CA Journal of Cultural Analytics

Self-Repetition and East Asian Literary Modernity, 1900-1930

Hoyt Long, Anatoly Detwyler, and Yuancheng Zhu

05.21.18

Peer-Reviewed By: Tomi Suzuki

Clusters: Genre

Article DOI: 10.22148/16.022

Dataverse DOI: 10.7910/DVN/DAKZHR

Journal ISSN: 2371-4549

Cite: Hoyt Long, Anatoly Detwyler, and Yuancheng Zhu, "Self-Repetition and East Asian Literary Modernity, 1900-1930," Cultural Analytics May 21, 2018. DOI: 10.22148/16.022

Histories of East Asian literary modernity have often begun as historiographies of the narrative self. For some scholars, the emergence of a decidedly self-referential mode of fiction in the early twentieth century is part and parcel of what defines this modernity.[1] In Japan there was the "I-novel"; and in China, Romantic fiction. The two are recognized as foundational genres that distinguished themselves from prior fiction by the adoption of a narrow autobiographical focus, ex-

[1] On the China side, see Robert Hegel and Richard Hessney, eds., *Expressions of Self in Chinese Literature* (New York: Columbia University Press, 1985), in particular the contribution of Leo Oufan Lee, "The Solitary Traveler: Images of the Self in Modern Chinese Literature" (282-307); Jaroslav Průšek, *The Lyrical and the Epic: Studies of Modern Chinese Literature* (Bloomington: Indiana University Press, 1980); and Lydia Liu, *Translingual Practice: Literature, National Culture, and Translated Modernity—China, 1900-1937* (Stanford: Stanford University Press, 1995). For Japan, see Karatani Kōjin, *Origins of Modern Japanese Literature*, ed. Brett de Bary (Durham, NC: Duke University Press, 1993); James Fujii, *Complicit Fictions: The Subject in the Modern Japanese Prose Narrative* (Berkeley: University of California Press, 1993); and Janet Walker, *The Japanese Novel of the Meiji Period and the Ideal of Individualism* (Princeton, NJ: Princeton University Press, 1979).

tended psycho-narration, and a new vernacular writing style. At the same time, others have struggled to define these genres in more precise stylistic or formal terms. Edward Fowler once said of the I-novel that "writing about [it] is not unlike pursuing a desert oasis…How is one to analyze a form that critics have debated for well over half a century but for which they have failed to come up with a workable definition?"[2] The case is similar for China's Romantic fiction, which, since the work of C.T. Hsia and Leo Ou-fan Lee, has been conventionally defined by its social milieu rather than through a coherent set of generic qualities.[3]

Definitional ambiguity is integral to how literary scholars understand genre: the identity of any one text will always be overdetermined. Equally important is the conceit that groups of texts can cohere in ways that differentiate them from others. In this paper we use computational methods to argue that a heightened tendency toward lexical repetition was a significant point of coherence for the narrative practices now captured under the signs of "I-novel" and Romantic literature. By tendency we mean something less than an essential trait found in all self-referential works but more than a minor feature found in only a few. The presence of this tendency in both cultural contexts prompts us to think about the role of repetition in literary style, but also of repetition *as* literary style. On the one hand, we suggest that repetition indexes specific formal transformations that I-novels and Romantic literature are commonly associated with: the vernacularization of writing and the adoption of Western grammatical structures. On the other, we propose that repetition also relates to changes at the level of content, in particular an emphasis on narratives of psychological realism and mental aberration. In this way, repetition as style becomes a way to identify, as a kind of surface phenomenon, a deeper and more complex set of interactions taking place between intellectual figurations of self and concrete linguistic strategies. Examining this surface through a computational lens, we propose, opens up a new comparative framework for analyzing the effects of these interactions across the space of East Asian literary modernity.

Our argument is divided into three sections. In the first we establish our rationale for associating repetitiveness with a set of qualitative traits that scholars have previously ascribed to Japanese I-novels and Chinese Romantic fiction. After collecting a set of measurable linguistic features that capture various kinds of repe-

[2] Edward Fowler, *The Rhetoric of Confession: Shishōsetsu in Early Twentieth-Century Japanese Fiction* (Berkeley: University of California Press, 1988), 3.

[3] Hsia suggests that, beyond the work of a handful of representative individuals, Romanticism's only distinguishing quality is a "maudlin sentimentality. . . completely deficient in restraint and objectivity." See *A History of Modern Chinese Fiction*, Second Ed. (New Haven and London: Yale University Press, 1971), 95. Likewise, Lee concludes that Romanticism was definable largely by way of group libraries and clashes of personalities. See *The Romantic Generation of Modern Chinese Writers* (Cambridge, MA: Harvard University Press, 1973), 22.

tition in language, we test the extent to which these features are characteristic of these genres as compared with contemporaneous works of fiction. In the second section, we discuss our empirical findings in light of past scholarly treatments of repetition in aesthetic, socio-linguistic, and psychological terms, where it has been recognized as fundamental to the construction of meaning. By assessing how literary critics and linguists have tried to model repetition in language and the advantages of doing so qualitatively versus quantitatively, we historicize our quantitative model and show how it is already bound up in previous efforts to read in the surface of language the symptoms of abnormal mental processes. In the third section we turn to several of the most repetitive passages identified in our analysis to consider how repetition as style can be read from textual surfaces. We assert that it can be read in multiple ways: as stylistic tendency that bridges literary relations across cultural and linguistic borders; as a tendency activated by writers to varying aesthetic ends; and finally as a supplement to comparative frameworks based on semantic meaning or ideology.

Repetition as Tendency

The modern project of literary self-fashioning began in earnest in Japan after the turn of the century and flourished in the 1910s. Much of this writing, retroactively collected under the label of "I-novel" (*shishōsetsu*), transformed the representational logic of Naturalism into an obsession with documenting the self's inner thoughts and daily experiences no matter how shocking or mundane. During this "age of confession," as one Japanese critic called it in 1909, many of China's May Fourth generation of writers were living in Japan as students.[4] In 1921, a group of them formed the Creation Society back home, a literary collective now closely identified with a romanticist interest in examining and exploring personal subjectivity. Together, these I-novel and Romantic writers produced a diverse collection of self-referential writing that is central to histories of modern Japanese and Chinese fiction.

Canons engender questions of coherence, however, and these groupings are no different. Decades of scholarship devoted to these writers suggests that there is little that singularly defines their fiction. Critics have disputed the coherence of these texts by isolating numerous ideological currents that run through them, by questioning their temporal cohesion, and by problematizing their status as nar-

[4] Shimamura Hōgetsu, "Jo ni kaete jinseikanjō no shizenshugi o ronzu" [By Way of a Preface: On Naturalism and my *Weltanschauung*]. Cited in Fowler, 100.

rative fiction.[5] Ambiguity as to whether the I-novel and Romantic fiction are meaningful generic labels has even led to extreme relativist claims that deny the existence of a coherent form or genre at all; these labels, so the argument goes, are mere discursive and ideological paradigms through which any text can be read.[6] Some scholars, while not denying there is variation within this literature, have proceeded from the opposite assumption, treating the I-novel and Romantic fiction as genres bound by distinct formal or empirical patterns. Focusing on narrative structure, rhetorical style, or social and media context, they try to isolate a set of features that hold these texts together.[7]

Our aim in this paper is not to rule over this long-running genre debate, a foreclosure that would in any case be antithetical to our role as literary critics. There is no single way to resolve such a debate because every attempt to argue for or against the ontological reality of these genre labels is predicated on different assumptions about the unit(s) of comparison. Is it the author? The ideal reader? Some aspect of the text? Here, we explicitly focus on shared linguistic patterns. They afford a scale of comparison that can encompass many hundreds of texts and multiple linguistic contexts. But they also provide a level of granularity through which we can potentially observe the stylistic tendencies that came together to instantiate the modern self as a literary construct. Or, to borrow from Franco Moretti's analysis of *bourgeois* style, as a "mentality" made of "unconscious grammatical patterns and semantic associations, more than clear and distinct ideas."[8] The first question we needed to answer was whether any such mentalities existed in the body of fiction grouped under the "I-novel" and "Romantic" labels.

As mentioned at the outset, there are several higher order phenomena that characterize this body of fiction. Scholars have long noted that its rise is intimately tied up with consolidation of the modern written vernacular under the *genbun-itchi* and *baihua* movements in Japan and China, respectively. Others point to the widespread experimentation with imported narrative techniques and Euro-

[5] For a thorough review of this criticism, particularly the contributions of Ito Sei, Hirano Ken, and Kobayashi Hideo, see Fowler, chapter 3; and Irmela Hijiya-Kirschnereit, *Rituals of Self-Revelation: Shishōsetsu as Literary Genre and Socio-cultural Phenomenon* (Cambridge, MA: Council on East Asian Studies, Harvard University, 1996), chapter 9.

[6] See Tomi Suzuki, *Narrating the Self: Fictions of Japanese Modernity* (Stanford, CA: Stanford University Press, 1996), 5-6.

[7] For China, see Edward Gunn, *Rewriting Chinese: Style and Innovation in Twentieth-Century Chinese Prose* (Stanford: SUP, 1991); Liu, *Translingual Practice*; Haiyan Lee, *Revolution of the Heart: A Genealogy of Love in China, 1900-1950* (Stanford: Stanford University Press, 2007); and Raymond Hsu, *The Style of Lu Hsun: Vocabulary and Usage* (Hong Kong: Centre of Asian Studies, University of Hong Kong Press, 1979). For Japan, see Fowler, *Rhetoric of Confession*; Hijiya-Kerschnereit, *Rituals of Self-Revelation*; and Barbara Mito Reed, "Language, Narrative Structure, and the *Shōsetsu*" (diss. Princeton University, 1988).

[8] Moretti, *The Bourgeois: Between History and Literature* (London: Verso, 2013), 19.

peanized grammar that happened in conjunction with vernacularization, but which was necessarily distinct from it.[9] On the one hand, these imports include things like free indirect discourse, lengthy interior monologue, and a rejection of emplotment.[10] On the other, they include use of personal pronouns, inanimate subjects, Western syntax, and an exaggerated specification of subject/object relations. Indeed, much attention has been given to how Japanese and Chinese, as non-inflected languages that traditionally allow for great flexibility in whether or not to specify the grammatical subject, were simultaneously leveraged and deformed in the creation of new structures of self-narration. In the Japanese case, it has been argued that this flexibility allows for the slippages between narratorial authority and character viewpoint that blur the I-novel's status as realist fiction.[11]

While these complex developments in literary language provide an important foundation for understanding what was unique about self-referential fiction, they do not scale terribly well as features nor do they necessarily separate out such fiction from other contemporary genres that similarly adopted a vernacular style or Western grammatical structures. Our goal was thus to find a set of quantitative measures that allowed us to compare hundreds of texts while potentially singling out linguistic tendencies that indexed these higher order phenomena in self-referential fiction. In practice, this meant creating effective proxies for these phenomena that captured some aspect of their impact on literary language. From the perspective of plot and narrative, we reasoned that the more intense psychological focus of these texts might lend itself to a narrowing of the semantic field and less lexical diversity as compared with plot driven works and their more dynamic narrative focus. In other words, did I-novels and Romantic fiction tend to concentrate their lexical attention on a smaller vocabulary? And, from the perspective of style, we hypothesized that one result of the shift to vernacular writing might be increased repetition or redundancy in the language. The adoption of

[9] For China, see Liu and Gunn. For Japan, see Kisaka Motoi, *Kindai bunshō seiritsu no shosō* [Various Aspects of the Formation of Modern Style] (Osaka: Wazumi shoin, 1988), Chapter 3. On the distinction between a vernacular style and shifts in the conceptual and grammatical structure of written Japanese, see also Karatani, 49-51. The two are typically seen as distinct, but interrelated movements in the development of the new literary language known as *genbun itchi*.

[10] With regard to emplotment, I-novels, for instance, have been described as "tedious" descriptions of "one's life and nothing else" (Yasuoka Shōtarō, 25); "fragmented and short-winded" (Yokomitsu Ri'ichi, 52) or otherwise "random" accounts of personal experience (Kume Masao, 46); a "medium for intimate expression that would suffer from too much attention to structure" (Ito Sei, 63); and "a string of impressionistic musings" (Uno Koji, 7). All cited in Fowler. On the China side, Yu Dafu's works have been singled out for emphasizing journeys that are "incomplete, aimless, and marked with uncertainties." Cited in Liu, 149. And Guo Moruo famously responded to early criticism of one of his works by saying "that it was a mistake to read his story as a straightforward narrative with a beginning, a climax, and an ending—he was trying to present the unconscious in the form of dream symbolism." Cited in Liu, 131.

[11] See Reed, 144-169; Fowler, Chapter 2; and Liu, 153-54.

Western grammatical features, in particular the tendency to render subject and object explicit in every sentence, would likely only further exacerbate this trend.

While some of these hypotheses were merely informed hunches, our hypothesis about vernacular writing has a long history in how orality has been understood in relation to the written word. If we understand redundancy to be the repetition of certain units of language (e.g., letters, phonemes, morphemes) either because they are contextually dependent on one another or because they enhance the reliability of a message, then all natural languages are inherently redundant.[12] They are built on rules and conventions that allow us to predict, for example, the word that follows another word or sequence of words, and consequently to leave out words that are implied by context. Many have argued that this built-in redundancy of language is even more extreme in spoken language, and oral culture in general. Building upon Milman Parry's study of contemporary Yugoslavian oral epics, Walter Ong has argued that formulaic expressions and repetition are an aid to memory in oral cultures and that in oral discourse "the mind must move ahead more slowly, keeping close to the focus of attention much of what it has already dealt with. Redundancy, repetition of the just-said, keeps both speaker and hearer surely on the track."[13] Linguists who study conversation suggest that "repetition is at the heart not only of how a particular discourse is created [between speakers], but how discourse itself is created"; this notion has been taken up by literary scholars too in the interest of identifying the linguistic markers of colloquial style in Western contexts.[14] To what extent, we wondered, does this repetitive quality of orality manifest itself in the new vernacular styles of Japanese and Chinese literature?

Fortunately, this longstanding interest in repetition by linguists has yielded a wide array of quantitative measures for capturing aspects of redundancy and lexical diversity. Many of these measures, especially those where the word is the primary unit of analysis, share a common origin in the field of psycholinguistics as it was practiced in America and Europe between the 1930s and the 1950s, a period marked by a broad interest in developing measures of lexical diversity for use in educational or clinical assessment. Researchers wanted to know, given a

[12] In fact, some have argued that the level of redundancy may even be stable across languages. See Marcelo A. Montemurro and Damián H. Zanette,"Universal Entropy of Word Ordering Across Linguistic Families." In *PLoS ONE* 6(5): e19875.

[13] Walter Ong, *Orality and Literacy: The Technologizing of the Word* [1982] (1991): 35-40.

[14] Deborah Tannen, *Talking Voices: Repetition, Dialogue, and Imagery in Conversational Discourse* (Cambridge: Cambridge University Press, 2007), 49. For a recent survey of work on repetition and colloquial style in literature, as well as a superb attempt to scale up this work quantitatively, see Marissa Gemma, Frédéric Glorieuz, and Jean-Gabriel Ganascia, "Operationalizing the Colloquial Style: Repetition in 19th-Century American Fiction," in *Digital Scholarship in the Humanities* 2015 fqv066 (doi: 10.1093/llc/fqv066).

particular sample of writing or speech, were more of the same words being repeated at a higher frequency or were many different words being used with less frequency? In 1935, George Zipf developed his eponymous law stating that the distribution of word frequency ranks in a given sample of natural language obey a power law, such that the frequency of any word is inversely proportional to its rank in the frequency table (i.e., the most frequent word will occur twice as many times as the second most frequent word, and so on). In 1938, John B. Carroll developed a diversity measure based on the observation that the growth of word diversity with text size must approach a limit. His measure focused on how often frequent words tended to be repeated in a passage, and he asserted that measures like this could help assess the relative adherence of one's verbal behavior to linguistic norms.[15] The next year, Wendell Johnson introduced the notion of type-token ratio (TTR): the number of unique words in a text divided by the total number of words. He suspected the ratio might serve as "a measure of degree of frustration, or of disorientation," and that it could serve to quantify the phenomena of "one-track mind," or "monomania."[16] The 1940s saw further attempts to build on these foundational measures in order to assess how repetitive, uniform, or concentrated was the vocabulary of a given segment of text. Some of these measures, to be described shortly, had the advantage of being less sensitive to variation in text length and of being able to dampen or ignore the influence of rare words.

Significantly, they also shared a mathematical relation to a measure that grew to become highly influential in the 1950s and represented a different approach to the problem of repetition: entropy. Following the work of Claude Shannon and Warren Weaver at Bell Labs, a number of psycholinguists began to approach repetition in more probabilistic ways, analyzing not just the diversity of words used but the predictability of their sequential order, or what were referred to as "transitional probabilities." They refocused ideas about repetition through the twin lenses of redundancy and information. In an information theoretic context, the amount of redundancy in a message (its entropy) reflects the amount of "information" in that message. Here, information means the likelihood of a message based on all the units available to constitute it, but also all the ways of combining these units given existing rules or patterns governing their arrangement. In short, information expresses how many different ways a message can be constructed given these initial constraints. An extremely information rich language, then, might be one where any given word is equally likely to appear next to any other. Every message in this artificial language would carry new information because

[15] John Carroll, "Analysis of Verbal Behavior," in *Psychological Review* 51 (March, 1944): 102-119.

[16] Wendell Johnson, *Language and Speech Hygiene: An Application of General Semantics, Outline of a Course* (Chicago: Chicago Institute of General Semantics, 1939), 11.

each would be as random and unpredictable as the one before it. The messages would also be wholly unintelligible, which is why all natural languages have some redundancy built into them.

While entropy proved a theoretically productive concept for many psycholinguists, it also proved very tricky to measure in any holistic way. Not only does it vary with the length of the text being measured, it also varies with the unit of analysis and the length of the sequence being considered. As a sequence grows longer, so too does the number of possible combinations with which to predict the randomness of the next item in this sequence. Entropy is thus biased by how much of a text or corpus is available to be measured and is also increasingly intractable as the number of units and their possible combinations increases. In practice, this meant that early applications of entropy to text were confined to smaller units of analysis (e.g., letters, syllables), because one could expect to see the fuller range of possible combinations in a given portion of text.[17] It also meant that focus remained on individual words or word pairs, such as in Gustav Herdan's use of entropy to reason about how writers manipulated the variability of expression in their writing to avoid undue repetition.[18] When confined to the individual word level, entropy simply captures the spread of the total words of a sample amongst the different words available in that sample. The highest entropy passage in this case is one where every word is unique and different; the lowest is a passage where every word is the same, and thus highly redundant.[19]

Despite the limitations of these various measures of lexical diversity and entropy, they do provide a baseline for quantifying the amount of repetition in a text. Using this baseline, we first determined whether I-novels or Romantic fiction show an exaggerated tendency to repeat as compared with other fiction written contemporaneously. Does the combination of a vernacular style, Western grammatical structures, and psychological focus translate into a narrower range of words being repeated more often? To answer this question, we first constructed corpora for each language. For Japan, we collected roughly 65 texts that scholars have specifically designated or read as belonging to the I-novel genre. We also included self-referential or psychological works by authors associated with the genre or by authors who briefly experimented with this mode of writing. The bulk of the works were published in the teens and twenties and represent about

[17] See, for example, the work of Wilhem Fucks who, in 1952, tried applying information theory to stylometrics and compared the entropy of syllables in prose versus poetry. "On the Mathematical Analysis of Style," in *Biometrika* 39, no. 9 (1952): 122-129.

[18] Gustav Herdan, *Language as Choice and Chance* (Groningen: P. Noordhoff, 1956), 167.

[19] For additional critiques of entropy as a valid measure of lexical richness or style, see, for example, P. Thoiron, "Diversity Index and Entropy as Measures of Lexical Richness," in *Computers and the Humanities* 20, no. 3 (1986): 197-202; and David Hoover, "Another Perspective on Vocabulary Richness," in *Computers and the Humanities*, 37, no. 2 (2003): 151-178.

30 authors. Next we assembled a popular corpus of similar size that we expected to diverge sharply at the level of content and narrative focus but not at the level of literary language. It mostly consists of highly emplotted historical and detective fiction from the 1920s and 30s written in the modern vernacular style.[20]

For China we adopted a slightly different approach due to the lack of an equivalent corpus of popular genre fiction. First, we identified over 100 Romantic texts by key May Fourth writers associated primarily with the Creation Society (*Chuangzao she*), including 1920s works from Yu Dafu, Guo Moruo, and Zhang Ziping. Our control group, however, was a set of 100 works of contemporary popular literature such as historical fiction and "Mandarin Duck and Butterfly" stories.[21] While these were chosen for their highly emplotted quality and lack of psychological focus, as in the Japanese case, most were also written in an older

[20] The I-novel corpus was created via secondary English and Japanese sources, including Fowler; Hasegawa Izumi, "Meijiā • Taishōā • Shōwa shishōsetsu sanjūgo sen" [A Selection of 35 I-Novels from Meiji, Taishō, and Shōwa], in *Kokubungaku: kaishaku to kanshō* 27, no. 14 (1962); *Wataskushi shōsetsu handobukku* [The I-Novel Handbook], Akiyama Shun and Katsumata Hiroshi, eds. (Bensei shuppan, 2014). Additional texts by authors associated with the I-novel were identified using the *Nihon kindai bungaku daijiten* [Encyclopedia of Modern Japanese Literature] and selected based on their degree of autobiographical content. Finally, several texts were included that are hallmarks of Naturalist style (e.g., Tokuda Shūsei's *Arakure*, Arishima Takeo's *Aru onna*) but not recognized as I-novels. Texts were acquired via Aozora Bunko (the Japanese equivalent of Project Gutenberg) or digitized on our own. The popular corpus was built from the Aozora Bunko archive and contains titles by genre authors like Unno Jūza, Kōga Saburō, Yoshikawa Eiji, Nakazato Kaizan, and Nomura Kodō. A complete list of corpus titles and associated metadata can be found in the Dataverse companion to this article. It should be noted that the need to maintain parity with the Chinese case restricted the kinds of comparative corpora we could use for this experiment. In the future, it will be important to compare the I-novels with a corpus of pure realist fiction from the same period. A corpus of proletarian fiction was also created for this project, but had to be bracketed out to simplify the analysis.

[21] The Romantic corpus takes as its core the texts and authors named by Zheng Boqi in his introduction to the volume on Creation Society literature in the seminal *Zhongguo Xinwenxue daxi* Vol. 5, ed. Zheng Boqi (Shanghai: Liangyou tushu yinshua gongsi, 1981). From this canonical collection, we largely focused on pre-1925 works so as to avoid mixing in the far more political and mass-inflected works that Guo Moruo began promoting after the May Thirtieth Incident. The core of the control corpus is based upon the titles listed in the seminal work on "Mandarin Ducks and Butterfly Literature," Wei Shaochang, ed., *Yuanyang Hudie pai yanjiu ziliao* Vol. 2 (Shanghai: Shanghai wenyi chubanshe, 1962). That said, many of the texts may not be strictly "Mandarin Ducks and Butterfly" works, but rather popular (and commercially successful) works of "historical fiction" in the vein of *Romance of the Three Kingdoms*. Originally our project was aimed at a triangular comparison between Romantic, popular, and socialist realist fiction from the 1930s. The latter corpus, however, proved difficult to distinguish from Romanticism along the line of repetition. This is partly due to the fact that the literary style of socialist realism of the 1930s was heavily influenced by stylistic developments of the May Fourth period. Wishing to avoid the questions of influence that a diachronic comparison raises, we bracketed out the socialist realist corpus. Future projects focusing on the interplay of Chinese genres will include this socialist realist corpus, but also the work of Lu Xun (whose early fiction was contemporary to Romanticism) and the self-obsessed fiction of the so-called Neo-Perceptionists (*Xinganjuepai*) of the late 1920s and early 30s.

style of vernacular (*jiu baihua*) markedly different from the vernacular modes developed by Romantic writers. Thus the comparison in this case was carried out along the dimensions of content *and* linguistic style. These differences notwithstanding, our goal in both cases was to determine whether various measures of repetition and redundancy were sufficient to identify a generically distinctive tendency in I-novels and Romantic fiction that transcended the meaning of the words on the page.

The next step was thus to apply these measures. Because measures like TTR and entropy tend to be highly correlated with the length of the passage being measured, we applied them such that the results would be independent of text length. For these two especially, this meant dividing texts into 1,000 word segments; measuring the TTR and entropy for these segments, including stopwords; and then computing the average, standard deviation, and cumulative sum across all segments of a text (Equation 1).

Suppose that in the chunk of length n there are m distinct words, $w_1, \ldots, w_m$, each appearing for $n_1, \ldots, n_m$ times. Therefore, we should have $n_1 + \cdots + n_m = n$. Let

$$\hat{p}_i = n_i/n, \quad i = 1, \ldots, m$$

be the proportion of appearance for word w_i. Our sample entropy of the word is calculated by

$$\widehat{\text{entropy}} = -\sum_{i=1}^{m} \hat{p}_i \log \hat{p}_i.$$

Standard deviation tells us about the variance of TTR and entropy across all chunks, while cumulative sum tells us how much higher or lower than average the values tend to be. Aware that our entropy measure was tied to the marginal distribution of individual words, we also calculated entropy based on the joint distribution of words, taking their sequential nature into account. This method, borrowed from Ioannis Kontoyiannis, adopts a non-parametric approach that captures long-range dependencies between sequences of words or characters.[22] Here we chose to focus on sequences of individual phonetic and Chinese characters, such that lower entropy means more repetition of the same sequences of characters. While the window size for finding matching sequences is still dependent on the length of our shortest texts, biasing the resulting entropy estimates

[22] I. Kontoyiannis, "The Complexity and Entropy of Literary Styles," in *NSF Technical Report*, no. 97 (June 1996-October 1997): 1-15. In this case, it is non-parametric in the sense that it is not bound to the smaller contexts (unigrams, bigrams, etc.) of Markov-based entropy measures. Thus, for each position *i* in a text's sequence of units (in our case, individual characters), the method looks for the longest sequence starting at *i* that does not occur prior to *i*. For example, at *i* = 100, it will scan for the longest sequence of characters that does not occur in the previous 100 characters. These lengths for various *i* are then used to estimate the entropy of the text as a whole.

to some degree, the estimates themselves are not correlated with text length.

Concerned that TTR and entropy alone provided too narrow a window onto repetition, we implemented two additional features related to entropy mathematically but originally created as indexes of lexical diversity. The first is George Yule's "Characteristic K," developed in 1944 to measure the repetitiveness or uniformity of vocabulary in a text. It relies on word rank and frequency for its calculation, relating the sum of all word frequencies to the number of words with a particular frequency, and was designed by Yule to be independent of sample size.[23] It also assumes that word occurrence in a given sample of text follows a Poisson distribution, treating words as fixed events that occur with a known average rate for any interval (i.e., the length of the sample). Herdan later corrected for this assumption, developing a modified K that was widely adopted in the 1960s as a stylistic measure for the concentration of vocabulary, including attempts to analyze schizophrenic language.[24] Another feature we included is an index of lexical concentration developed, also in 1944, by French linguist Pierre Guiraud. "Guiraud's C," as it is known, expresses the proportion of a text's cumulative word frequency taken up by its most 50 frequent "content" words. A high value of the index implies that "an author concentrates his attention on a relatively narrow range of words with full meaning," which in turn testifies to "thematic compactness, to the concentration on the main theme, [and] in some cases also to stock phrases."[25] This measure is more sensitive to text length than Yule's K, and thus has less explanatory power, but its explanation is more intuitive. Both have the benefit of not requiring the splitting of texts into smaller chunks. And both, importantly, are akin to entropy in that they depend on the sums of relative word frequencies.[26]

Examining these measures individually, we find that nearly all are good at distinguishing I-novels and Romantic fiction from their popular contemporaries. The distributions of average TTR and entropy for the Japanese corpora indicate

[23] See George Yule, *The Statistical Study of Literary Vocabulary* [1944] (Hamden, CT: Archon Books, 1968). The measure is calculated as follows: 10,000 x (M_2 - M_1)/(M_1 x M_1). M_1 is the number of word tokens. M_2 is calculated by multiplying the number of words at a given rank frequency by the square of that rank (e.g., all words occurring 2 times multiplied by 2^2) and then summing over all of these values.

[24] Juhan Tuldava, "Stylistics, Author Identification," in *Quantitative Linguistics: An International Handbook*, ed. Reinhard Köhler, et. al (Berlin: Walter de Gruyter, 2005), 374. See also Arthur Holstein, "A Statistical Analysis of Schizophrenic Language," in *Statistical Methods in Linguistics* 4 (1965): 10:14.

[25] Tuldava, 375. Guiraud's C is derived by summing the frequencies of the top 50 most frequent words and dividing through by the total number of words.

[26] On the relation of Yule's K to entropy measures, see Kumiko Tanaka-Ishii and Shunsuke Aihara, "Computational Constancy Measures of Texts," in *Association for Computational Linguistics* 41, no. 3 (2015): 481-502.

that I-novels generally score lower on both counts, indicating less lexical diversity and more repetition. Overall, we found that most measures pointed toward greater repetitiveness in this mode of fiction and that, surprisingly, this tendency seemed to hold true across languages.[27] Rows correspond to the assigned genre labels and columns correspond to the predicted genre labels. In the Chinese case, the separation is similarly distinctive (Figure 1).

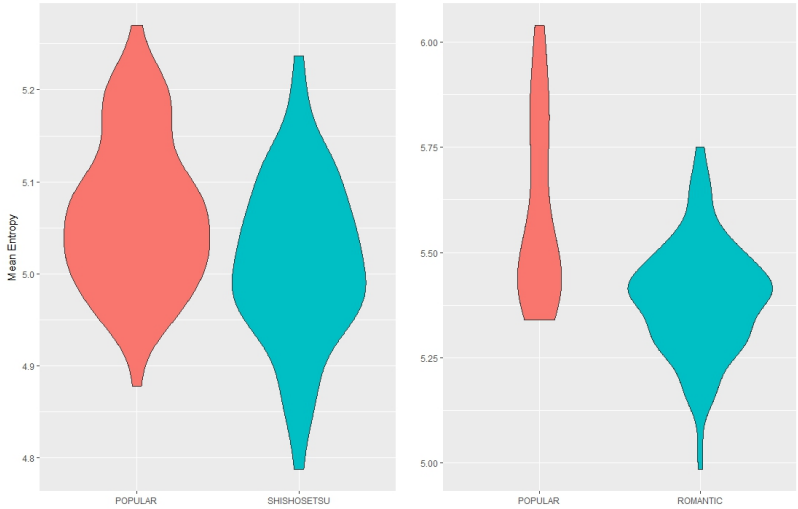


Figure 1. Violin plots representing the distribution of mean entropy by genre. Width indicates the relative proportion of texts in a genre that have a particular mean entropy. In the Japan case (left), we see a narrower band of I-novels with lower entropy than Popular works. For the China case (right), we see a much larger proportion of Romantic works with lower entropy than Popular works.

Yule's K and Guiraud's C reveal statistically significant differences in both cases as well, indicating a tendency toward lexical uniformity and compactness in self-referential fiction.[28] Interestingly, the self-referential fiction also tends to have more extreme fluctuations in repetitiveness, as indicated by their higher standard deviation of TTR and entropy. These texts are more repetitive on average, but they also exhibit more drastic shifts between less repetitive passages and more

[27] We used a pairwise t-test with a Bonferroni correction to determine significance between the distributions of each feature. Significance indicates that the mean value of each feature is not equal in the two samples being compared. Significance was assessed at the p <= .05 level.

[28] These two measures were less reliable in the Chinese case in that both were correlated more heavily with length. This likely has to do with the greater variance in text length in the Chinese corpus, which includes some very short and some exceptionally long texts.

repetitive ones. A measure that did not show significant difference across categories was Kontoyiannis's entropy measure, suggesting that no group of texts had significantly more long-range dependencies than the other. It did, however, when analyzed in combination with other features, help to identify some self-referential texts that were repetitive in ways our word-based measures could not capture, a point to which we will return. Overall, we were surprised to find that most measures pointed toward greater repetitiveness in this mode of fiction and that, importantly, this tendency seemed to hold true across languages.

Because these measures alone gave no indication as to the possible reasons for increased repetition, our next step was to triangulate them with finer-grained lexical and grammatical features. That is, we sought additional proxies for the higher order phenomena of vernacular style, grammatical structure, and self-referential content. This included obvious things like the mode of narration (whether first person or not) and the ratio of verbs related to thought and feeling.[29] It also included features likely to be associated with the influence of Western grammar and translated works: ratio of first or third person pronouns; ratio of punctuation; ratio of only periods; and ratio of grammatical function words (stopwords). On their own, all these features, aside from mode of narration, turned out to be reliable indicators of overall generic difference. We assumed this would be true of pronouns and "thought/feeling" verbs given the confessional and solipsistic nature of I-novels and Romantic fiction, but it was not obvious that this would be true for stopwords (which are more frequent in these works) as well as punctuation (which are less). A possible reason for the latter, at least in the Japanese case, is that the works contain less dialogue.[30] Self-contemplation, we can imagine, does not leave time for small talk. Plotting these finer-grained features against our measures for repetitiveness, the most interesting finding was a correlation between entropy and the ratio of verbs signifying acts of contemplation, feeling, and mental attention. This relation holds for both Japan and China regardless of whether the work is narrated in the first or third person, but also holds *within* each genre (Figure 2).

[29] For Japanese, the words we included were the following: 思, 感じ, 考え, 心持, 気分, 心配, 気持, 考へ. On the China side, we included the following words: 想, 觉得, 知道, 心里, 晓得, 精神, 想起, 感到, 觉, 感觉, 思想, 感情.

[30] We were not able to confirm this in the Chinese case because of less reliable OCR results for some of the popular texts. Further correction is necessary to ensure that punctuation accurately reflects the original texts.

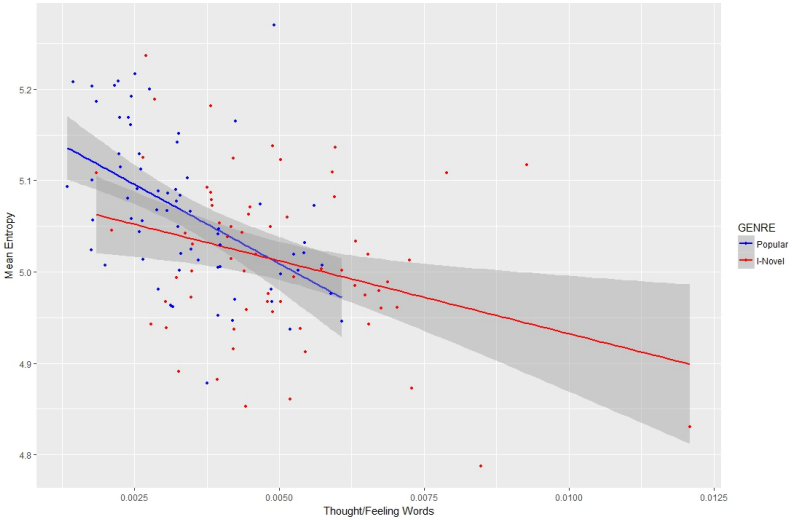
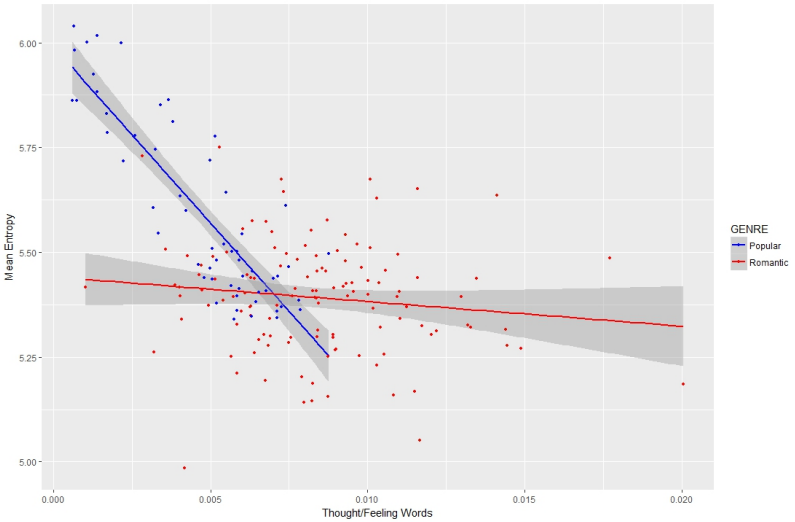


Figure 2. Plots for the ratio of "thought/feeling" words against average entropy for Japan and China, with linear regression lines fitted by genre. In both cases, we can observe that as the ratio of "thought/feeling" words increases (horizontal axis), the mean entropy of the texts decreases (vertical axis), indicating more lexical repetition.



A comparison of the 100 most redundant passages with the 100 least redundant

passages in the I-novel and Romantic texts reveals that some "thought/feeling" verbs are uniquely distinctive to the most redundant passages.[31] These results suggest a strong association between simple lexical repetition and the representation of cognition.

The final step in confirming repetition as a marked tendency in I-novels and Romantic fiction was to combine all of these individual features into a single model in order to evaluate their relative weight in distinguishing this fiction from popular works. We wanted to know how well such a model would predict the genre of a text based solely on measures like entropy, TTR, proportion of "thought/feeling" words, and so on. Using a logistic regression classifier with best subset selection, we confirmed what we saw with the individual features.[32] In the Japanese case, the classifier guessed the assigned genre of the text with 80% out-of-sample accuracy. In fact, it needed only the Kontoyiannis entropy measure together with the ratios of thought/feeling words, stopwords, and periods to achieve this accuracy. This does not mean that the other features were not also discriminative, only that the classifier could perform equally well without them. On the China side, the model guessed the correct genre nearly every time (Table 1), needing only average entropy and Yule's K to do so. Here, redundancy and uniformity of vocabulary alone are enough to separate the two corpora. Unfortunately, unlike in the Japanese case, our inability to control for linguistic difference makes it difficult to determine if the repetitiveness is mostly an effect of language or if the impact of psycho-narration is also playing a part. Nevertheless, both results support the notion that repetition was fundamental to the experiments that I-novelists and Romantic writers were conducting. The aesthetic currents that converged to produce these genres of self-referential writing appeared to manifest across diverse cultural and linguistic contexts as a compulsion to repeat oneself.

Japanese Corpora:

[31] To establish distinctiveness, we compared word frequencies in the 100 most entropic chunks and the 100 least entropic chunks using a chi-squared test. Words occurring four or fewer times were excluded. The thought/feeling words most distinctive to low entropy I-novel passages were 考え (think) and several inflections of 思う (think), while for the Romantic passages they were 心 (heart/mind) and 知道 (know). All of these words were in the top 5% of most distinctive words as determined by the chi-squared test score.

[32] A logistic regression classifier uses a set of independent variables (our features) to make a categorical decision about the class (or genre) label of a work. It looks at the distributions of these features across a subset of the corpus and determines whether they differ significantly between genres. Best subset selection will attempt every possible combination of features in order to identify the combination that is most discriminating of the two groups of texts. Although computationally difficult with more than ten features (e.g., 1000 combinations), we had a relatively small feature set. We ran the classification multiple times using a different set of starting features and the "best" features were almost always the same, thus giving us confidence in the stability of the procedure. The classifier uses these features to make a decision about the category of a work that it has not seen previously.

	Popular	I-novel
Popular	5.1	1.9
I-novel	0.9	5.

Chinese Corpora:

	Popular	Romantic
Popular	12.3	0.1
Romantic	0.1	5.6

Table 1. Confusion matrices for our logistic regression classifier. These matrices were produced using ten-fold cross validation and represent how often, on average, the classifier predicted the assigned class label. In the Chinese case, we can see that "Popular" works were almost never classified as "Romantic" works, and vice versa. In the Japanese case, "Popular" works were slightly harder to distinguish from "I-novels."

Reading Repetition

Having discerned such a compulsion in early-twentieth-century self-referential fiction, it remains to be seen what this tendency means at the level of style or in terms of generating a new kind of "mentality." And given the restricted definition of repetition we are using, this tendency needs to be situated against other ways for demarcating and reading the meaning of repetition. What our measures precisely capture is the relative degree to which a writer repeats the same limited set of words within a 1,000 word window. The more he or she does so across many such windows in a given text, the more repetitive is the text overall. Our goal is to understand whether this sustained compression of vocabulary corresponds to particular linguistic modes, narrative situations, or subject matter, but also whether it generates particular aesthetic effects.

Of course, readers do not read a text in discrete, 1,000 word chunks. Repetition as we measure it represents a narrow sliver of the many kinds of repetition that might interest literary scholars. J. Hillis Miller catalogs other alternatives in *Fiction and Repetition*: "On a small scale, there is repetition of verbal elements: words, figures of speech, shapes or gestures, or, more subtly, covert repetitions that act like metaphors…On a larger scale, events or scenes may be duplicated within the text…A character may repeat previous generations, or historical or mythological characters…Finally, an author may repeat in one novel

motifs, themes, characters, or events from his other novels."[33] Miller goes on to suggest that we interpret novels in part by noticing these recurrences, for "any novel is a complex tissue of repetitions and of repetitions within repetitions, or of repetitions linked in chain fashion to other repetitions."[34] The problem, of course, lies in this noticing. As Gilles Deleuze observes, repetition of some thing or event is essential to it acquiring a fixed identity in one's mind—and so too the mind of the reader—but this identity is always virtual to the extent that repetition itself is posited by way of abstraction. We abstract out the infinite variations that intervene between one occurrence of a thing and the next in order to make the idea of repetition possible.[35] As readers, our noticing of repetition in a literary text is always predicated on some method for delimiting the boundaries of repetition and for holding at bay all the myriad dimensions along which any two instances of a thing or event can differ.

This method is easy to articulate when one is working with individual texts or focusing on smaller units of analysis, like phonemes or words. Studies of alliteration, parallelism, or rhyme in poetry are exemplary in this regard. It becomes harder, however, as these units grow in complexity and as one tries to follow that repetition across more than a handful of texts. To trace the repetition of a theme or motif, for example, requires significant abstractions in order to fix the identity of that theme or motif across many instances. The less consistency there is in these abstractions, the harder it is to assert that the same thing is being repeated, and the harder it is to provide a quantitative interpretation of this repetition, since repetition is meaningful to the extent that something is repeated more (or less) often than might be expected. Linguists who work on repetition are especially attuned to this fact, and thus take great care to clearly articulate both the object that is being counted and the background against which these counts acquire significance. A recent methodological survey, for instance, outlines no fewer than ten forms that repetition might take, including absolute repetition (a simple frequency); positional repetition (an unexpected higher or lower frequency at a given position in a text); associative repetition (two things coinciding more often than expected in a given frame); and repetition in blocks (a thing is repeated according to a lawful distribution over blocks of text).[36] In each case, importantly, it is assumed that repetition makes quantitative sense only relative to existing patterns of usage, whether in terms of the thing itself, its use with re-

[33] J. Hillis Miller, *Fiction and Repetition: Seven English Novels* (Cambridge, MA: Harvard University Press, 1982), 1-2.

[34] Miller, *Fiction and Repetition*, 2.

[35] James Williams, *Gilles Deleuze's Difference and Repetition: A Critical Introduction and Guide* (Edinburgh: Edinburgh University Press, 2013), 11-12.

[36] For the full list, see Gabriel Altmann and Reinhard Köhler, *Forms and Degrees of Repetition in Texts* (Berlin: Walter de Gruyter, 2015), 5-6.

spect to some context, or its use with respect to time.[37] Such strict assumptions may limit the kinds of things one can count, but the advantage over qualitative approaches is that they allow one to scale up one's analysis and to reason about relative degree of repetition across a larger number of texts.

At the same time, this advantage does not make quantitative approaches to repetition any less "virtual," in Deleuze's sense, nor does it help to interpret the linguistic function or symbolic effect of such repetition. As linguists themselves have been careful to point out, there are many reasons why repetition occurs. There are external structural factors, of course, such as natural limitations imposed by grammar or the lexical inventory of a language. Repetition may also be used intentionally and strategically to establish thematic bonds, provide rhetorical emphasis, for stylistic effect, or even to control information flow. At a more granular level, it is used in conversation to aid in comprehension; to increase efficiency by providing a frame for new information; and to enhance the feeling of mutual participation in a conversation and thus to strengthen social bonds. It can even be unconscious, such as when a speaker repeats what is being said with a split-second delay or otherwise imitates the speech of others. When this imitation becomes obsessive, or else automatic in the sense of not being motivated by external stimuli, then the interpretation of repetition veers toward the psychological and towards mental or neurological maladies.[38] This last kind of reading is especially relevant for our study given the overt psychological orientation of much of the material.

Freud was one of the earliest to closely consider the psychological function of repetition by interpreting acts of reproduction as the psychic mechanism's resistance to confronting an unpleasant, repressed memory.[39] He treated the topic at length in *Beyond the Pleasure Principle* (1920) and pursued different explanations for the compulsion to repeat, variously ascribing it to the subject's attempt to gain mastery over a situation, the expression of a repression in the subject's ego, and the "death drive," a kind of instinctual desire, rooted at the cellular level, to return to a pre-organic state. For all his attention to the role of language in psychopathology, however, the founder of the "talking cure" largely overlooked

[37] Deborah Tannen refers to these multiple contexts of repetition as "dimensions of fixity," noting that while "all expressions are relatively fixed in form, one cannot help but notice that some instances of language are more fixed than others. This may be conceived as a number of continua reflecting these dimensions. There is, first , a continuum of relative fixity in form, another of relative fixity with respect to context, and a third with respect to time." See her *Talking Voices: Repetition, Dialogue, and Imagery in Conversational Discourse* (Cambridge: Cambridge University Press, 2007), 55.

[38] For an extensive list of possible interpretations of repetition, see Altmann and Köhler, 2-3; and Tannen, chapter 3.

[39] Freud, "Remembering, Repeating, and Working-Through"[1914], in *Standard Edition* vol. 12, 145-157.

specific acts of repetition in speech or prose, instead focusing on dreams, games, and other forms of acting-out or repression.

The importance of linguistic repetition gained traction with the rise of psycholinguistics in the 1940s and 1950s, whereby the interpretation of repetition as a window into human psychology took a strongly quantitative turn. As William Levelt notes in his comprehensive history of the field, "it had suddenly become possible to quantify the amount of information transmitted between sender and receiver, its redundancy, transmission rate and noise in the channel, and so on."[40] George Zipf's research on word frequencies was an early precursor to this transformation, and his now famous law was motivated by his belief in a deep property of mind he called "the principle of least effort." He derived this property from a model of communication in which speakers benefit from reducing "the size of [their] vocabulary to a single word" while listeners prefer to "increase the size of a vocabulary to a point where there will be a distinctly different word for each different meaning."[41] It is the balancing out of these two forces in communication that generates the smooth rank-frequency relationship described by his law. Yet this norm was defined by observed deviations from it. Specifically, Zipf analyzed the recorded speech of autistic and schizophrenic patients and argued that a sharper negative slope in the rank-frequency relation meant a smaller set of words being overloaded with a greater set of meanings, suggesting that such patients were less inclined to adjust their private languages to a common cultural vocabulary.[42]

So, too, were other early psycholinguists like John Carroll and Wendell Johnson drawn to lexical repetition and diversity as indexes of deviation from social norms. Johnson participated in several studies in the early 1940s that used his TTR measure, among others, to compare speech and writing between adults and children, age groups, IQ groups, sexes, schizophrenics, and normal adults.[43] These studies found that higher IQ correlates with higher lexical diversity and higher TTR; that the college freshman's TTR is slightly higher than the schizophrenic's; and that speech on the telephone is more repetitive than schizophrenic speech. The notion that lower diversity in word use, and greater repetition, signaled abnormal conditions of some sort (e.g., less education, less ability to relate to others, or extreme orality) played an important part in early

[40] Levelt, *A History of Psycholinguistics: The Pre-Chomskyan Era* (Oxford: Oxford University Press, 2013), 5.

[41] George Zipf, *Human Behavior and the Principle of Least Effort* (Cambridge, MA: Addison-Wesley Press, 1949), 21. The thesis was originally formulated in *The Psycho-Biology of Language: An Introduction to Dynamic Philology* (Boston: Houghton Mifflin Company, 1935). His theories are summarized in Levelt, 453.

[42] Zipf (1949): 285-87.

[43] Levelt, *A History of Psycholinguistics,* 456.

psycholinguists' ideas about language and cognition. Later on, entropy too, and its companion redundancy, would become compelling frameworks for thinking about the psychology of language, whether in Roman Jakobson's musings on language as a code whose conventions differ between inner, affective language (which tends to be more redundant) and exteriorized, intellectual language; or Anthony Wilden's use of redundancy to reinterpret Freud's description of psychic symptoms as revealed in multiple, over-determined ways. The compulsion to repeat, he argues, is really a safeguard against inner mental noise.[44]

Thus do forms of repetition help define and even construct the modern, psychological subject. This brief history adds another essential dimension to the rich hermeneutic space through which repetition can be read. As we have seen, it has offered a scale along which to imagine differences between orality and writing; between inner language and exteriorized speech; between isolating psychological conditions like schizophrenia and normative, socially-aware subjectivity. By quantifying the repetitive tendencies of I-novels and Romantic fiction, we gain access to this space at the scale of hundreds of texts. Our measures also help us to orient texts within this space along a continuum. We can do so in terms of their relative redundancy, but also by considering the extent to which their measured features, as a composite, cohere with the features observed in one genre and not the other. The following plot shows the Japanese texts most likely to be "I-novels" as judged by our classifier and the features in our model (Figure 3). The higher a work appears in the plot, the more confident the classifier is that the work shares the quantitative tendencies observed in other "I-novels" in our corpus.

[44]Roman Jakobson, "Langue and Parole: Code and Message," in *On Language*, eds. Linda R. Waugh and Monique Monville-Burston (Cambridge, MA: Harvard University Press, 1990): 97-98; and Anthony Wilden, *System and Structure: Essays in Communication and Exchange* (London: Tavistock Publications Limited, 1972), 35-37. More recently, the Stanford Literary Lab, in a study of the differences between popular and canonical novels, has hinted at a potential link between repetition, as measured by TTR, and narratives of trauma. See Mark Algee-Hewitt, et al., "Canon/Archive" (2015), 9-10.

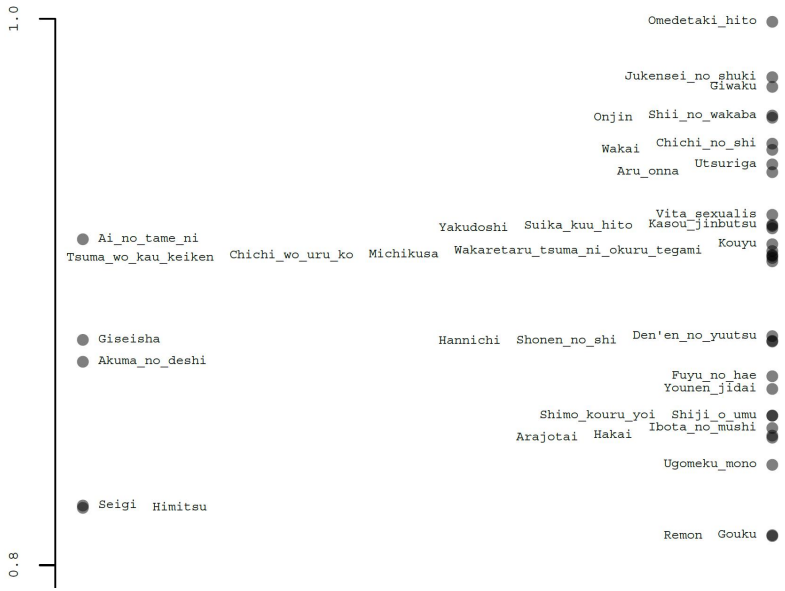


Figure 3. The most "I-novel" like titles based on our classification model. Texts on the right are those originally labeled as "I-novels." Texts on the left are "Popular" works. The closer a title is to 1.0, the more the model thinks it is an I-novel based on what it has learned from the corpus.

While being able to reorient texts along these sorts of continuums generates new kinds of comparison, it is still up to us to navigate the hermeneutic space of repetition. Can the repetition we are capturing quantitatively be explained merely by the desire for a colloquial style or by adoption of foreign grammar? Can we read into it a strategy for linking repetitiveness to interior mental processes and possibly mental breakdown? We have argued that repetition as style is an epiphenomenon of all of these things, but only by examining individual texts can we understand how they are interacting with specific models of literary subjectivity across diverse cultural and linguistic contexts. This step is also crucial because repetition alone, as we have measured it, hardly captures all the differences separating self-referential fiction from other modes of writing. Our results indicate that some I-novels and Romantic fiction are relatively un-repetitive, while a few popular works (mostly detective fiction in the Japanese case) are nearly as repetitive as the most self-redundant works. Exploring such exceptions is important for future work, but here we will consider texts that push the tendency to repeat to its extremes. They help tighten the focus of the comparative lens that repe-

tition as style provides and allow us to examine the particular aesthetic uses to which it was being put.

Repetition as Style

Within this narrower lens appear the works at the top of Figure 3 deemed most "I-novel" like due to their propensity for repetition and certain lexical and figural items (i.e., thought words, stopwords, periods). Several of these confirm a reading of repetition as the surface effect of transformations in literary language combined with attempts to narrate psychological interiority and mental disorder. "Shii no wakaba" (Young Pasania Leaves, 1924), for instance, a late work by Kasai Zenzō, is noteworthy for having been dictated to a scribe over a twelve-hour period.[45] As Edward Fowler observes, this lends the work an orality that "asserts itself emphatically and repeatedly," not least because Kasai refused to read over what he had dictated. This left him, like a meandering storyteller, to constantly hark back to his previous utterances through his selective memory of them, leading to "increasingly redundant summations…[and] frenzied yet almost formulaic musings about insanity."[46] The result was a highly "disconnected" narrative style that shifted haphazardly from one episode to the next even as it churned over the same mental ground. In *Giwaku* (Suspicion, 1913), dubbed by critic Hirano Ken to be Japan's first "true" I-novel, we are presented with a similar kind of "narrative claustrophobia," as Fowler puts it.[47] The novel's author, Chikamatsu Shūkō, is well known for protagonists who exhibit a "myopic preoccupation with private life" and revel in "self-engendered doubts," producing an "isolated (as opposed to an individuated) consciousness" that is almost entirely cut off from political, social, or familial concerns and which dwells repeatedly on certain periods in the author's life.[48] As is true for *Giwaku*, the periods usually involve abandonment by a former lover and the feelings of disgust, rage, and desperation that ensue as the protagonist combs his memory for past evidence of deceit. But in this case, such feelings are all we get. The entirety of the novel's action takes place in the narrator's mind, a fact foreshadowed in its opening lines: "Usually I hid under my quilt and, in my mind, imagined and redrew the scene of your murder and of my imprisonment; imagined and redrew it. While contemplating whose

[45] The author was frequently bedridden during his last years due to excessive drinking and dictation was the only way that publishers could extract material from him. See Fowler, 272-73.

[46] Fowler, *Rhetoric of Confession*, 274.

[47] Fowler, *Rhetoric of Confession*, 151. Hirano made this claim in his *Geijutsu to jisseikatsu* [Art and Private Life] (1964), cited in Fowler, 150-51.

[48] Fowler, *Rhetoric of Confession*, 151-52.

wife you'd become and how to find you, day after day I could think of nothing better but to imagine the same scenes over and over again, almost as if I were being suffocated."[49]

Giwaku appeared at a time when narratives of mental breakdown were seen by some as the hallmark of literary value. As the writer Funaki Shigeo remarked in the same year *Giwaku* was published, of the highest quality works being produced now "there were none that did not acknowledge the operation of the nerves (*shinkei*) to some degree."[50] Funaki was himself a contributor to a sub-genre of self-referential fiction known as the "neurasthenia novel" (*shinkei suijyaku shōsetsu*), which helped to reinforce the conceit that psychological deterioration and mental anguish were the proper source and subject of modern art.[51] This was a conceit shared with certain strains of European Naturalist writing as well. Charles Baguley, in his study of French naturalist writing, argues that while naturalist novels were too general in scope to give rise to specific thematic determinants, one of their characteristic movements was in the "direction of disintegration and confusion"; "from order to disorder, from mental stability to hysteria and madness."[52] Repetition as style was one way that writers like Kasai and Chikamatsu could inscribe this movement toward mental breakdown in the very language used to describe it.

The most "I-novel" like text in our corpus is Mushanokōji Saneatsu's *Omedetaki hito* (A Blessed Man), from 1910. Though it has not fared well in critical history, Mushanokōji himself was retroactively recognized in the 1920s as a founder of the I-novel. As novelist Uno Kōji put it, his "remarkable style," combining a true colloquial style with reforms to the written language was, "in a certain sense, the origin of the I-novel."[53] Indeed, Naturalist writers like Mushanokōji have, as a group, been closely linked to both the development of colloquial language and the adoption of Western syntax and expressions.[54] That a work by one of the

[49] Translated from Chikamatsu Shūkō, *Chikamatsu shūkō shū* [Chikamatsu Shūkō: An Anthology], ed. Hirano Ken, in *Nihon bungaku zenshu*, vol. 14 (Shūeisha, 1974), 100.

[50] See Hibi Yoshitaka, *'Jiko hyōsho' no bungaku-shi* [A Literary History of 'Self Representation'] (Tokyo: Kanrin shobō, 2002), 228. The statement was made in the context of a review of recent works by Shiga Naoya, who went on to become one of the most recognized I-novelists of the era.

[51] Hibi, 228-34. See also work by Christopher Hill, "Exhausted by their Battles with the World: Neurasthenia and Civilization Critique in Early Twentieth-Century Japan," in *Perversion and Modern Japan*, ed. Nina Cornyetz and Keith Vincent (London: Routledge, 2009); and Pau Pitarch-Fernandez, "Cultivated Madness: Aesthetics, Psychology and the Value of the Author in Early 20th-Century Japan." PhD dissertation, Columbia University, 2015.

[52] David Baguley, *Naturalist Fiction: The Entropic Vision* (Cambridge: Cambridge University Press, 1990), 207.

[53] Uno Kōji's essay, "Watakushi shōsetsu shiken" [Personal View of the I-novel], is cited in Lippit, 29.

[54] Kisaka, 382-83. As Tomi Suzuki has noted, Tanizaki Jun'ichirō also made this connection in a

genre's recognized founders—which opens with a note from the author stating, "I believe in the existence of a selfish literature, a literature for the self"—shows up as most "I-novel" like gives us confidence that repetition is capturing a definite tendency in the genre.

How *Omedetaki hito* handles mental breakdown presents another variation on the theme and is worth analyzing at greater length. The madness depicted in this novella is once again both mono-maniacal and narcissistic, with repetition and redundancy of internal thought driving much of the work's psychological description. So much so that the reader is made to feel the text's repetitiveness in an extremely visceral way. In what amounts to an account of one man's pathetic attempts to win the attention of a girl, we are told by the narrator five times in the first page that, "I am starved for women." Each time he repeats the phrase nearly word for word. Also apparent in the first pages is an excessive use of the first-person pronoun *jibun*, which is used as a subject marker in almost every other sentence and in ways that are completely unnecessary grammatically. It is as if the narrator feels compelled to discursively reaffirm his self presence at every instant, lest the reader forget who is narrating. This compulsion becomes particularly acute in moments of prevarication that perpetually delay actual encounters with the woman, Tsuru, by whom he is so smitten.

> I hear that Westerners think Friday is taboo. And so for the past two or three years, when I want to go meet her, I make it a point not to go and meet her on Fridays. But there are times when I think this superstition is bad and go out. But that makes me feel a little odd. Since she moved, I have to travel a little further to meet her. Thus it bothers me to go out of my way and go on a Friday. But there are times when I think that's just superstition, superstition isn't good, and I go anyway. At those times I've even thought that it's probably better I don't meet her. I feel more upset about going on Friday than about the fact that I'm meeting Tsuru after not seeing her for nearly a year. But I want to meet her. Then I think that, after all, since I haven't seen her until now, it's better not to meet her, whether I've talked myself into it or not. So finally I give up on going to meet her.[55]

Here, and throughout the text, Tsuru is merely a screen for reflecting the narra-

1929 essay, where he argued that "those who had contributed most to the westernization and artificiality of the modern vernacular style were novelists in the Japanese Naturalist movement," most of whom continued to move toward westernization of the written language. Suzuki, 176.

[55] Mushanokōji Mushanokōji, "Omedetaki hito," *Gendai Nihon bungaku zenshū*, vol. 40 (Chikuma Shobō, 1973), 7-8.

tor's convoluted mental deliberations, his desire for her seeming to be motivated apropos of nothing and growing more intense the longer he manages to avoid physical encounter with her. She becomes an excuse for flights of fancy that have the narrator pondering the nature of lust, the nature of self, and the possible consequences for his own self-infatuation should he find a way to marry her. Needless to say, no such marriage comes to pass. The only time they actually meet is a chance encounter on a train, where once again the narrator fails to turn thought into action.

> I stood up just before the train got to Yotsuya. I looked at Tsuru. My eyes met with Tsuru's. Tsuru quickly turned her eyes away. I decided then to pass in front of her and stop. The train was coming to a stop, but Tsuru didn't stand. And her face was turned away from me. Finally the train stopped. I tried to pass in front of Tsuru. Then she suddenly stood. I put my hand on Tsuru's back. I decided to follow her off the train. Then, near the exit, a man with his child stood up. I didn't have the courage to rudely push past the man and follow right behind Tsuru. I let the two come between Tsuru and I.[56]

Repetition not only suggests inner turmoil, but also serves to slow down the action, linking each step to the next while preserving a sense of singular focus. When the narrator finally summons the courage to call out her name, she responds with a curt, "Can I help you?," before walking off in another direction. Despite all indications to the contrary, he takes this as a sign of her love for him despite this being the last we see of her. This is as it must be, however, for were Tsuru to enter the story as a living, breathing character, it would only derail the narrator's one-track mind. A repetitive and redundant style has here given birth not to madness, per se, but to a radically self-centered narrative mode with no precedent in Japanese fiction.

[56] Saneatsu, "Omedetaki hito," 25.

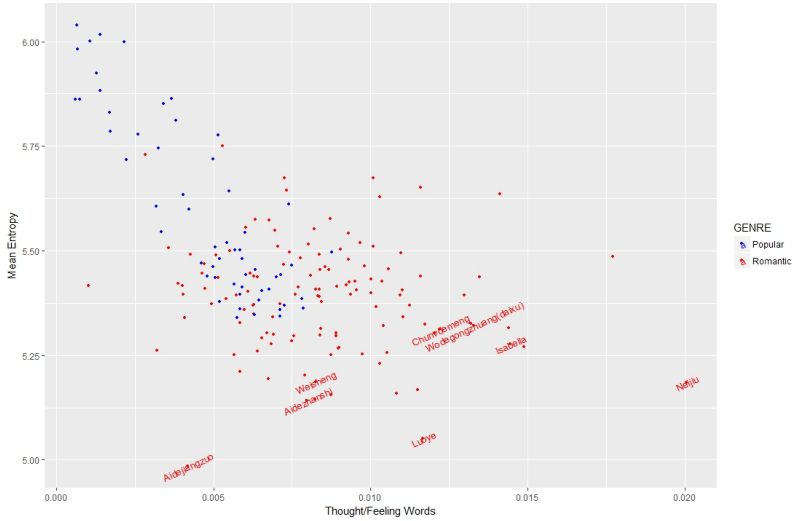


Figure 4. The ratio of "thought/feeling" words plotted against average entropy for the Chinese texts, with some outlier works highlighted by title.

On the China side, we again find many self-obsessed works at the intersection of high repetition and excessive thought in our Romantic corpus (Figure 4). Curiously, most of the extreme outliers belong to the same author: Ye Lingfeng (1905-1975). Ye occupies a shifting and uncertain place in the historiography of modern Chinese literature. A relative late comer to the scene of Romantic writing, Ye joined the Creation Society in 1925, a time when the group was already pivoting away from indulgent narratives of the self toward a politically-inflected interest in national identity and class consciousness. Eager to capture a place in the burgeoning field of modern literature, Ye marketed himself as a dandy and iconoclast, and he rapidly found success writing fiction that featured titillating love triangles, urban decadence, Freudian-inflected depictions of sexual desire, and focused on bodily and psychological "abnormalities" (from masturbation, castration, and homosexuality to bisexuality, suicide, and incest) in such a way as to "achieve mental confusion."[57] Such narratives would earn him a central place in the scene of the high modernist literature of the 1930s promoted by the so-called Neo-Perceptionist writers like Mu Shiying, Liu Na'ou, and Shi Zhecun. But in the 1920s, his work was still aligned with Romanticism, earning Ye the status of being a "2nd generation Romantic." His liminality made Ye difficult to classify in his own time, and he remains somewhat of an understudied figure in

[57] Yingjin Zhang, *The City in Modern Chinese Literature and Film: Configurations of Space, Time, and Gender* (Stanford: Stanford University Press, 1996), 211.

scholarship today.[58]

One of Ye's most repetitive texts is his 1928 story "Aidezhanshi" ("Warrior for Love"), a representative piece featuring a love triangle between urban youth. The narrative alternates points of view between a young female author, Sophie, and her smooth-talking beau, Ping, who fancies himself a Napoleon in the realm of love. Displaying little dialog or character interaction, each section is essentially an insular contemplation of the character's inner state. The story opens with Sophie writing in her journal, pining away for Ping:

> I don't believe this is spring. Lower and lower, this bolt of grey canvas is already so low as to push down upon my head. I don't believe this sky has ever had a golden-red sun, I don't believe this sky has ever had a mirror-like moon. It's incapable of having them. These are all dreamed-of things, these are all lies told by fortunate people to unfortunate people. Do you believe? Do you believe this ground could have a strip of verdant-green grass, the blue-black stream at its side a pink-blossomed peach tree, under whose flowers a young man is just now embracing—What? I don't believe. I don't have the courage to write on.[59]

Close up, it is easy to see how the recurrence of words like "lower," "I," and "believe," would culminate in a low entropy score. Such rhythm is sustained throughout the narrative and augments the excessively sentimental tone of the two narrators, thereby imparting the text with a strong dose of imaginative interiority. As the story progresses, Sophie becomes increasingly disillusioned by the affair, while Ping, conversely, grows ever bolder in his lecherous pursuit of women. The story ends darkly when, to take revenge on Ping for his infidelity, Sophie gets him drunk and stabs him to death, triumphantly declaring herself a "warrior for love."[60] This violent overreaction retroactively makes Sophie's repetitive writing style appear as dangerously off-kilter and manic. Sophie is not a particularly sympathetic character, and the story should not be read as a feminist indictment of patriarchy. Instead, repetition stylistically augments Ye's penchant for the sensational, a flourish which one scholar has characterized as "kitschy."[61] This kitsch

[58] Ye's marginalization in scholarship is also due to his attacks on Lu Xun in the late 1920s, and his politics during the 1930s. See Zhang, 208.

[59] *Ye Lingfeng xiaoshuo quanbian* Vol. 1 (Shanghai: Xuelin chubanshe, 1997), 168.

[60] This ending foregrounds Ye's debt to Oscar Wilde's *Salome* and its depiction of a femme fatale who kills in the name of love. See Xiaoyi Zhou, "Salome in China: The Aesthetic Art of Dying," in *Wilde Writings: Contextual Conditions*, ed. Joseph Bristow (Toronto: University of Toronto Press, 2003), 295-316.

[61] "Although Ye Lingfeng's fiction contains"newness" and experimentation, he is clearly committed to kitsch, which suggests repetition, banality, triteness. . . . [his stories are] at once avant-garde

can be read as an attempt to push the rhetoric of the modern self to excess, building upon and amplifying the repetitive style of psycho-narration characteristic of his Romantic predecessors.

Does such heightened repetition help explain Ye's outlier status in the literary field? Or is it the other way around? Both interpretations are possible. But our larger point is that such a correlation illuminates how the linguistic tendencies of Creationist Romanticism were pushed to their extremes just as the genre was on the verge of a major transformation.[62] That a writer on the fringes of self-referential Romantic writing should be the most extreme example, rather than a writer at the origins, as in the Japanese case, suggests that our measures might be capturing different moments in the respective trajectories of these modes of writing. In this case, the surprising degree to which Ye stands out raises new questions about the relationship of lexical redundancy to commercial viability of May Fourth literature in 1920s Shanghai, and the extent to which followers of a literary movement will exaggerate stylistic trends developed by originators of that movement.[63]

While Ye's stories are, on average, the most repetitive, some important figures in the Romantic movement emerge when we examine other highly redundant passages. Yu Dafu's "Jiedeng" ("Street Lamps," 1926), for example, and Guo Moruo's novel *Luoye* (*Fallen Leaves*, 1925), contain some extremely repetitive sections. On the whole, these works share with Ye's stories a number of characteristics, such as being written in the present tense and having a minimized plot.[64] They also have the tendency to adopt a narrative mode emphasizing direct address, either in epistolary and diaristic framing devices, or in the form of extended blocks of reported speech. Such vernacular style and ample reported speech likewise feature prominently in works with low entropy as determined by Kontoyiannis's measure, which captures repetition of longer sequences of words rather than just

and receptive to mass culture." See Jianmei Liu, "Shanghai Variations on 'Revolution Plus Love,'" in *Modern Chinese Literature and Culture*, 14, no. 1 (Spring, 2002), 82 and 84.

[62] Shu-mei Shih describes Ye's emergent literary group in the late 1920 as taking "Guo Moruo's 'explosion of the self' and Yu Dafu's self-indulgence to an extreme, aggrandizing the self in defiance of all constrictive norms and celebrating sexuality without the kind of anxiety that had troubled their May Fourth predecessors." See *Lure of the Modern: Writing Modernism in Semicolonial China, 1917-1937* (Berkeley: University of California Press, 2001), 255.

[63] Because the genre of Romantic literature declined rapidly in the late 1920s with the politicization of the Creation Society, our corpus doesn't go beyond 1928. As such, more work would need to be done to measure the evolution of redundancy in Ye's later work, or to compare him to his 1930s contemporaries.

[64] Which is not to say that the narrators don't reminisce about the past. In stories such as "Remorse," "Madam Chrysanthemum," and *Fallen Leaves*, the reminiscent narrators are very self-conscious of the temporal gap between their memories and the putative present in which they are recording their narratives.

individual words (Table 2).

Genre	Author	Title	Entropy Score
Popular	Zhang Henshui	Tixiaoyinyuanxuji	1.315763
Popular	Youqihong	Bimenghen	2.796662
Romantic	Ye Lingfeng	Aidejiangzuo	3.579148
Romantic	Ye Lingfeng	Luoyan	3.947289
Romantic	Cheng Fangwu	Bei'aideangeèr	4.212046
Romantic	Guo Moruo	Yeluotizhimu	4.264046
Romantic	Guo Moruo	Shizijia	4.290531
Romantic	Teng Gu	Baizuchong	4.317347
Romantic	Yu Dafu	Jiedeng	4.485555
Romantic	Tao Jingsun	Muxi	4.485555
Romantic	Ye Lingfeng	Kouhong	4.544576
Romantic	Ye Lingfeng	Mingtian	4.574672
Romantic	Zhang Ziping	Aizhijiaodian	4.60517
Romantic	Ye Lingfeng	Jiulumei	4.60517
Romantic	Guo Moruo	Tingzijianzhong	4.636077

Table 2. The fifteen lowest entropy texts in the Chinese corpus based on non-parametric entropy measure. The highest entropy texts by this measure range from 7.1 to 8.4, and are thus twice as redundant. The median score for each genre is 5.7.

Here too we see works by Ye Lingfeng, but also several Romantic short stories by Guo Moruo: "Shizijia" ("Crucifix," 1924), "Ye Luoti zhi mu" ("Ye Luoti's Grave," 1924), and "Tingzijian zhong" ("Within the Garret," 1925). These latter texts feature repetition in ways even more obvious to a human reader, whether in the frequent exclamatory or emphatic adverbial modifiers such as "extremely carefully, carefully. . . lightly, lightly" ("Tingzijian zhong"), or the nearly verbatim duplication of sentences: "When the nurse reached out her hand to take his pulse, in a state of half-consciousness he instead said 'Ah, many thanks, auntie.' When the nurse again reached out her hand to place the thermometer under his right armpit, he again said 'Ah, many thanks, auntie' " ("Ye Luoti zhi mu"). From the pen of a canonical author like Guo Moruo, then, comes an excessively repetitive style often seen as emblematic of vernacular writing of the period.

But while Guo Moruo's stories all invoke a strongly maudlin atmosphere characteristic of the frustrated May Fourth individual (and, coincidentally, feature death), it is important to point out that many of the Chinese texts with the highest rates of repetition are not limited to evocations of trauma or suffering. In fact, we found that words evoking "kumen," an influential sentiment connoting suffering/despair, do *not* meaningfully correlate with low entropy, unlike the "thought/feeling" words. Instead, as we see in Ye's work, redundancy is also an important source (or effect) of sexual titillation, sentiment, and free love. Such repetition works in the opposite direction of the Freudian death drive, and is instead in accordance with a kind of pleasure principle.[65] This pleasure may be

[65] Freud is especially warranted here because Ye Lingfeng was himself a devotee of Freudian psy-

celebratory, guilty, narcissistic, libidinal, or manic, indicating a richly complex relationship between repetition, interiority, and sentiment that helped define literary subjectivity in Romantic fiction.[66] But a more thorough exploration of repetition's relationship either to specific Freudian psychic mechanisms or to sentiment in general falls outside the scope of this paper.

Conclusion

What does the notion of repetition as style have to offer histories of East Asian literary modernity? As stated at the outset, these histories have alternatively upheld the coherency of a body of self-referential literature that joined linguistic transformation to psychological narration and also emphasized the impossibility of reducing it to a single set of formal characteristics. Our goal in this paper has been to take some of the consistencies associated with this literature and transpose them into a quantitative register, thus creating an alternate framework with which to explore the gap between generic coherency and generic ambiguity. This framework, while taking advantage of what scholars already know about this literature, was meant to use the affordances of computation to extend this knowledge in hitherto unexplored comparative directions.

What scholars already knew is that Japanese I-novelists and Chinese Romantic writers converged in the early twentieth century around a new grammatical mentality that synthesized linguistic change and a fascination with the psychologized self. Thus on the one hand works by writers like Mushanokōji and Ye illustrate the stylistic revolution that joined a colloquialization of the written language with the adoption of Western grammatical concepts and structures. Naturalist writers came to epitomize this revolution in Japan, while in China it was May Fourth writers (including the Romantic authors) who were advocates for vernacularization and Europeanization of Chinese writing. Aimed at providing a model of writing commensurate to the subjectivity of a modern individual, such language experiments in fact provided the conditions for the "discovery" of the self. As scholars such as Karatani Kōjin and Lydia Liu have shown, in the development of a modern interiorized subject in both Japanese and Chinese literary history, vernacular writing was generative rather than reactive: the new literary language

chology. See Jingyuan Zhang, *Psychoanalysis in China: Literary Transformations 1919-1949* (Ithaca: Cornell University Press, 1992).

[66] On sentiment, see Haiyan Lee's engagement with the "structures of feeling" of modern Chinese literature. As Lee states: "The modern subject is first and foremost a sentimental subject." *Revolution of the Heart: A Genealogy of Love in China, 1900-1950* (Stanford: SUP, 2007), 7.

was not so much a response to new desires to express the psychic interior as it was the condition which made such expression possible.[67] Knowing that Chinese Romantic authors drew direct inspiration from their Japanese counterparts and that I-novels were widely translated into Chinese during this period only reinforces the sense of a shared aesthetic and ideological project that cut across spatial, linguistic, and cultural divisions.

One comparative approach would be to show how these divisions particularize, in myriad ways, the manifestation of broader currents of linguistic transformation and psycho-narration in individual texts. Here we have taken a different tack by trying to inductively identify a trait that links the interaction of these currents across different individual expressions. By examining the I-novel and Romantic literature at scale and in comparative context, we have discovered their tendency toward repetition. To be sure, our quantitative model of repetition can only be a loose proxy for the much broader and more complex transformations of literary writing of this period. However, our initial results suggest that it captures something of the interaction of these currents since it identifies works that exemplify the grammatical mentality born of this interaction. Moreover, precisely because repetition is not identical or coextensive with these currents, our model provides a new and unfamiliar instrument for comparing and relating self-referential works to one another. Mushanokōji and Ye's texts have never appeared together in comparative histories of East Asian literary modernity, but it is hard to ignore their uncanny overlap as narratives of sexual desire and fantasy. This overlap is not something our model was designed to capture. It knows only that the texts tend to be more repetitive, and that repetition loosely correlates with words related to cognition. But by isolating repetition as a stylistic feature we were able to survey this more abstract textual register—the "unconscious history" of scale, as Braudel put it—to set familiar and unfamiliar works in new comparative context.[68] This context at once singles out individual authors like Mushanokōji and Ye (or Chikamatsu and Guo Moruo) who deliberately employed an extremely repetitive style in the service of self-discovery. At the same time, it reveals this choice to be a shared symptom of changing grammatical mentalities driven by psychological pressures from within and socio-linguistic pressures from without. It allows us to scan the textual surfaces of East Asian literary modernity and consider some of the broader ripples stirred up by deeper architectonic shifts taking place below.

[67] Karatani, 61.

[68] See "History and the Social Sciences: The *Longue Durée*," trans. Sarah Matthews, in *On History* (Chicago: University of Chicago Press, 1980), 25-54.