passages that specifically embody this narrative trope. Typically, SOC appears in flashes, rarely establishing the overall form of the novel. Thus we extracted SOC passages in two ways.

First, we took what scholars have explicitly identified as SOC sequences of writing from thirty novels. On average, these passages were twelve hundred characters long, and we collected five from each novel for a total of 150. We then looked through the other thirty novels and pulled out five passages from each (of similar length) that we jointly agreed were SOC, yielding 300 passages in all. Such passages come from archetypal novels such as *Ulysses*; novels influenced by SOC but less canonical in their reception, such as Conrad Aiken's *Blue Voyage*; English-language translations, such as Marcel Proust's *Swann's Way*; and novels in which SOC passages appear only infrequently, such as Jean Toomer's *Cane*.

Next we had to build a corpus against which to compare these passages and identify which features most distinguish SOC as a mode of narration. In literary scholarship, SOC is sometimes viewed as a stylistic break with literary realism in both American and British traditions (Scholes and Kellogg 1966: 193–206), suggesting that there exists a set of literary features that reliably differentiates one from the other. Even granting that certain forms of narrated consciousness are present in realist works, SOC is often regarded as marking a new, if nevertheless related, stage in the evolution of these forms. To test this hypothesis, we constructed a corpus of sixty realist novels that are generally viewed as not making significant use of SOC and randomly sampled five passages from each, creating a parallel corpus of 300 passages drawn chiefly from canonical works, such as *Middlemarch*, *House of Mirth*, *Sister Carrie*, *The Portrait of a Lady*, *Wuthering Heights*, and *Great Expectations*.[1] While we acknowledge that the distinction between SOC and realism is not as firm as some scholars have argued, we refer to this corpus as "realist" for the sake of convenience.

The next step was to select and extract a set of textual features with which to build our model of SOC. We selected the features that literary

[1] In contrast to our careful curation of SOC passages, the sampling method used for realist novels was crude. After dividing them into equal-size passages fifteen hundred characters in length (roughly a page of text), we had the computer randomly select five from each. There are obvious limitations to this approach, in particular the fact that it does not distinguish between passages focused on dialogue and those focused on description. In the future, we hope to implement a more sophisticated method that would exclude dialogue.

critics have identified as most prevalent in SOC and those that we could operationalize or quantify. Some were easy: we can measure the amount of onomatopoeia in a text by simply counting its instances.[2] Others, such as third-person narration, proved too difficult. Last, there were features that did not immediately lend themselves to quantification yet could be represented by proxy as numbers. For example, we can capture semantic complexity in part by taking the number of different words in a text (types) and dividing it by the total number of words (tokens). A high type-token ratio indicates a large degree of lexical variation in a text, while a low ratio indicates the opposite. We assume that seemingly crude empirical values, like type-token ratio, index latent stylistic qualities in a literary text. We identified thirteen features with which to compare SOC against realism.[3]

Finally, we wrote a program to tabulate the proportion of each feature appearing across the entire population of SOC and realist passages. An algorithm uses this table to learn which features are associated with each type of passage and then tries to classify passages as either SOC or realist in a process called cross validation. Essentially, the algorithm iteratively trains on a subset of the data and is then shown passages from which identifying labels have been removed. It tries to predict whether these masked passages are SOC or not based on the previously observed data. The more accurately it predicts the correct label, the more confident we are that our model features reliably distinguish between the two categories. An accuracy rate of 80–85 percent is typically a strong indicator of statistical reliability. When we tested our model in this manner,

[2] To identify onomatopoeia, we relied on a dictionary-based approach. This meant gathering lists of onomatopoeia from various sources and searching for these terms in the selected passages. Because these sources are skewed toward the contemporary period, we are probably missing historical instances of onomatopoeia that are no longer in use. A more historically precise list is a goal for future projects.

[3] The thirteen features are (1) average sentence length, (2) proportion of nominalized sentences, (3) proportion of verbless sentences, (4) proportion of sentences beginning with a gerund phrase, (5) ratio of personal pronouns to all words in a passage, (6) proportion of sentences beginning with a personal pronoun, (7) type-token ratio, (8) type-token ratio without stop words, (9) type-token ratio without proper nouns, (10) ratio of onomatopoeic words to all words, (11) ratio of neologisms to all words, (12) ratio of ellipses to passage length, and (13) proportion of sentences using simple free indirect discourse.

we achieved a rate of about 95 percent, which means that it is excellent at distinguishing SOC from realism.[4] This model confirms our original intuition. Wildly different results would suggest that the model was incorrect. Yet our results also bring to light an important finding: SOC, compared to realism, depends on a specifiable set of linguistic traits. While scholars have debated whether SOC is a unified style or infinitely varying, our model points to some degree of formal unification.

This core discovery facilitates additional insights in that the classification results also reveal the relative importance of each feature in distinguishing SOC. Each feature receives a different weight in the process because some features appear prominently in SOC and thus help distinguish SOC from non-SOC passages, while others appear infrequently and thus are less useful in the classification process, or else appear more frequently in the non-SOC passages. After classification the relative weight of each feature is reported. In our tests, for example, we learn that the features associated with pronouns are statistically insignificant in predicting whether a passage is SOC or not. If we simply remove these features from the model, it will perform equally well. This result runs counter to some existing scholarship (Cohn 1978: 94; Dahl 1970: 16). We also learn that the most distinguishing feature is type-token ratio. If the type-token ratio increases by 1 (i.e., increasing lexical diversity), the odds that a passage is SOC increases by a factor of 18. This feature is very important in recognizing SOC style. Scholars have intuited the importance of type-token ratio in SOC, but never with such precision or on such a scale (Dahl 1970; Steinberg 1973: 155–58).[5] Finally, and most

[4] The determination of "excellent" or "poor" accuracy greatly depends on the underlying data and the nature of the research question. There is no universal standard. Typically, the researcher determines a reasonable baseline accuracy based on knowledge of the data. Before running our tests, we posited that SOC and non-SOC passages would look very different in our model and that the classifier would be able to identify them more than 50 percent of the time (better-than-random guessing). Further, we posited that SOC is distinct enough that a 75 percent accuracy rate would be unacceptable. Because our result of 95 percent is well above this threshold, we deem the outcome "excellent."

[5] According to our model, seven features indicate a higher likelihood that a passage is SOC: sentence length, verbless sentences, sentences beginning with gerund phrases, high type-token ratio without stop words, onomatopoeia, neologisms, and free indirect discourse.

critically, our model tells us that our list of features alone is almost always enough to differentiate SOC from realist passages. The number of things we need to know about a passage to identify it as SOC is finite.

This analysis is in and of itself productive, but our article focuses on a series of broader historical and sociological questions. To restate the initial hypothesis, borrowed from Moretti: After the eighteenth century "world literature" is like a "wave." It starts in places like France ("the core") and emanates to nations like Japan ("the periphery"), and as it moves, it is increasingly defined by "sameness." Literatures that appear at the periphery converge on the same forms that appear at the core (Moretti 2011: 70–71). Before testing this broader hypothesis, however, we look at SOC's diffusion into more linguistically proximate regions. In particular, we trace SOC's diffusion into the anglophone "semiperiphery," a region that mediates between core and periphery. At the same time, we track SOC's diffusion as it travels from high modernism to the broader field of literature, including popular fiction. These are different but equally important modes of diffusion. Thus we add to our initial hypothesis two more specific subtheses: (1) SOC diffuses from core to semiperiphery, as well as from core to periphery, and (2) SOC diffuses from high modernism to popular genres of writing.

To track SOC's diffusion, we created a corpus of seventeen hundred anglophone novels primarily from the United States but also from England, Ireland, Scotland, Australia, Canada, and South Africa.[6] We removed any novels already identified as SOC and then sampled five passages each from roughly eight hundred of the remaining novels, exactly as we had done with the realist novels. We then used the same classification process as before to distinguish SOC from non-SOC passages written between 1923 and 1950. The earlier test had confirmed that SOC and realist passages (the latter taken largely from late nineteenth-century novels) represent measurably distinct forms of writing. Using the same set of features, we found here too that SOC passages are highly distinct from non-SOC ones; the classifier separated them at a

[6] Using WorldCat records, we identified the ten thousand novels in this period from these seven nations that were most commonly held in US libraries. We then winnowed this list down to seventeen hundred by selecting those novels that existed in digital format. While this is by no means a complete representation of literary output in this period, we believe that it provides a reliable and diverse sample.

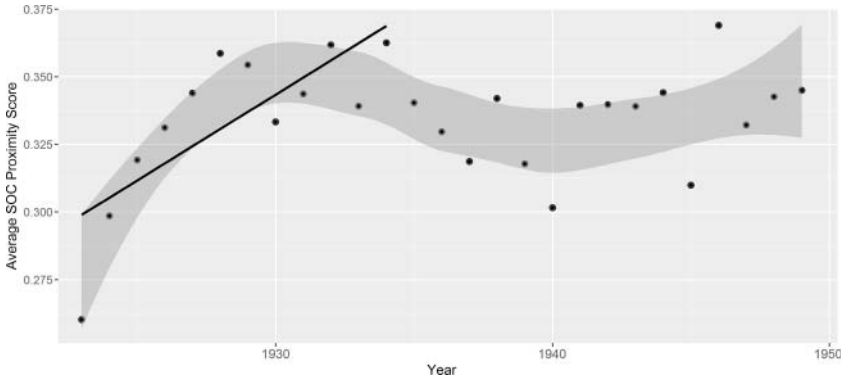

Figure 1. Average SOC proximity score per year for all sampled novels, 1923–50. Each point represents the average score for all sampled novels published in that year. The black line is the fitted regression line for 1923–34, showing a statistically significant upward trend. The gray bar represents the overall trend of the data.

rate above 90 percent. The results from the classifier can also be used to predict how close any individual passage is to the ideal SOC passage. Averaging these scores across the five passages from each novel, we obtain a proximity score for that novel between 0 and 1. The closer to 0 a novel is, the more likely it is to have elements of SOC; the closer to 1 it is, the more likely it is not to.

A plot of the average proximity score for all sampled novels by year shows a statistically significant increase in SOC-like novels after 1923 (fig. 1). For clarity, the scores are inverted in this graph, so that 1 now indicates proximity to SOC. Between 1923 and 1934 the value of each point increases toward 1, indicating that the novels increasingly contain SOC-like passages. This trend flattens out after 1935. These results suggest that more writers experimented with SOC after its first appearance in the early 1920s, but this initial moment of diffusion had its limits. One can easily speculate why. Literary historians have argued that the Great Depression pushed American and British writers to reject the previous decade's interest in modernist techniques and embrace explicitly outward, political forms of writing (Denning 1997; Foley 1993). The flattening of our trend line perfectly coincides with this narrative. But it does not indicate an absolute decline. SOC merely enters a holding pattern.

If SOC diffuses more broadly into the literary field between 1923 and 1934, then what specific factors drive that diffusion? To test more specific hypotheses about national or generic influence, we collected data about the nationality of each author in the sampled corpus and used a standard statistical model, logistic regression, to determine if nationality predicted the likelihood that a novel was more proximate to SOC. The results showed nationality to be statistically insignificant, meaning that it cannot predict whether a novel will contain more SOC passages or fewer. It appears that anglophone writers were equally likely to adopt SOC between 1923 and 1934, the period of its greatest increase in anglophone literatures. Moreover, after controlling for gender and race, we found that these—whether in combination with nationality or as independent variables—have no role in predicting proximity to SOC. Strikingly, nationality, race, and gender all lack explanatory power as variables.

What about genre? Critics argue that SOC represents an avant-garde, high literary style that over time remains the exclusive property of "serious" writers. If it diffuses through the literary field after its initial appearance in the early 1920s, it does so through a narrow range of novelists, such as Thomas Wolfe and Djuna Barnes (Friedman 1955). From this perspective we would expect popular genres of writing, such as romantic fiction, to be resistant to adopting such an avant-garde form. SOC is more likely to be picked up by the spiritual and aesthetic heirs of James Joyce. Our results suggest the opposite. Collecting genre data for each novel, we used the same model described above to determine if genre could predict more SOC-like novels. The genres represented in the corpus include modernism, realism, historical fiction, romance, adventure stories, science fiction, and several others, with the most dominant being detective fiction (roughly 33 percent of the corpus). Despite such generic diversity, genre too proves insignificant. Here, however, the ramifications of this insignificance are more profound. It suggests that writers of high and low, modernist and popular fiction, were equally likely to experiment with SOC between 1923 and 1950. There is no indication that SOC exclusively diffuses within an elite and narrow band of writers. Indeed, the only genre that resists SOC significantly is detective fiction, where SOC is scarce relative to the overrepresentation of this genre in the corpus. Otherwise, the form's diffusion knows no generic bounds. Romance writers were just as likely to take it up as Barnes.

That neither nationality nor genre organizes the diffusion of SOC after 1922 raises the question of what our model is in fact identifying as SOC. Looking at an example of a passage classified as SOC is helpful. We could examine passages from novels that both our model and traditional scholarship single out as having elements of SOC, such as Wolfe's *Look Homeward, Angel* (1929) or Barnes's *Nightwood* (1936). Rather than confirm existing critical judgments, however, we want to consider how our model broadens an understanding of SOC, and its materialization in the overall literary field, by pointing to novels that fall well outside SOC scholarship.

A good example is Jeffery Farnol's *Way Beyond*, a novel published in 1933 by Little, Brown. It is ranked the most SOC-like novel by our model, even higher than works by Barnes and Wolfe. Farnol was a popular British author who wrote more than thirty genre novels, primarily romances. *The Way Beyond* is indeed a work of romantic fiction, a midcentury best seller of the genre. Unsurprisingly, we find no scholars linking SOC to Farnol's work. Scholars of modernism and SOC tend to follow Joyce's contempt for the romance genre. In an often-cited interview Joyce starkly contrasts literary modernism against romance, denigrating the latter. If novels like *Ulysses* realistically illuminate life as "jugs, and pots and plates, back-streets and blowsy living-rooms inhabited by blowsy women, and . . . a thousand daily sordid incidents," the romance novel merely provides "flimsy drop scenes" (quoted in Power 1974: 75). Since *Ulysses* the ordinariness of high modernism has been pitted against the unreality of romances.

But consider the following passage from *The Way Beyond*, a common scene rendered in language typical of the novel:

> Thus then they sat, Rosemary staring down at the bonnet strings her strong, shapely fingers were twisting and Richard gazing at her beautiful, down-bent face, whose loveliness was made even more alluring by its sudden, bewildering changes, or so thought Richard: This nose, for instance, though perfect in itself, yet because of its delicate, so sensitive nostrils, became positively adorable; this rose-red mouth, with its sweet, subtle curve of mobile lips, broke his heart when it dropped . . . and, by heaven, it was dropping now! He seized her hands to kiss and kiss them, he lifted her head that he might look down into her eyes, and gazing into these tender deeps, he questioned her in a voice anxious and a little uncertain. (Farnol 1933: 30)

While the narrative content is overwrought, stylistically this passage would not be wholly out of place in a novel by Virginia Woolf. It bears rhetorical features familiar to SOC: the long, meandering sentences; the alliterating language of description ("rose-red mouth"); the use of ellipsis to mark a character's state of reflection or contemplation; and the most definitive hallmark of SOC, free indirect discourse. This last facilitates a key moment in the passage—"and, by heaven, it was dropping now!"—the moment in which our protagonist turns physically passionate. The high SOC score assigned this novel by our model thus seems far from arbitrary. The text emits a set of rhetorical signals that allies the novel with more canonical examples of SOC, such as *Mrs Dalloway* or *Nightwood*. Of course, Farnol's novel does not embody SOC in the same way as Woolf's or Barnes's. But this novel and others like it, bearing the traces of specific stylistic features, fall well within the penumbra of SOC that our model defines, sometimes squarely so.

This passage from Farnol offers some useful granularity to our otherwise confounding results. Our model confirms the general hypothesis that SOC diffuses through an integrated literary market by borrowing and consolidating specific stylistic features. Yet there is no evidence that this diffusion is organized by nationality (the core-to-semiperiphery thesis) or genre (a standard claim of scholarship). SOC is distributed across the anglophone field without obvious bias. In Farnol we catch a glimpse of how that distribution plays out at the level of narrative style and technique. *The Way Beyond* indexes SOC's evolution in the decade after *Ulysses*, the way it radiates beyond a coterie of high modernist writers in England and America while retaining certain core linguistic features, such as free indirect discourse. In this case, nationality and genre do not help us grasp the contours of that diffusion. But individual texts do, particularly as delivered up by the naive but principled reading practices of computational models.

Modeling SOC in Japan

We began with the hypothesis that SOC moved across the world in a wavelike pattern through processes of formal integration and consolidation. Thus far, however, our view of the world has been quite limited. To fully test our hypothesis, we need to look beyond anglophone texts to

see if "the *streamlining of formal solutions* imposed by the world market" holds true (Moretti 2011: 71). To date, very few computational studies have attempted to identify the movement of literary forms across languages.[7] Technical complexity and unevenness in digital corpora are partly to blame for this, but a more fundamental issue is that as formal and stylistic elements move through the world literary system, those less independent of language encounter "all sorts of obstacles" (74). For his part, Moretti draws a line between "plot" and "style" in thinking about what gets lost in the global circulation of forms. Here we begin with the assumption that some elements of form do travel more easily than others and that these can be modeled across languages. Building on our anglophone model, we test the extent to which the rhetorical features of SOC persisted as it entered new linguistic contexts. Which were singled out as unique and necessary to SOC, and which fell prey to linguistic obstacles? How did they fit in the existing ecology of literary forms?

Japan is an interesting test case in this regard because of the stark grammatical differences between English and Japanese and because of the profound, if fleeting, impact that SOC had on Japanese modernist writers starting in the late 1920s. In 1929 a critical review and partial translation of *Ulysses* set this wave of influence in motion (Doi 1929). By 1930 a complete translation of *Ulysses* began serialization in an avant-garde literary journal, followed by another in 1932. Additionally, partial translations of *To the Lighthouse* (1931), *Swann's Way* (1932), and Paul Morand's *Open All Night* (1929) appeared alongside dozens of essays attempting to define SOC and establish its literary significance.[8] Within months of the 1929 review, up-and-coming writers began experimenting with SOC, producing nearly twenty works of short fiction inspired by the technique.

A popular theme in the critical literature was the question of how to domesticate a technique that, primarily through *Ulysses*, was considered a radical innovation in psychological realism. Contributors to this debate saw SOC as a definite, concrete force to be reckoned with. Yet for all the attention they paid it as a new way to represent the complexities of

[7] One important exception is Piper 2015.

[8] Ōta Saburō (1955) compiled a bibliography of essays and translations related to Joyce and more generally to the SOC style. He lists over two hundred items between 1918 and 1941, the bulk of them between 1929 and 1933.

the human mind, or as a style worth assimilating, they rarely discussed technical specifics. Critics cited passages from core European models of SOC (e.g., by Dujardin, Joyce, Woolf, and Proust), translated canonical SOC novels, and even tried the style out for themselves. But the lack of technical discussion leaves open the question of what writers and translators saw when they looked at SOC. Did those who devised formal solutions to the problem of adapting SOC to Japanese rely on the same features that tended to differentiate it in English? Our goal was to test whether some combination of the previously demonstrated features was preserved in this new context.

Doing so required translating our original model into this new linguistic context without significantly altering its core features and our ways of measuring them. This process was made easier by our model design, which ignores semantic content and identifies grammatical features with direct or approximate equivalents in Japanese. It is simple to measure sentence length, extract dialogue, and count ellipses, for example, because writers in Japan had by the early 1900s incorporated most Western punctuation practices. Parts of speech are not as well defined in Japanese as in English, but it is still grammatically meaningful to track the use of personal pronouns as well as verbless or nominalized sentences. The latter two features similarly serve as a rough proxy for incomplete or fragmented sentences. Moreover, the word-based features in our model are relational, relying on lexical categories (e.g., types, tokens, neologisms, onomatopoeia) rather than on meaning. Finally, while free indirect discourse is marked in Japanese by a different set of grammatical and lexical patterns, these are still empirically detectable and thus provide an adequate proxy for this feature.[9]

The first critical test was to see if our model reliably distinguished between Japanese SOC texts and select works of fiction produced prior

[9] We were able to replicate all but one feature used in the English model. What we could not track were sentences beginning with a gerund or adverbial phrase. While grammatically possible in Japanese, these phrases never occur at the head of a sentence. Also, in detecting free indirect discourse, we do not account for whether a work is narrated in the first or the third person. Instead, we identify nondialogue sentences whose ending is marked by grammatical indicators of interior monologue or personal address. These can be questions, exclamatory statements, statements of supposition, or volitional phrases.

to its introduction. Thus we chose thirty titles published before the SOC boom that reflected the kind of narrative style that early adopters of SOC sought to overcome. This served as an analog realism corpus. While Japan did not have European-style realism, it had things like it, developed in response to literary currents in the West, in particular a mode of psychological realism characterized by intensely self-referential and confessional accounts of individual lives that became a direct counterpoint to the mode that SOC adopters attempted to introduce. We thus populated the realism corpus with canonical texts that belonged to this dominant strain of early twentieth-century Japanese letters.[10] The second test was to see how well our model distinguished Japanese SOC texts from works produced contemporaneously. Thus we identified thirty titles to serve as a control corpus. These texts are temporally proximate to the SOC adaptations and translations but come from popular genres and writers unaffiliated with the technique. Detective fiction comprises about 80 percent of this corpus; works of historical and proletarian fiction make up the rest.[11] For the SOC corpus, we identified thirty titles from which to extract passages evocative of SOC style. This included recognized or self-proclaimed SOC experiments; translations of canonical SOC works, such as *To the Lighthouse* and the two translations of *Ulysses*; and a handful of titles from before 1928 that were treated as harbingers of a modernist turn in the representation of subjectivity.[12]

We then classified the Japanese SOC passages against the realism and control corpora in the same way as in the anglophone case. With the realist texts, the model guessed the true class label with astonishing 97 percent accuracy. Furthermore, only seven features were needed to accomplish this. What appears to distinguish SOC in Japan, as compared with earlier narrative styles, is the presence of onomatopoeia and

[10] Authors represented include Kasai Zenzō, Natsume Sōseki, Shimazaki Tōson, Tayama Katai, and Tokuda Shūsei.

[11] Authors represented include Hamao Shirō, Kobayashi Takiji, Kosakai Fuboku, Kunieda Shirō, and Unno Jūza. The decision to include this much detective fiction was based on an initial hunch that it offered the best example of a contemporary, nonelite form distinct enough in narrative mode but not too linguistically distant from SOC. Because this biases our results, however, future tests will need to include a more generically balanced control corpus.

[12] Authors represented include Hori Tatsuo, Itō Sei, Kaji Motojirō, Kawabata Yasunari, and Yokomitsu Ri'ichi.

neologisms, the use of ellipses, high type-token ratio, nominalized sentences, and the use of free indirect discourse. These stylistic traits, viewed in combination, inscribe a decisive gap from earlier modes of writing, even though much of that writing was similarly concerned with narrating internal thoughts. Indeed, the gap is so great that when the scores for each passage were averaged over all the passages from a given title, only one text fell below the halfway point (0.5) nominally separating SOC from non-SOC texts.[13] The vast majority (twenty-four titles) scored above 0.8. The most SOC-like title is Natsume Sōseki's famous *Kokoro* (1914), a novel with large portions of first-person narration that evoke Laurence Sterne in how they preserve fragmentation, digression, and insignificance as tokens of authentic emotional expression and spiritual depth (Brodey 1998: 208).[14]

The degree of separation here reinforces much of the commentary about how and why SOC diffused into Japan. Many of the earliest and most avid adopters were younger writers just getting their footing in the literary establishment. The appearance of this internationally recognized and radically distinct style was an opportunity to expand, as Pierre Bourdieu (1996: 125) puts it, the local space of literary possibilities at a moment when "revolution" still imposed itself as "the *model* of access to existence in the [literary] field." A contributor to the movement, Nagamatsu Sadamu (1931: 264), said as much, emphasizing that the "method" of SOC signaled the "departure point of a new literature" that helped young writers "quickly and eagerly catch up" with a new and necessary way of viewing contemporary reality. Later critics downplayed these efforts as a superficial formalist obsession leading to exaggerated and haphazard use without the integration into established narrative structures that could force a deeper rethinking of the relation of self to society (Hojō 1980; Ōta 1955; Sharif 2003). While both perspectives

[13] This halfway point is not arbitrary but statistically determined. It is based on the observed distributions of our classification results and indicates the dividing line between where most SOC- and non-SOC-classified passages fall. In all our cases, it happens to fall at the 0.5 mark. This becomes the predictive threshold for SOC passages.

[14] Sōseki, notably, was a key Japanese interpreter of William James and used his concept of "stream of consciousness" in his major theoretical work, *Bungakuron* (1907), to explicate ideas about the relation of literature to psychology.

acknowledge the distinct formal break that our computational approach confirms, neither speaks to the precise nature and degree of that break. They rely instead on aesthetic explanations that reduce the complex fact of diffusion to judgments of success or failure.

In contrast, a computational approach shows that the break was engineered through a selective appropriation of several features that distinguished SOC in the anglophone context. Some features continued to be positively associated with SOC in Japan, while others took on an inverse association with the form.[15] Some features, like sentence length and verbless sentences, were not important at all. What we have, in short, is empirical evidence of formal continuity across languages, but also of divergence. It is a more precise account of how Japanese writers and translators collectively reached a formal solution to the problem of moving SOC into a new linguistic context. It is a means of seeing at scale the "complex selection process" that Kirsten Silva Gruesz (2002: 28) describes as a way of taking "what [was] useful from the model being copied, and leaving the rest."

Our model also clarifies the differing methods and degrees by which SOC texts achieved a formal solution. It reveals an underlying structure in the group, but also internal variance, allowing us to compare passages based on how close to or far from the ideal SOC model they fall. For instance, of the four titles predicted to be most SOC-like, two are short stories by Nagamatsu and two are different translations of the "Proteus" chapter in *Ulysses*—a chapter well known for its SOC representation of Stephen Bloom's meandering thoughts as he lies atop the rocks at Sandymount Strand. That these texts are judged to be similar is, in retrospect, easily explained. Nagamatsu (1931: 264) not only included his stories in a list of SOC-inspired works but also worked directly on one of the translations of *Ulysses*.[16] History thus already points to the

[15] Specifically, high type-token ratio, onomatopoeia, and free indirect discourse are positively associated with SOC in Japan, as in the anglophone context. Nominalized sentence endings, neologisms, and ellipses have inverse associations; the first two are positively associated with SOC in the Japanese case and negatively associated in the anglophone case. The converse is true for ellipses.

[16] Nagamatsu collaborated with Itō Sei and Tsujino Hisanori on the first translation of *Ulysses*, which began serialization in the journal *Shi genjitsu* in 1930. A second team of translators led by Morita Shinpei began to serialize a translation of the novel in the journal *Bungaku* in 1932.

possibility of a stylistic affinity between these texts. That they are more similar than any of the other texts in the SOC corpus, however, is not a conclusion that could have been reached without a specified model of SOC and without the scale of comparison it affords. It brings the relation of these texts into a different kind of focus, from which new readings and comparisons can begin.

Consider an excerpt from Nagamatsu's "Portrait of Mademoiselle Mako" (1931), which consists of several pages of extended interior monologue. The narration, which begins with the female narrator's musings about a recent love letter, quickly grows more fragmented and dispersed, slipping into and out of multiple recollected conversations before spiraling into a series of repetitive existential questions. Along the way are found scattered ruminations on geopolitics and the fate of Japanese women:

> Ecstasy of the movie theater. True *paradise*. Intoxication that smothers the breast. It's a fact that the ancients clasped their chests with iron rings so that their hearts wouldn't fracture on account of love. Figure of life. *American life*. How bright life is in America. "The Gloryfication [*sic*] of the American Girl"!. Pitiable Japanese girls. Who on earth will glorify Japanese girls? The *dingy* Japanese youth? Oh, Americans' *nonchalance*, Americans' boisterousness, Americans' recklessness, Americans with their hats off to the side, America's money, money, money—the Americans are all taking over the world. And then in order to compete with them, the red *Soviets*. *G.P.U.* 5 Year Plan. Innumerable translations of *Marxism*. And the laborers' light blue overalls. Wills of iron, *Marx's* iron laws. (Nagamatsu 1970: 475–76)

"Words words words words" (Moretti 1996: 134). This is what Moretti sees in *Ulysses* as a symbol of how "advertising and stream of consciousness pursue and implicate one another throughout" Joyce's novel (135). Here it is not just the words themselves but their quality and presentation that are noteworthy: they are marked as foreign, they are repeated, they stand on their own as sentences. Neologisms, a high type-token ratio, and noun-ending sentences are all indicative of SOC in Japan, and they appear in their highest concentrations in Nagamatsu's works and in translations of *Ulysses*. One might read this as the symptom of a deeper formal affinity between the two writers in terms of their understanding and deployment of SOC.

Whether they were deploying it to similar ends is a question our model cannot answer. On this score, Nagamatsu arguably has more in common with Woolf or Proust, since SOC in "Portrait of Mademoiselle Mako" functions as a style for "exceptional circumstances" (e.g., a lover's panic and increasing delirium) and not for "absolute normality," as it does in *Ulysses* (Moretti 1996: 174). Nagamatsu himself insists on his aesthetic autonomy. He writes that "Portrait of Mademoiselle Mako" should not be considered a complete example of SOC (Nagamatsu 1931: 264), nor should Joyce's technique for psychological realism be thought of as "an absolute ideal which we should all imitate. On the contrary, all technique is style, and to the extent that style belongs to each individual author, if our style should tend a little bit in that direction, that means nothing more than that our worldview and sensibilities have also turned in that direction" (262). Influence, in other words, is merely a shifting frame of mind. But our model points to a yet deeper layer where influence operates—a layer where the desire for stylistic affinity and sameness expresses itself as a quantitative sensibility: how much or how little a specific formal feature should be included. At this level Nagamatsu's sensibility is more in sync with *Ulysses* than he might admit, at least in its Japanese translation, as he falls closer to it on the SOC spectrum than any of his contemporaries. Computation can reveal how literary grammars were similarly weakened and deformed by world writers facing urban capitalism's onslaught of words, but it can also order these changes by their patterns of variance within this shared structure of influence.

Our model has shown how sharply SOC texts differed from earlier fiction and made visible their degree of internal variation. But it is possible that the gap separating them from naturalist works merely reflects a historical shift in narrative and rhetorical style. To control for this possibility, we repeated our classification tests with a corpus of fiction contemporaneous with the SOC boom. SOC passages were only slightly more difficult to classify against this corpus. The accuracy dropped to 93 percent, meaning that the SOC passages were still very distinct based on the features in our model. Most of these features were predictive in the same way, and to the same degree, although with some variation. Type-token ratio, onomatopoeia, neologisms, and noun-ending sentences still weigh heavily toward SOC passages, although neologisms

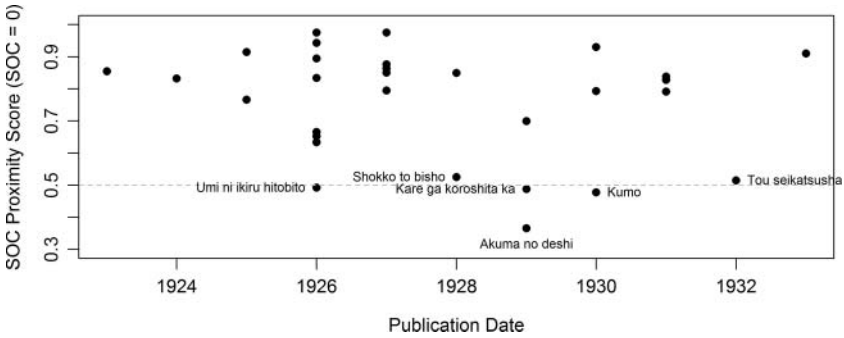

Figure 2. Proximity of control corpus titles to SOC. Each point represents the average predicted score for all passages sampled from that title. Titles closest to zero are those with the most SOC-like passages.

much less so than in the realist texts. Free indirect discourse is no longer a significant predictor of SOC on its own, although a personal pronoun at the head of a sentence is, if by a small factor. These variations hint at underlying differences between the earlier realist titles and the contemporary popular fiction. The latter texts used neologisms and free indirect discourse to a greater degree but tended to put first-person pronouns at the beginning of sentences less often. Despite these variations, the continued distinctiveness of the SOC passages rules out the possibility that a general linguistic drift in Japanese fiction is the only causal factor.

That said, the slight decrease in classification accuracy warrants further analysis. Of the thirty titles in this control corpus, six lie on or below the halfway mark separating SOC from non-SOC passages (fig. 2). Half of these are works of detective fiction (*Kare ga koroshita ka*, *Akuma no deshi*, and *Kumo*), and the other half are typically categorized as proletarian literature. Originally we chose to include a larger proportion of detective fiction in our control corpus based on the hunch that there would be some interesting stylistic overlap. Japanese detective novelists in the 1920s were fascinated with Western psychological theory, especially that of Freud (Kawana 2008: 29–68; Saito 2012: 235–76). It has been suggested that at the end of the decade there emerged among detective novelists and elite writers an understanding of the text "as a site where sensational 'surfaces' as expressions of the unconscious are

posited as standing in for the subjective core of our being" (Saito 2012: 274). Yet, although 80 percent of the control corpus was detective fiction, the fact that only three novels are proximate to SOC indicates a negative correlation with the genre. This may be because many of the detective novels were published just as the SOC wave was gaining momentum, and thus the genre could not draw on the many translations and adaptations being produced. More speculatively, the results may hint at the same generic resistance observed in the anglophone case. The most striking result is that all three proletarian works included in the control corpus fall near the SOC threshold. The limited sample size means that the association may be coincidental. Nevertheless, it is provocative, considering the sharp ideological divide known to separate modernist and leftist writers in Japan at this time. A closer inspection of these works may reveal that the stylistic tics chosen to narrate emergent class consciousness have something in common with those selected by SOC writers and translators to narrate the workings of the unconscious mind.

We began this section wanting to know how much language intervened in the diffusion of literary form. We found that it threw up obstacles in some places and not others and that these patterns of interference internally varied. In this last step, almost by accident, we situated this variance in a larger field of domestic and imported literary styles and thus positioned SOC within a spectrum of Japanese aesthetic responses to the psychological conditions of the modern subject. Its location in this spectrum highlights where our model captures the impact of shifting grammars of literary subjectivity across languages and genres. At the same time, it demarcates the edges beyond which our model fails and where the construction of new models more responsive to local literary and linguistic contexts will be necessary.[17]

Turbulent Flow

In this article we have sought to construct a computational model of SOC style to study its diffusion as a world literary form. We find support for our

[17] For instance, there are likely to be stylistic features, whether homegrown or imported through different channels, that our model fails to capture because it relies entirely on the original anglophone model. One goal of future work will be to identify such features and study how they alter the model's performance.

initial hypothesis that SOC followed a wavelike pattern of dispersion from the world literary system's core to its semiperiphery and periphery. Yet, at each stage of our analysis, our model charts the broad contours of this diffusion while exposing how this diffusion is marked by constant, heterogeneous variance. We do not see a single, monolithic pattern of diffusion but *patterns* of dissemination. In other words, we find patterns of difference (or variation) in sameness. This is an idea that is not extrinsic to computational or statistical methods but is deeply embedded in them. Indeed, among humanists, a common misunderstanding of modern statistical modeling is that quantitative models seek to explain everything about a social phenomenon and leave no room for interpretive ambiguity or indeterminacy. The opposite is true. A key feature of every statistical model is an error term that captures precisely what the model cannot explain. Moreover, a common reflexive technique in modeling is to estimate a model's own inability to fully measure the underlying processes that generate a data set. Modeling is thus deeply invested in indeterminacy, whether of itself or of the data to which it is applied.

To capture this notion of "indeterminacy within structure" that computation reveals, we posit a new conceptual term for the study of world literature: *turbulent flow*. It is a term we take loosely from physics, where turbulence describes a process in which an otherwise linear and coherent flow of objects or things becomes radically unstable, heterogeneous, and chaotic. We find this concept useful because it takes as its assumption the existence of a general regime of movement—for us, the spread of literary forms—that breaks apart as it progresses. The stimuli that drive this dissipation can be internal or external. The sheer speed and distance by which the form travels may cause it to disintegrate, for example, or it may be subject to various external pressures present in the form's new host country. Whatever the cause, turbulence is inevitable in the movement of literary forms as they circle the world, and this leads to their inevitable fracturing. But we cannot comprehend this turbulence without assuming some underlying measurable structure. Structure and variance, that is, go hand in hand. To understand the ways that SOC breaks apart into infinite other forms, we must measure and understand its underlying structure of diffusion. Conversely, to understand how the diffusion of SOC operates like a wave, we must read closely the individual

instances of transformation wherein SOC diverges or is rerouted from its overall directional flow. Computation allows for both, offering a vision of world literature as simultaneously structure and variation, flow and turbulence.

Hoyt Long is associate professor of Japanese literature at the University of Chicago. He is author of *On Uneven Ground: Miyazawa Kenji and the Making of Place in Modern Japan* (2012) and author, with Richard Jean So, of "Literary Pattern Recognition: Modernism between Close Reading and Machine Learning," in the Winter 2016 issue of *Critical Inquiry*. He codirects the Chicago Text Lab with So.

Richard Jean So is assistant professor of English at the University of Chicago. His book *Transpacific Community: America, China, and the Rise and Fall of a Cultural Network* is forthcoming, and he is also author, with Hoyt Long, of "Literary Pattern Recognition: Modernism between Close Reading and Machine Learning," in the Winter 2016 issue of *Critical Inquiry*. He codirects the Chicago Text Lab with Long.

References

Bourdieu, Pierre. 1996. *The Rules of Art: Genesis and Structure of the Literary Field*, translated by Susan Emanuel. Stanford, CA: Stanford University Press.

Bowling, Lawrence. 1950. "What Is the Stream of Consciousness Technique?" *PMLA* 65, no. 4: 333–45.

Brodey, Inger Sigrun. 1998. "Natsume Sōseki and Laurence Sterne: Cross-Cultural Discourse on Literary Linearity." *Comparative Literature* 50, no. 3: 193–219.

Casanova, Pascale. 2004. *The World Republic of Letters*, translated by M. B. Debevoise. Cambridge, MA: Harvard University Press.

Chatman, Seymour. 1978. *Story and Discourse: Narrative Structure in Fiction and Film*. Ithaca, NY: Cornell University Press.

Cohn, Dorrit. 1978. *Transparent Minds: Narrative Modes for Presenting Consciousness in Fiction*. Princeton, NJ: Princeton University Press.

Dahl, Liisa. 1970. *Linguistic Features of the Stream-of-Consciousness Techniques of James Joyce, Virginia Woolf, and Eugene O'Neill*. Turku: Turun Yliopisto.

Denning, Michael. 1997. *The Cultural Front*. London: Verso.

Doi Kōchi. 1929. "Joisu no yurishizu" ("Joyce's *Ulysses*"). *Kaizō* (*Reconstruction*), February, 24–47.

Edel, Leon. 1964. *The Modern Psychological Novel*. New York: Dunlap.

Farnol, Jeffery. 1933. *The Way Beyond*. Boston: Little, Brown.

Foley, Barbara. 1993. *Radical Representations: Politics and Form in U.S. Proletarian Fiction, 1929–1941*. Durham, NC: Duke University Press.

Friedman, Melvin. 1955. *Stream of Consciousness: A Study in Literary Method*. New Haven, CT: Yale University Press.

Gruesz, Kirsten Silva. 2002. *Ambassadors of Culture: The Transamerican Origins of Latino Writing*. Princeton, NJ: Princeton University Press.

Hojō Fumio. 1980. "Jeimuzu Joisu to Nihon kindai shōsetsu (I)" ("James Joyce and the Modern Japanese Novel (I)"). *Publications of the Institute for Comparative Studies of Culture affiliated to Tokyo Woman's Christian College* 41: 35–53.

Humphrey, Robert. 1954. *Stream of Consciousness in the Modern Novel*. Berkeley: University of California Press.

Kawana, Sari. 2008. *Murder Most Modern: Detective Fiction and Japanese Culture*. Minneapolis: University of Minnesota Press.

Moretti, Franco. 1996. *Modern Epic: The World-System from Goethe to García Márquez*, translated by Quintin Hoare. London: Verso.

———. 2000. "Conjectures on World Literature." *New Left Review*, no. 1: 54–68.

———. 2011. "World-Systems Analysis, Evolutionary Theory, *Weltliteratur*." In *Immanuel Wallerstein and the Problem of the World: System, Scale, Culture*, edited by David Palumbo-Liu, Nirvana Tanoukhi, and Bruce Robbins, 67–77. Durham, NC: Duke University Press.

Nagamatsu Sadamu. 1931. "Nihon ni okeru 'ishiki no nagare' shōsetsu" ("'Stream-of-Consciousness' Novels in Japan"). *Shinbungaku kenkyū* (*New Literature Studies*), April, 260–64.

———. 1970. *Nagamatsu Sadamu sakuhin shū* (*Selected Works of Nagamatsu Sadamu*). Tokyo: Gogatsu Shobō.

Ōta Saburō. 1955. "Joisu no shōkai to eikyō" ("The Introduction and Influence of Joyce"). *Gakuen* (*Campus*) 175, no. 4: 203–25.

Piper, Andrew. 2015. "Novel Devotions: Conversional Reading, Computational Modeling, and the Modern Novel." *New Literary History* 46, no. 1: 63–98.

Power, Arthur. 1974. *Conversations with James Joyce*. London: Millington.

Saito, Satoru. 2012. *Detective Fiction and the Rise of the Japanese Novel, 1880–1930*. Cambridge, MA: Harvard University Asia Center.

Scholes, Robert, and Robert Kellogg. 1966. *The Nature of Narrative*. Oxford: Oxford University Press.

Sharif, Mebed. 2003. "Shōwa shoki ni okeru 'ishiki no nagare' juyō wo megutte" ("On the Reception of 'Stream-of-Consciousness' in Early Shōwa"). *Issues in Language and Culture*, no. 4: 5–16.

Steinberg, Erwin R. 1973. *The Stream of Consciousness and Beyond in "Ulysses."* Pittsburgh, PA: University of Pittsburgh Press.