The Journal of Asian Studies (2022), **81**, 267–288
doi:10.1017/S0021911821002278

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Quantifying Quality: A Computational Approach to Literary Value in North Korea

Benoit Berthelier

Faculty of Arts and Social Sciences, University of Sydney, Australia
Email: benoit.berthelier@sydney.edu.au

Abstract

This article analyzes questions of literary production and literary value in the Democratic People's Republic of Korea (DPRK) using quantitative methods and data sets covering a significant part of the country's literary and intellectual output over forty years (1977–2016). The first part of the article adopts a thematic approach, using topic modeling and time series, to map the DPRK's literary output and investigate the influence of genre on literary consecration. The second part looks at questions of style and literary quality in the DPRK through a comparative analysis of multiple novels grouped by degree of literary prestige. It introduces a method to evaluate the originality of an author's writing style using a masked language model (BERT) and uses it to assess the role of conformity and originality in the DPRK's literary field. Even though writers consecrated by the state display a high level of stylistic conformity when writing mandated biographies of the ruling family, their independent output ranks higher in originality than nonconsecrated writers.

Keywords: authoritarian fiction; BERT; digital humanities; Korean literature; literary originality; literary value; language model; North Korea; topic modeling

"It is a law of nature that quantitative change brings about qualitative change," North Korean writer Pan Sangsŏ (2013, 32) once noted, paraphrasing a precept of dialectical materialism. North Korean literature's quantitative turn is older than the country itself: in 1946, two years before the Democratic People's Republic of Korea (DPRK) was formally established, the North Korean Writers' Union introduced production objectives. Like several other parts of the North Korean economy, literary production became centrally planned, with yearly quotas by genre and length and regular campaigns to increase production of works on themes mandated by the Korean Workers' Party (Park 1999; Shin and Oh 2000).

Outside the DPRK, the qualitative impact of this number-driven mode of literary writing has generally been assessed negatively: works written "less for the reader and other writers than for the eyes of the party hierarchy" (Pihl 1977, 81), "boring, similar to an assembly line" (Choi 1994, 81), and "repeating the same themes and material over and over" (Lee 2013, 181). Yet, a number of scholars have nonetheless argued that North Korean literature still exhibits literary qualities in spite of these constraints (David-West 2007; Epstein 2002; Immanuel Kim 2012; Ryu 1998), and they have sought to extract select examples of well-wrought narratives, complex characters, or poignant metaphors from the country's textual production. Despite being seemingly diametrically opposed, these two ways of assessing the value of North Korean literature nonetheless share a common premise about what literary writing is or ought to be. Both rely on a normative conception of literature as autonomous, free from the external political pressure of a government (Bourdieu 1991): isolating the literary bits from the heteronomous political discourse transparently reflects the binary logic of autonomy while characterizing North Korean literature as merely propaganda amounts to denying that any sort of literariness can be salvaged from its politically laden discourse.

Autonomy, however, is foreign to the literary tradition of the DPRK, and therefore it can be of only limited analytical value when trying to understand the dynamics and standards of the North Korean

© The Association for Asian Studies, Inc. 2022

https://doi.org/10.1017/S0021911821002278 Published online by Cambridge University Press

literary field. This article attempts to move beyond these limitations by mapping some of the features of literary texts that are actually valued (as evidenced by external signs of literary recognition) in the DPRK, regardless of their status in other, perhaps more familiar, systems of aesthetic valuation. How is literary quality assessed within the DPRK? If the political field plays an important role in the way a literary work is evaluated and valued, how is this relationship between politics and literary valuation concretely structured and realized? And if repetitiveness and conformism do not seem to be anathema, what is the place of originality in the North Korean economy of literary value?[1]

Autonomy, Autonymy, Automation

The absence of scholarly inquiry into the specific mechanics of the literary system in North Korea is symptomatic of a more general tendency to either dismiss politically laden, state-sponsored literature as presenting little artistic interest or to study it outside the literary institutions in which it was produced. It is not that authoritarian literature as such is not an object of research. Much to the contrary, a vast body of the scholarship on the topic has developed since the numerous landmark studies of the subject were published in the 1980s (Clark 1981; Gardner 1983; Günther 1990; Suleiman 1983). But the works studied are treated as ideological tools, independent of the system of publishers, critics, prizes, and academies in which they were originally embedded. On the other hand, theories of "world literature" (Buchholz 2017; Casanova 1999, 2005; Moretti 2013) conveniently ignore vast areas of literary history, such as East Asian proletarian literary movements (Cho 2018) or the USSR's "People's Republic of Letters" (Djagalov 2018), and end up offering a very Eurocentric and implicitly normative conception of literature (Friedman 2013; Neuman and Rippl 2017). Only when they can be opposed to dissident authors are authoritarian writers considered part of a literary system, but one in which autonomous actors struggles against heteronomous ones. In the case of Stalinist literature in the USSR, for instance, the dogmatists of socialist realism such as Gribachev or Kochetov are opposed to autonomous dissidents such as Akhmatova or Mendelstam.[2] Yet the latter were effectively excluded from the period's literary scene, while the rules and forces that allowed the former to rise to the heights of literary fame over a multitude of other zealous Party apologists remain largely ignored.[3]

The opposition between autonomous and heteronomous literature is more complex than the common distinction between political and pure literature (Goldstone 2013; Van Rooden 2015). After all, Mendelstam wrote a considerable amount of political verse, as did even more openly militant yet still highly regarded poets such as Mayakovsky. But there is a difference between the autonomous engagement (Bourdieu 1992, 442) of writers committing to a cause they believe in (often clashing with authorities in that process), and the bureaucratic labor of writing for the state according to policy objectives, from potato yield increase to electricity saving campaigns. In the former case, literature is still produced according to a paradigm of literary autonomy, which favors distinction and experimentation, and these literary qualities can be acknowledged by critics regardless of the work's political dimension. In the latter case, however, while originality may not be nonexistent, other factors come into play, such as how relevant the work is to current policy or how precisely or comprehensively government objectives are communicated. In the autonomy paradigm, which considers literature and politics as separate fields, these factors are not literary—but they are in North Korea. That is why trying to isolate "purely" literary aspects in North Korean works can only be a frustrated endeavor. Whatever aspects are isolated will necessarily appear inferior compared with works produced under the autonomy paradigm, since they only constitute a part of what makes the work valuable in the

[1]The term "economy of literary value" refers to the different mechanisms by which literary value is assessed in a manner that is distinct but not necessarily unrelated to other valuation systems such as commodity markets (Casanova 2005, 13–14; Herrnstein Smith 1983).

[2]It is worth clarifying that the opposition between autonomous and heteronomous does not necessarily map to the distinction between "pure" and political literature (cf Bourdieu's [1992, 551] discussion of Zola's autonomous engagement). The distinctive criterion is the submission to state-imposed rules and standards.

[3]The work of Alla Latynina (1996) on Stalin Prize recipients is a notable exception.

https://doi.org/10.1017/S0021911821002278 Published online by Cambridge University Press

DPRK. And, more importantly, the method will overlook the other (for instance, political, moral, or ideological) valuation criteria actually used in the country.

Indeed, literature being propagandistic does not preclude the existence of a valuation system (or, in Bourdieusian terms, of symbolic capital): there can be good or bad propaganda, and the criteria separating one from the other are open to interpretation and competition. Some topics of propaganda might carry more value than other. For instance, it is not inconceivable that writing about war and writing about potatoes would carry different levels of literary prestige. Furthermore, when the same heteronomous norm (e.g., loyalty to the Party or any political leader) applies equally to all participants in the field as a prerequisite to entry, it loses its power as a positive valuation factor. Instead, it will primarily work negatively, as a factor of exclusion—for instance, when writers are purged for failing to toe the Party line. It therefore remains necessary to understand the standards against which literary works and their authors are judged once political orthodoxy and loyalty are accounted for.

This is not to advocate for North Korean literature or to call for a reevaluation of the literary heritage of Stalinist poetry. Both tasks entail a judgment of taste that is best left to critics or readers themselves. Rather, the point is to develop a way to analyze the workings of literary systems under authoritarian regimes—and, more broadly, in any historical or cultural context in which the separation of social activities into autonomous spheres is not the norm—beyond the binary perspective of autonomy and heteronomy. This would make it possible to consider political didacticism and plot construction, state policy and thematic choices, cult of personality and writing style not as elements to be disentangled but as fundamentally intertwined and equally integral components of literary evaluation. This does not imply that what is valued in authoritarian literary systems is necessarily radically different from other literary systems. North Korean literature is not the absolute other of literature in liberal capitalist countries; similar aspects of literary writing could be valued in both systems, but according to different modalities and in different proportions.

In order to map North Korean literary standards, this article relies on an approach that may be termed "autonymy."[4] This is not a new, alternative theory of literary value but rather a methodology for understanding existing systems of literary evaluations and how they distinguish between good and bad literature. I do not posit a priori that the criteria for this distinction are radically different—or similar—from criteria used elsewhere. Rather, I intend to map these criteria either from scratch or through hypothesis testing, based on what is labeled good or bad literature in North Korea. This approach is therefore fundamentally nominalist in that it rejects abstract universals (Quine 1953) and asserts that properties such as "literary quality" exist in name only (there is no universal standard of literary value). It is also pragmatist in that considers the meaning of a word to be determined by its use in language within a certain cultural context (Wittgenstein 1953, 20): what is good North Korean literature is what North Koreans call good literature. Therefore, autonymy does not offer a definition of literary value, but merely acts as a framework to understand how literary value is defined by the agents of a specific literary system. It does not seek to give a definite, comprehensive exposition of literary standards, but rather favors induction from a set of necessarily partial, but empirically grounded, observations.

Such an approach necessarily means prioritizing primary sources from within the DPRK. But it also implies favoring certain sources over others, as the question of what constitutes a valid, reliable label of literary quality is far from straightforward. A traditional approach in literary studies would be to turn to criticism, literary theory, and writers' essays on the nature of their craft. Yet such sources are usually normative or aspirational and therefore may not reflect the effective standards of the field. A classic example, from Korean literary history, is novelist Lee Kwangsu, who condemned literature's submission to morality and the cheap sentimentality of popular literature in his essays (Lee 1916), all the while publishing didactic, melodramatic fiction to popular and critical acclaim. In the context of North Korea, the strong incentive to appear to follow or to prescribe ideological orthodoxy likewise means that the theoretical definitions of literary value put forth by critics are quite removed from

[4]I use the term autonymy (i.e., "self-naming") specifically for the purpose of this study and with no reference to previous usages of the term in other fields such as linguistics (Barthes 1964; Carnap 1964) or narratology (Dällenbach and Tomarken 1980).

https://doi.org/10.1017/S0021911821002278 Published online by Cambridge University Press

the actual practices of the field. It is nonetheless possible to address this issue by focusing on nontextual symbols of literary value—literary prizes, institutional appointments, publication in more or less prestigious venues, commendation by the Party or the Leader—over textual ones. Indeed, these multiple acts of designation constitute realized instances of literary judgment. Textual sources, of course, need not be dispensed with entirely. For instance, when other forms of data are scarce or inaccessible—a common conundrum in the study of North Korea—literary criticism might be the only source from which to infer the state and rules of the literary system.

There would seem to be a natural methodological affinity between such an approach and the practice of distant reading (Moretti 2013, 48; 2015, 688), which aims to eschew "direct textual reading" and underline the primacy of "abstract institutions." Yet, as critics have pointed out (Goldstone 2015; Merriman 2015), distant-reading-based scholarship has still mostly concerned itself with textual properties rather than the "the institutional matrix that actually produces and circulates them." Conversely, this article is mostly concerned with understanding a literary system and its economy of literary value, rather than producing knowledge about the specificities of individual texts or authors. Furthermore, the context in which distant reading was originally conceived—the study of world literature—presupposed the universality and transparency of the notion of "literature" and its associated lexicon. Franco Moretti's famous and oft-debated (Beecroft 2014, 182; Milner 2005, 136; Spivak 2003, 107–9) conclusion that all modern novel forms are "variations" from an "Anglo-French core" (Moretti 2013, 35) is here exemplary. The observation is true, but tautological and normative (since foreign texts too different to be considered "variations" will be excluded), and it is precisely the pitfall that the methodological decision to study a literature on its own terms—by focusing on autonymy and the way things are labeled in their original context (here *munhak* 문학[5] as a category corresponding to literature)—is trying to avoid in the first place.

Distant reading has also become shorthand for studies that work on a large scale using computation—although, as Ted Underwood (2017) points out, neither large corpora nor computational methods are necessarily specific or inherent to it[6]—and this work is certainly indebted to the quantitative studies done at the Stanford Literary Lab as well as to other traditions of computational criticism. As the goal of the present study is to map the dynamics of an entire literary system, a large scale of analysis seems a prerequisite. It may be possible to reach similar conclusions through the study of a more restricted, representative corpus, as is often done in traditional cultural studies. But such small corpora present a higher risk of selection bias and their actual representativeness is hard to evaluate. A larger amount of data, on the other hand, minimizes the risk of relying on anecdotal evidence and lets us evaluate the statistical significance of our observations. Computational methods, in turn, allow us to work efficiently at such a scale by automating the labor of reading and counting necessary to draw conclusions from large volumes of textual data.

Data-driven approaches are not, of course, a panacea. As Judea Pearl and Dana Mackenzie (2018, 22) remind us, data are "profoundly dumb," and at least until more sophisticated causal models become commonplace, the task of interpretation and explanation still falls to the human scholar. But the idiocy of "offensively simpleminded" (So 2017, 668) computer models may also present an advantage for the analysis of literatures that are understudied and easily subject to cultural bias: first because they require framing questions simply, as hypotheses whose validity is easy to test and contest and can substitute

[5]This is not a linguistic but a cultural issue: what is considered *munhak* in the North is sensibly different from the South's conception of it, even though both countries use the same term. Attempts to develop concepts that would encompass both literatures (e.g., "national literature"/*minjok munhak* 민족 문학 or "Unification literature"/*ton'gil munhak* 통일 문학) only illustrate the reality of that gap.

[6]It is worth noting here that several macro-surveys of North Korean literature have been conducted prior to the advent of distant reading. In the first study of North Korean literature published in English in 1977, for instance, Marshall Pihl compiled a corpus of short stories published between 1959 and 1973 and used it to generate statistics on literary production by genre and period. Computational approaches are rarer but not unheard of, including a comparative survey of the epic poet Cho Kich'ŏn published in South Korea (Kang 2003) and, in the North, a work of computational linguistics that happens to rely on (among others) a corpus of literary texts (Mun 1993). However, both works were entirely descriptive, underlining the need to couple computational methods with well-defined research questions and a critical outlook (Broadwell, Chen, and Shepard 2019; Underwood 2017).

https://doi.org/10.1017/S0021911821002278 Published online by Cambridge University Press

empirical validation to intuitive induction, and second because this idiocy is also a great revealer of idiosyncrasies. Computer models are "stochastic parrots" (Bender et al. 2021) that reproduce and amplify the cultural specificities and biases of the data it was trained on. While this naive form of data mimicry has deleterious effects in practice, it also offers an unfiltered picture of a society's discriminatory actions —of what is deemed normal, valuable, or original and what is not.

Politics and Literary Prestige in Early North Korean Literary History

Defined by Kim Jong Il as an "an ideological weapon under the control of the Party" (Kim 1992b, 14), North Korean literature has been a political endeavor ever since its inception after Korea's liberation from Japan in 1945. In 1946, the North Korean Writers' Union (Chosŏn Chakka Tongmaeng, 조선작가동맹) was established as the country's sole literary organization under the tutelage of the Party's department of agitation and propaganda (Shin 1988). Now state employees, writers received guidelines from the Party instructing them to produce works on current political issues, from labor laws and land reform to alphabetization campaigns. Apolitical writing was rare to begin with—most North Korean authors had been active members of the Korean Artists Proletarian Federation (KAPF)[7] during the colonial period—and increasing censorship and public condemnations quickly rooted it out entirely. By the end of the 1940s, virtually all published literature contained an ideological message in line with the Party's current policies, while dissident writing remains unheard of to the present day (Immanuel Kim 2012).

The complete politicization of the field, however, did not mark the end of debates regarding what literature was or ought to be. Different factions continued to vie for prestige. The main one, composed of former KAPF writers such as Han Sŏrya and Yi Kiyŏng, had come to fame during the colonial period. Dominating the country's artistic institutions, their aesthetic strategy was, first and foremost, a patrimonial one, aiming to preserve their legacy by designating their own works from the colonial period as the standard to be emulated (An 1946; Myers 1994). Their main competitors were a group of Soviet-Korean Red Army officers and their young North Korean protégés. Their literary careers either nonexistant or unknown in the peninsula in 1945, the Soviet-Koreans built their claim to artistic legitimacy by leveraging the literary capital of the Soviet Union and presenting themselves as proponents of orthodox Soviet socialist realism, upholding the "noble aesthetic tradition of the developed country of the USSR" (Min [1949] 1993, 384; see also Kang 1993; Lee 1950). Their role in the Soviet cultural apparatus in the DPRK also helped them secure literary prestige by giving them the financial resources to establish the country's first literary prize—a legitimizing symbol that they could, and regularly did, award to members of their faction to build their reputation. Other smaller groups—for instance, writers such as Im Hwa who were affiliated with the Southern Communist Party and its leader Pak Hŏnyŏng—defended their own literary positions, but regardless of their size, all groups mapped to political forces within the DPRK.

The field became more homogenous after the Korean War, as power was concentrated in the hands of Kim Il Sung and most literary factions lost their political patrons. The Southern communists were purged at the end of the Korean War and their representative writer, Im Hwa, was executed (Shin and Oh 2000, 164–67), while the Soviet-Koreans were purged along with Maoists at the end of the 1950s (Lankov 2002). Former KAPF writers were now in a monopolistic position, in control of the field's entryways and responsible for the training of new and aspiring writers. But despite their newly gained authority on literary matters, writers were now subject to much more Party supervision than they had been before the war. By 1967, Kim Il Sung had removed his remaining political rivals and called for a cultural revolution to secure the establishment of a monolithic ideological system within the DPRK

[7]"Korea Artista Proleta Federacio," the KAPF was the main Korean proletarian artistic organization from 1925 until its dissolution in 1935. Not unlike North Korean literature, the KAPF's output was long dismissed (and censored) as exhibiting little artistic quality even by its own former members. Recent efforts to rehabilitate the KAPF's legacy, for instance, by Sunyoung Park (2015, 11), have focused on the group's contribution to Korean intellectual history and its ability to articulate an original socialist thought but have explicitly avoided questions of literary value.

https://doi.org/10.1017/S0021911821002278 Published online by Cambridge University Press

(Person 2013). His son, Kim Jong Il, who was appointed head of the arts and letters section of the Party's agitation and propaganda department, was in charge of outlining what the revolutionary literature of this new political era ought to be (Kim [1967a] 1987).

The transformative effect of the events of 1967 on the literary field can hardly be overstated. In June, Kim Jong Il overtly criticized the "anti-Party anti-revolutionary elements and their followers" who insisted on "upholding the 'tradition of the KAPF'" instead of the "glorious tradition of revolutionary art and letters established by the Great Leader during the time of armed resistance to the Japanese Empire" (Kim [1967b] 1994, 242). The literary tradition of the KAPF represented more than an aesthetic position. It was a reminder that other resistance movements to Japanese colonization had existed besides Kim Il Sung's, and a competing historical narrative with the potential to cast shade on Kim's claims of being the sole defender and liberator of colonized Korea. The "glorious tradition" that was to replace it, on the other hand, was indissociable from the figure of the North Korean Leader. Originally a loose and largely oral compilation of music, theater, and tales produced in Manchuria during the 1920s and 1930s that focused on the activities of local partisans, the new revolutionary classics would be adapted to several media forms (such as revolutionary operas, cinema, or novels), and their authorship, often uncertain, attributed to Kim Il Sung himself (Kim 1994, 200–211; Shin and Oh 2000, 290). Upholding the legacy of this new canon meant writing "Leader literature" (*suryŏng hyŏngsang munhak* 수령형상문학) that represented the ruling Kim family and their historical role in the resistance to the Japanese colonizer and the construction of a revolutionary Korea.

Kim Jong Il's pronouncements redefined literary value—effectively pegging it to the figure of Kim Il Sung—and institutionalized it. Kim created a subgroup within the Writers' Union called the April 15 Literary Production Unit (4.15 문학창작단, hereafter April 15 LPU), which was tasked with producing a series of hagiographic novels depicting the revolutionary exploits of the North Korean Leader. The unit was staffed with the country's most talented writers, and being nominated to it constitutes, to this day, the highest sign of literary recognition in the DPRK (Immanuel Kim 2016; Paek 2017). While the unit's writers are free to write on topics of their own choosing, they have an obligation to produce works, and particularly full-length novels, on the lives of the DPRK's ruling family. These works are collected in a series called *Immortal History* for novels about Kim Il Sung, *Immortal Guidance* for novels about Kim Jong Il, and, recently, *Immortal Journey* for Kim Jong Un. All of the works in these collections hold a canonical position in the DPRK's literary pantheon.[8]

The cultural revolution of the late 1960s therefore seems to have rendered the issue of literary quality straightforward: good literature is Leader literature, and good writers are those who write it. But that simplicity is misleading. That writers recognized as talented get tasked with writing about the Leaders implies that the recognition of their literary skills is based on their output prior to their appointment as specialized hagiographers—that is, on writings that may have little to do with the Leaders. More generally, the majority of North Korean literary production comprises texts that do not deal with the cult of personality of the Kim family. How do we account for these works in the general economy of literary value in the DPRK? Furthermore, equating quality with cult of personality tells us something about the content of literature, but little about its form. How does the style of canonical hagiographical novels compare with other forms of literary writing? Finally, as the "immortal" epithet of the collection of Leader literature indicates, the North Korean canon presents itself as the embodiment of ahistorical, immutable aesthetic standards. But this is an aspiration, a claim, rather than an observed reality. One is reminded here of Andre Schmid's (2018) observation about the risk of reproducing features of North Korean propaganda within one own's analytical framework by taking its normative claims at face value. Rather than consider aesthetic standards to have been made immutable, is it possible to historicize them and perhaps detect trends and changes in their application?

[8]Throughout the rest of the article, I use the expression "*Immortal* series" to refer to the works published in these collections, while the term "Leader literature" is used to refer to the whole genre of writing about the Leaders' lives (in novels, poetry, or short stories). The term "canonical novels" is used later in the article to refer to a specific corpus of novels drawn from the *Immortal* series, and it is used as a representative sample for statistical analysis. Put differently, "canonical novels" are a subset of the "*Immortal* series," which is itself a subset of "Leader literature."

https://doi.org/10.1017/S0021911821002278 Published online by Cambridge University Press

A writer's admission to the April 15 LPU gives us some potentially exploitable information. Like a literary prize, it signals that a writer has achieved a certain level of recognition for the literary quality of their work. While this fact, in and by itself, provides insufficient ground for analysis, it can nonetheless be cross-referenced with a quantitative survey of these writers' outputs throughout their careers. Thus, by leveraging a single marker of literary quality, one can trace the careers of writers who achieved literary fame, map the themes that brought them recognition, and question their propensity to conform to or depart from the literary standards of their day.

Genre and Literary Recognition: Topic Modeling Forty Years of North Korean Literature

North Korean literary prose falls within a limited number of thematic genres, such as the production novel dealing with workers and factory life, the agrarian novel about collective farming, or the Leader literature depicting the deeds of the Kim family.[9] With the exception of Leader literature—explicitly considered the highest form of art—genre is only a descriptive concept in North Korean literary criticism. The quality of a work is instead assessed through more abstract criteria such as "relevance to contemporary times" (*hyŏndaesŏng* 현대성), "veracity" (*chinsilsŏng* 진실성) or "party spirit" (*tan'gsŏng* 당성).[10] This, in turn, would suggest that literary quality in the DPRK is agnostic to genre and that works are evaluated solely based on the same ideological and aesthetic features regardless of their thematic content. But these terms are far from unambiguous, and relying on them to understand literary quality may tell us more about what some critics think literature should be rather than what is actually valued.[11] Following the methodology outlined earlier, I opt to overlook critics' opinions in a favor of a more empirical approach. Instead, taking inclusion in the April 15 LPU as a marker of literary quality, I compare the generic breakdown of the output of writers from the April 15 LPU over their careers with the production of other writers to see whether there are any discrepancies that could hint at certain genres carrying more or less prestige than others.

To do so, I compiled a corpus of 1,492 short stories published in the DPRK's main literary journal, *Chosŏn Munhak*, between January 1977 and December 2016. Short stories are an important element of the North Korean literary field. They are the format through which writers make their literary debut and account for the majority of writers' early-career output. Senior writers, too, regularly produce short stories for literary magazines or anthologies, even if they have already published several full-length novels. The corpus was obtained from a DVD commercially sold in the DPRK that contains digitized copies of the journal and a reader application.[12] To break down the corpus by genre, I use topic modeling, a set of algorithmic methods that can help "discover hidden thematic structure in large collections of texts" (Blei 2012). Topic modeling generates topics, or clusters of words likely to appear in the same context, without human supervision.[13] The topics are simply automatically

[9]While genre certainly entails other features besides theme, the two are sufficiently linked in North Korean literature that the latter can be used as a proxy for the former.

[10]See, for instance, Kim Jong Il's *Theory of Juche Literature* or Kim Il Sung University's *Treaty of Juche Literature*, in which most of the discussions focus on these concepts, some of which can be traced back to Soviet socialist realism. Only a few enumeratory paragraphs are dedicated to the question of genres (*kallae*)—save for Leader literature, "the supreme task of our literature" (Kim 1992a, 1; Ri 1996).

[11]Myers (1994, 41, 52, 81), for instance, describes how the critical vocabulary of Soviet socialist realism was mobilized to praise or criticize the work of other writers along factional lines and with little relation to the actual content of the works.

[12]The content was extracted and decrypted using reverse engineering. Code for the extraction algorithm is available, along with the rest of the code used for this article, at https://github.com/anonymousjournalsubmission/nk_code.

[13]Topic modeling is an unsupervised machine learning method, meaning that its output categorization—topics—is not specified a priori but inferred from the data. Likewise, the words associated with each topic are extracted without prior human input. By contrast, a supervised approach would require to manually annotate a subsample of the data set by associating each story in it with one or several predetermined genres then use these examples to train a model to predict the genres of unseen stories. Supervised approaches have the advantage of offering clearer evaluation metrics (one can, for instance, estimate the proportions of misclassifications) but are less suited to the discovery of unknown patterns. Supervised methods also require the definition of an a priori taxonomy and a necessarily subjective annotation process, which offer additional avenues for introducing biases compared to an unsupervised approach. For a recent example of an approach mixing both supervised and unsupervised methods to investigate the question of genre, see Antoniak, Walsh, and Minno (2021).

https://doi.org/10.1017/S0021911821002278 Published online by Cambridge University Press

generated collections of related words, as well as their associated probabilities, and it is up to a human interpreter to determine their common underlying themes and give them appropriate labels. Running the algorithm on the corpus produces the seven topics listed in table 1, along with their twenty most representative words.[14]

The topics generated largely map to genres manually identified in a previous survey of short stories published in *Chosŏn Munhak* during the 1980s (Choi 1994).[15] They show little overlap, save for topics 2 and 4, which pertain to the DPRK's ruling family and, for the rest of the present analysis, are merged under the label "Leaders" (narratives about each Leader are distinctive enough thematically for the algorithm to group them in separate clusters, but the distinction is of little relevance to our present analysis). It must be noted that topics do not exactly correspond to literary genres per se. In topic modeling, a document is always a mix of different topics. For instance, a short story about a soldier writing to his family about his experience fighting alongside Kim Il Sung might be broken down as being 50 percent "Military," 30 percent "Leaders," and 20 percent "Family." To get from topic to genre, I make the assumption that the topic most prominently represented in a short story is its genre.

Having thus labeled the topics and assigned a genre to each story, one can plot the proportions of each genre in the corpus (figure 1) as well as their diachronic variations over the period considered (figure 2). This first overview of the corpus confirms some of the intuitions outlined earlier. The 20 percent share of Leader literature highlights the importance of looking at North Korean literature beyond the cult of personality in order to account for the remaining 80 percent of literary production. The most common genre in the corpus, education, is almost entirely absent from academic research on North Korean literature. Such a discrepancy may be explained by the scholarly practice of using North Korean literature as source material for the study of internal and external politics. Stories that can be used to directly illustrate Party directives, agrarian reforms, or military policies seem to gather more scholarly attention than works about learning and studying. This, in turn, reinforces the idea that North Korean literature is predominantly concerned with these themes. On the other hand, the significant changes in generic trends over time underscore the need for historical perspective and proper contextualization. For instance, only a decade after the instauration of Leader literature, the genre was almost dormant, but it regained a dominant share in the last three decades, with pronounced peaks at times of national crises (the famine of the 1990s, the breakdown of the Agreed Framework and economic sanctions in 2002, and the death of Kim Jong Il and transition of power to Kim Jong Un in 2010).

These insights, however, are merely quantitative and do not tell us anything about the different degrees of literary prestige attached to each of the genres. That "Education" is the most commonly represented genre in the corpus may tell us something about its importance to Party officials who set production quotas, but that does not necessarily imply that the genre is particularly highly regarded within the literary field. In order to examine the relationship between genre and prestige specifically, I selected twenty-six authors known to be members of the April 15 LPU (detailed list available in appendix 1) and created a set of the short stories they published "pre-recognition" (i.e., before being admitted to the group; 62 stories), as well as a set of stories published "post-recognition" (46 stories). The goal is to see whether the thematic structure of their literary production evolves before and after they achieve literary recognition and how it compares with that of non-LPU writers.[16] To make sure these differences are not products of random chance (as the proportion of each topic in the subset of short stories written by recognized authors might actually be within the range of variation that could be expected of any

[14]The words are ranked by their probability of occurring in each topic. Details of the preprocessing, algorithm, and hyperparameters are available in the article's code repository. The number of topics was selected iteratively to avoid junk topics, limit redundancy, and optimize granularity.

[15]The six main categories used by Choi (1994) in his taxonomy were "Kim II sung and his family," "rural community development," "economic and industrial development," "advancement of science and technology," "civil-military relations," and "beautiful people." Four other minor categories were mentioned: "overseas Koreans," "history," "South Korea," and "unification" (Choi 1994).

[16]I rely on the data compiled by Kang Chinho and Imha Yu (2009) for the admission dates. When no data were available for a specific author, I used the publication date of their earliest known work published in one of the *Immortal* book series.

https://doi.org/10.1017/S0021911821002278 Published online by Cambridge University Press

Table 1. Topics and their most representative words

Topic #	Topic Label	First Twenty Keywords (in order of descending importance)
1	Education	Father, teacher (honorific), teacher, young woman, husband, student, Korea, sea, youth, Japan, motherland, school, research, school employee, university, university professor, wife, tree, grandfather, letter 아버지, 선생님, 선생, 처녀, 남편, 학생, 조선, 바다, 청년, 일본, 조국, 학교, 연구, 교원, 대학, 교수, 안해, 나무, 할아버지, 편지
2	Leaders (1)	General, comrade (honorific), Kim Jong Suk, speak (honorific), Kim Jong Il, Kim Il Sung, people, mother (honorific), commander, respect and love, best, army, Kim Jong Un, give (honorific), revolution, liberation, factory, Japanese (derogatory), great, squad 장군, 동지, 김정숙, 말씀, 김정일, 김일성, 인민, 어머님, 사령관, 경애, 최고, 일군, 김정은, 드리, 혁명, 해방, 공장, 왜놈, 위대, 부대
3	Family	Mother, father, son, mom, husband, grandmother, hometown, village, uncle, kid, child, mother (honorific), letter, little brother, big brother, female proper noun, grandfather, male proper noun, kid, school 어머니, 아버지, 아들, 엄마, 남편, 할머니, 고향, 마을, 아저씨, 자식, 아이, 어머님, 편지, 동생, 오빠, 옥이, 할아버지, 철이, 애가, 학교
4	Leaders (2)	Grand leader, speak (honorific), Kim Jong Il, parents, comrade (honorific), people, great, leader, army, dear, comrade leader (honorific), dear, give (honorific), revolution, committee leader, adjutant, motherland, history, car, administration 수령, 말씀, 김정일, 어버이, 동지, 인민, 위대, 지도자, 일군, 친애, 지도자동지, 친애하는, 드리, 혁명, 위원장, 부관, 조국, 력사, 승용차, 관리
5	Industry	Manager, factory, workplace, technician, production, technic, comrade (honorific), plan, Party secretary, coal mine, secretary, young woman, equipment, mine, crane, youth, construction, blueprints, worksite, worker 지배인, 공장, 직장, 기사, 생산, 기술, 동지, 설계, 당비서, 탄광, 비서, 처녀, 설비, 광산, 기중기, 청년, 건설, 도면, 현장, 일군
6	Agriculture	Old man, group leader, agricultural sub-workteam, committee leader, administration, agricultural sub-workteam chief, farm, workteam, young girl, farmwork, secretary, tractor, seedbed, amateur art troupe member, farmworker, fertilizer, tree, manure, village, corn 아바이, 반장, 분조, 위원장, 관리, 분조장, 농장, 작업반, 처녀, 농사, 비서, 뜨락, 모판, 소조원, 농장원, 비료, 나무, 거름, 마을, 강냉이
7	Military	Company leader, young girl, platoon leader, company, comrade (respectful), warrior, squad leader, soldier, platoon, unit leader, battle, unit member, war, hill, division, commander, frontline, officer, soldier, shock brigade 중대장, 처녀, 소대장, 중대, 동지, 전사, 분대장, 병사, 소대, 대장, 전투, 대원, 전쟁, 고지, 분대, 사령관, 전선, 군관, 군인, 돌격대

subset of the general corpus), I use a permutation test with a significance level of 90 percent because of the small sample sizes (Bradley and Tibshirani 1994; Jockers and Mimno 2013).

The results are presented in figure 3 for short stories published before the writers' admission to the LPU and in figure 4 for short stories published afterward. The statistically significant differences are the overrepresentation of Leader literature output post-recognition (after writers joined the April 15 LPU) and the overrepresentation of industrial production stories and military stories pre-recognition. The high prevalence of the "Leaders" topic post-recognition is unsurprising: writers from the April 15 LPU are, after all, in charge of producing the full-length biographic novels of the *Immortal* series. The data nonetheless show that this duty extends to the form of the short story, and that writers still produce a large amount of works that do not focus on the Kim family after being accepted into the group (about 40 percent of their output). More importantly, this reveals that the writing of works praising the Leaders is more a symptom of literary recognition than a way into it. The proportion of works dedicated to the Leaders is roughly the same for LPU writers pre-recognition than it is for writers overall, meaning that one does not gain recognition by specializing in the production of cult of personality stories. Rather, the specialization comes as a consequence of being a famous writer.

https://doi.org/10.1017/S0021911821002278 Published online by Cambridge University Press

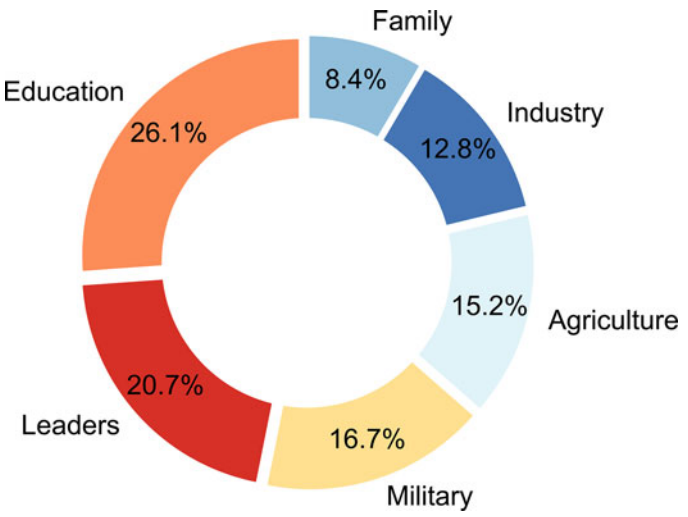


Figure 1. Distribution of genres in the short stories published by *Chosŏn Munhak*, 1977–2016.

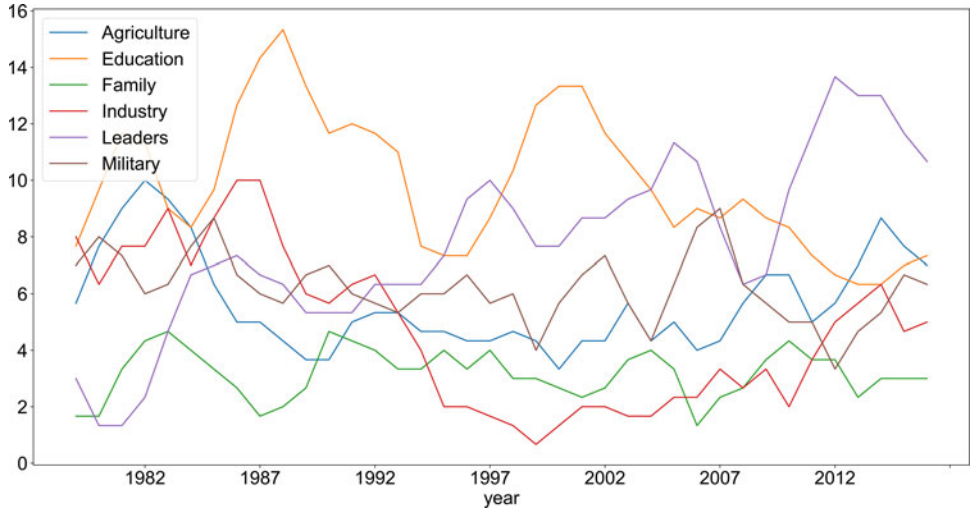


Figure 2. Yearly number of short stories published in *Chosŏn Munhak* for each genre considered (three-year moving average smoothing).

On the other hand, the fact that writers who were eventually admitted to the LPU published significantly more stories about the military and the industrial working class indicates that these two genres may be more highly regarded and more likely to lead authors to literary recognition. The valuation of these genres intersects with cultural and ideological hierarchies of class and gender. At least half the writers considered worked on industrial sites or served in the military prior to their literary career. Furthermore, writers from the working class have historically enjoyed more opportunities to start a literary career than agrarian workers or service sector employees. As social groups and professional activities, both the military and the working class have been predominantly gendered as male in textual and visual representations.[17] The patriarchal structure of North Korean society may thus skew literary capital distribution toward themes that are predominantly male in both their content and authorship (every member of the April 15 LPU analyzed here is male).

[17]For an exploration of these differences, see, for instance, https://digitalnk.com/blog/2017/12/23/north-and-south-korea-through-word-embeddings/ and https://digitalnk.com/blog/2017/09/30/gender-distribution-in-north-korean-posters/.

https://doi.org/10.1017/S0021911821002278 Published online by Cambridge University Press

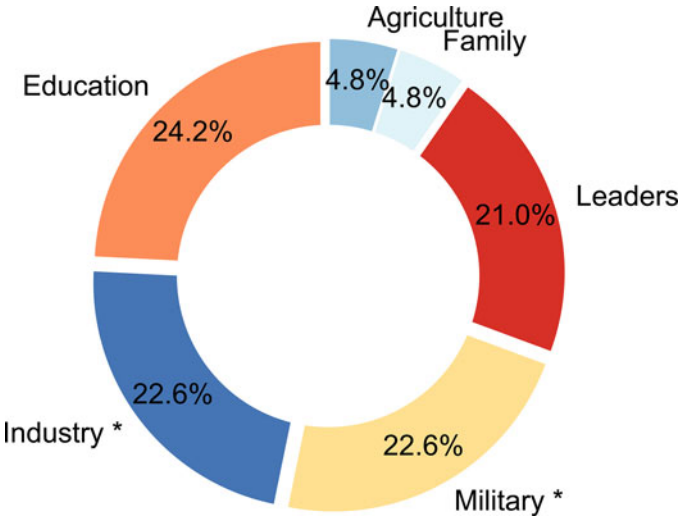


Figure 3. Distribution of genres in short stories of writers from the April 15 LPU *before* their admission to the group. Statistically significant differences with non-LPU writers are marked with an asterisk.

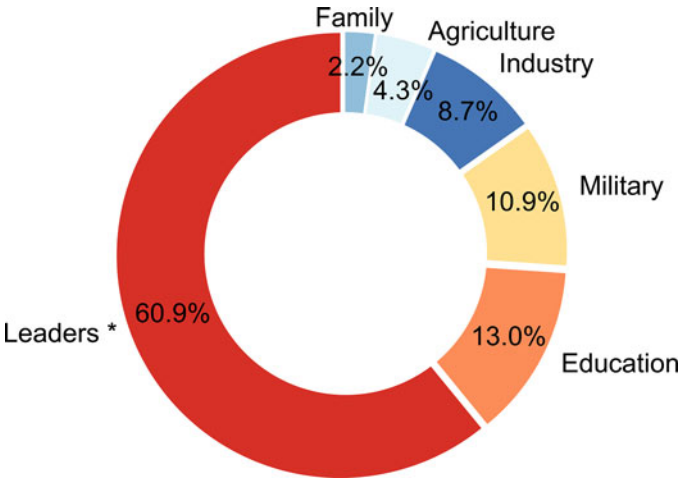


Figure 4. Distribution of genres for short stories of writers from the April 15 LPU *after* their admission to the group. Statistically significant differences with non-LPU writers are marked with an asterisk.

These findings, however, lack a diachronic perspective. The oldest members of the April 15 LPU joined upon its creation in 1967, and the most recent member in the data set joined in 2015. While literary standards have been shown to change slowly in other Western contexts (Underwood and Sellers 2016), the same can not assumed to hold true in North Korea, where writers' interconnectedness with politics may put them on a time scale closer to that of a quinquennial plan. To get a sense of this change, I divide the group of prestigious writers into an "old guard" and a "young guard," based on the date they joined the April 15 LPU. I set the threshold date for the separation between the two groups at 1991, as the end of the Cold War outside the Korean peninsula marked the beginning of a period of radical change for the DPRK (Lankov 2014, 77–111). Like before, only their output prior to their admission to the April 15 LPU is considered.

Comparing the two groups (figures 5 and 6), two genres—agrarian and military fiction—show statistically significant differences. While some of the younger generation of elite writers tried their hand at agrarian fiction in their early careers, the genre was completely ignored by the older generation. Both groups still produced less agrarian fiction (although the difference lacks sufficient statistical

https://doi.org/10.1017/S0021911821002278 Published online by Cambridge University Press

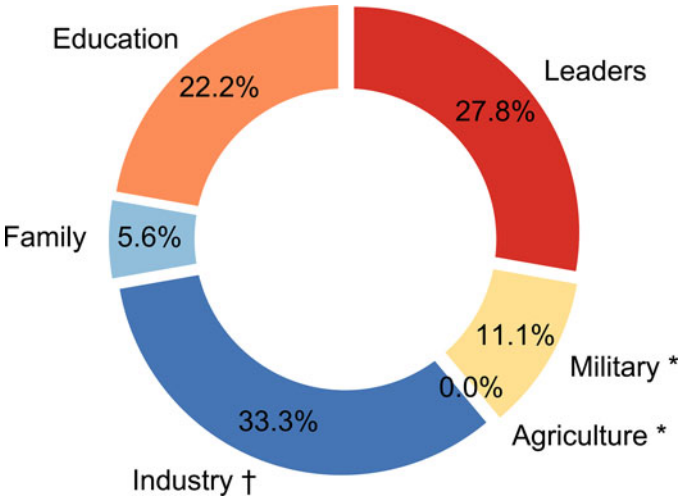


Figure 5. Distribution of genres in short stories of "old guard" writers *before* their admission to the LPU ($N = 18$). Statistically significant differences with the "young guard" are marked with an asterisk. Statistically significant differences with the output of non-LPU writers before 1991 are marked with a cross.

significance to be considered conclusive) than regular writers, suggesting that the genre could carry less prestige—perhaps a reflection of the traditionally low status accorded to the peasantry in the DPRK.[18]

The difference in the proportion of military fiction, higher by a factor of two among younger writers, shows the rising value of the genre over the last three decades. Young writers admitted to the April 15 LPU produced significantly more military fiction than writers who never joined the LPU, indicating that more literary prestige is attached to this particular genre. The rise of the military genre can be explained by the advent of "military first" (*sŏn'gun* 선군) politics in the 1990s. Following the tremendous social, economic, and political upheaval caused by the combined events of the dissolution of the USSR, the death of Kim Il Sung, and the famine, Kim Jong Il ushered in a new ideological line prioritizing defense and the military to ensure the survival of the regime (Smith 2015, 235). The new policy made the military a demographically larger and politically crucial social class, eventually supplanting the working class as the "main force of the revolution" (Kim 2007, 1).

The declining political weight of the working class seems to be reflected in the decreasing volume of industry-themed short stories from the end of the 1980s onward. Yet despite this quantitative slump, there is still a hysteresis effect in the amount of literary prestige associated with the genre. Indeed, young writers admitted to the April 15 LPU still wrote industrial production stories on a much larger scale than regular writers. While they did write less industrial fiction compared with older members of the April 15 LPU, the difference, although sizable, is not statistically significant,[19] which makes it impossible to confidently conclude that the stock of the genre may have eroded between the two generations.

That so much of the relationship between literary genre and prestige revolves around social classes raises the question of whether class is really just a confounding variable. As previously noted, most of the April 15 LPU members came from either the working class or the military. And, more generally, because of the realist framework governing literary production in the DPRK, writers tend to produce stories that reflect their social background. As both the military and the working class have historically been politically key social groups in North Korea, the higher prestige conferred to military and industrial fiction in the literary field may simply reflect the privileges enjoyed by these two classes. One

[18]Although in theory, the North Korean *juche* ideology emphasizes that, unlike Marxism, it does not consider the working class as the sole leading agent of the revolution, but rather the popular masses (*inmin taejung* 인민대중) as a whole, in practice, the country has always relied on a hierarchical class system. While this hierarchy has changed throughout history, with the army taking over the working class's leading role in the 1990s, the peasantry, as a class still needing to be "working-classized" (*rodong kyegŭphwa* 로동계급화) or "revolutionized" (*hyŏngmŏnghwa* 혁명화), has consistently remained at the bottom (Kim Chaeung 2012; Kim Chinhwan 2016, 241).

[19]The statistical significance of a difference (whether it is due to chance or not) is different from its magnitude.

https://doi.org/10.1017/S0021911821002278 Published online by Cambridge University Press

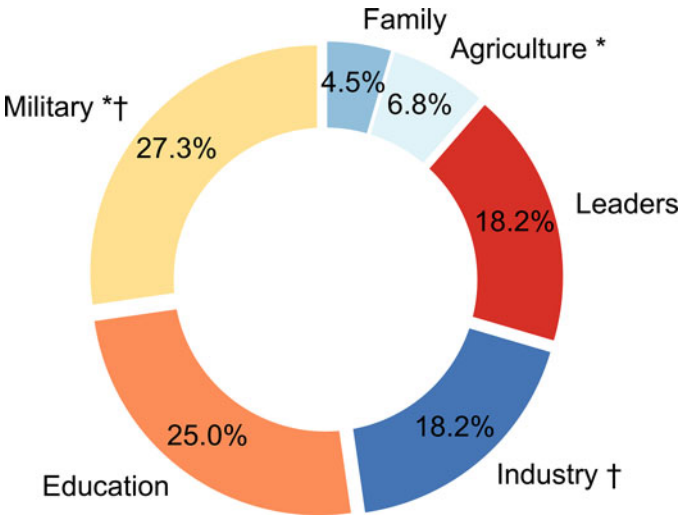


Figure 6. Distribution of genres in short stories of "new guard" writers *before* their admission to the LPU ($N = 44$). Statistically significant differences with the "old guard" are marked with an asterisk. Statistically significant differences with the output of non-LPU writers after 1991 are marked with a cross.

might see this as yet more evidence of the heteronomy of the North Korean literary field and its lack of purely literary standards. But social determinations weigh upon all literary fields, even those traditionally considered autonomous, and the brokers of literary value (critics, publishers, or prize committees) that inhabit them. For instance, as Richard Jean So (2020) has demonstrated, the simple factor of whiteness remains the best predictor of literary success in postwar American literature, despite an allegedly autonomous critical apparatus. What seems more necessary, therefore, is to highlight the institutional, systemic mechanics by which social forms of discrimination are translated into the language of literary standards.

Modeling Stylistic Originality

Literary consecration through admission to the prestigious April 15 LPU comes with a mandate to produce works featuring the Leaders. This task has been described as forcing writers to "sacrifice their creativity as artists" (Immanuel Kim 2016) since it imposes a subject matter and requires them to follow specific conventions: ways of describing the Leaders, grandiloquent metaphors, plot conventions, or stock characters. It would then *seem* that the North Korean literary field operates in a manner that is diametrically opposed to the paradigm of modern Western literature. Whereas in the latter, originality, or the ability to distinguish oneself from the rest, is essential to the logic of aesthetic evaluation (Bourdieu 1979), in the former, it appears that quality is determined by conformism to existing writing patterns. This would conform to the characterizations of North Korean literature by foreign readers as boring and repetitive. Yet such a hypothesis cannot be taken at face value without being tested and investigated. That North Korean literature operates according to a different system of literary value does not necessarily imply that this system is a perfect negative image of the modern Western literary valuation system. The importance of conformity, for instance, does not necessarily prevent some creativity from being valued under certain modalities.

That all novels from the *Immortal* collections share roughly the same theme would certainly make them more homogenous than general literature and its multiple genres. But does this homogeneity extend to style? Are the canonical cult of personality novels of the *Immortal* collections really less stylistically innovative than the rest of North Korean literary production? As noted previously, Leader literature may represent the pinnacle of literary value, but it is not necessarily what writers are evaluated on before their consecration. To account for this, it is also necessary to consider the literary output of April 15 LPU writers besides their contributions to the *Immortal* collections and try to infer whether these works, too, exhibit a higher degree of conformity.

https://doi.org/10.1017/S0021911821002278 Published online by Cambridge University Press

To answer these questions, I begin with a formal definition of literary creativity as a deviation from a syntactic or lexical norm. This differential perspective is often used to define literary language in computational research (Louwerse, Benesh, and Zhang 2008; Van Cranenburgh, van Dalen-Oskam, and van Zundert 2019) and formalist theories of literature such as Roman Jakobson's (1960) poetic function or the concept of "foregrounding" (Fricke 2008). Such a definition is limited in that it does not account for narrative aspects of literary singularity, but it still offers a way to model and potentially measure the originality of writing style.

How to measure a text's deviation from linguistic norms? I propose a solution based on a probabilistic way to represent language known as language modeling. Language models are used to predict the probability that a certain word will appear based on the words that precede or (in so-called masked language models) surround it. For instance, a language model can assess that the word "literature" is more likely than the word "car" to follow the sequence of words "I want to read Korean" (Jurafsky and Martin 2019). Likewise, the same language model could tell us that the whole sentence "I want to read Korean literature" is more likely to be said by an English speaker than the sentence "I want to read Korean car." Traditional language models learn these probabilities by counting the frequencies with which words appear alongside each other within a very large corpus of texts.[20] These probabilities thus approximate likelihood that a sequence of words will be uttered by an "average" speaker of the language.[21] For the purposes of the present study, I used the deep-learning-based BERT language model (Devlin et al. 2019),[22] which I trained[23] on a corpus of roughly 4.5 million sentences extracted from various North Korean language sources (a general description of the corpus' content is given in appendix 2).

I posit that language models' ability to assign probabilities to sequences of words is similar to the concept of a norm against which to measure the deviations of literary language. Indeed, the probabilities returned by a language model line up with a basic model of literary creativity: they reflect lexical sophistication (given two synonyms, the rarer one will receive a lower probability) as well as the unexpectedness of word associations common in literary language. BERT has been shown to capture both semantic and syntactic aspects of language (Clark et al. 2019; Jawahar, Sagot, and Seddah 2019). One might say, following Jakobson, that its probabilities are computed following paradigmatic (predicting a word over others) and syntagmatic (based on its context) axes, which the poetic function subverts.

To illustrate this premise, let us take a sentence from a North Korean novel with a literary simile: "It was the jewel-like voice of little girls" (처녀아이들의 구슬 같은 목소리였소). A language model can predict the probability of the word "jewel" appearing given the words that surround it (a sort of fill-in-the-blank exercise). Doing so returns a certain probability for the word "jewel," and that probability will be higher compared with almost all other words in the Korean lexicon. Some words, such as "fire" (불), "storm" (폭풍), or "iron and stone" (철석), would be assigned slightly higher probabilities. This, however, is a feature, not a bug: expressions such as "a voice like iron and stone" or "a voice like fire" are simply more common clichéd similes in North Korean literature. A word producing an original, unique simile, on the other hand, would have be assigned a much lower probability. Instead of

[20]This description is meant to offer a basic intuition of how language models work, but it only applies to traditional language models. More recent models such as BERT rely on a more sophisticated attention mechanism that helps the model focus on important parts of the input data.

[21]The use of "average" here is, of course, highly contentious—a language model is ultimately a reflection of the texts it was trained on. These texts, typically Wikipedia entries, newspaper articles, websites, are really a subset of a written language determined by class, gender, and access to technology. For an overview of how these various biases become embedded in language models, see Blodgett et al. (2020).

[22]BERT can also be trained on a "next-sentence prediction task," and the probabilities given by the model that a certain sentence follows another have been used to experiment with assessing the predictability of fiction at the narrative level (Underwood 2020; Wilmot and Keller 2020). My approach derives from this insight but seeks to work at the level of sentences and style instead.

[23]The process of training an artificial intelligence model involves providing it with data to learn from. For instance, if one wanted an artificial intelligence model to distinguish between pictures of dogs and cats, one would need to compile a data set of pictures of both animals and then train the model to learn the differences between the two based on the examples provided in this data set.

https://doi.org/10.1017/S0021911821002278 Published online by Cambridge University Press

simply evaluating a single word, I aggregate the probabilities of different words within a sentence to compute an "originality score" for that sentence.[24]

The advantage of the method is not only that it provides a quantifiable measure of originality, but also that it allows us to avoid the subjectivity and biases that may come with having individual readers evaluate the originality of a sentence. The issue of bias is, of course, all the more complex in the case of North Korean literature, for which North Korean assessors would be hard to find. A reader familiar with South Korean language and culture but unfamiliar with North Korean literature may judge a sentence to be original simply because it is linguistically different from what they are used to. Or, conversely, they could judge a sentence to be overly trite because it sounds more stilted than what is common in South Korean literature. Furthermore, evaluations would differ because of individual proclivities, requiring the use of several readers and a way to aggregate their responses. On the other hand, a large language model acts as a "model reader" that would, thanks to its training corpus, be familiar with the North Korean language, but also with a good portion of its literary output.

Having thus constructed a way to estimate the stylistic originality of a sentence, it is now possible to attempt to score sentences taken from books of different levels of literary prestige to investigate the relationship between literary quality and creativity in the DPRK. To do so, I take three corpora of fiction:

1. Novels taken from the *Immortal History* (22 titles) and *Immortal Guidance* (16 titles) collections, hereafter referred to as *canonical novels*.
2. Novels written by members of the April 15 LPU on themes of their own choosing. These works do not depict the lives of the North Korean Leaders and therefore do not have the same canonical status as works from the *Immortal* collections. Their symbolic status is nonetheless still different from that of other regular novels due to the prestige of their authors. I will refer to these texts as *prestigious novels*.
3. Novels published by literary publishing houses in Pyongyang. These works are not of low quality, but their authors have not achieved the same level of recognition as the authors of the two previous corpora. I will refer to these texts as *regular novels*.

I also include a corpus of poetry and a corpus of articles from the Korean Central News Agency (KCNA) from the year 2000 to provide elements of comparison. None of the data were included in the data set that the language model was trained on. I sample 10,000 sentences from each of these five corpora and give every sentence an originality score. A low score means that the words in a sentence are highly predictable and therefore unoriginal. A high score means a less predictable and more original word choice. The results are given in figure 7 and table 2.[25]

That there would be little literary originality in KCNA articles compared with literary writing might seem unsurprising, but it nonetheless tells us something about literary language in the DPRK. Namely, even if conformity were valued in Leader literature, it is still conformity to literary standards—that is, to a language that differs significantly from the everyday language of a news dispatch. The poetry corpus, on the other hand, displays the highest average level of originality among the four literary corpora. However, one needs to distinguish here between the mean and median. While the mean score of poetry is comparatively extremely high, the median is roughly equal to that of prestigious novels. Furthermore, the 25th percentile of the poetry corpus is lower than two of the fiction corpora. This

[24]Because BERT is a masked language model, rather than a language model in the traditional sense, I rely on a pseudo-perplexity score based on the work of Salazar et al. (2020) with adaptations for the task at hand. For more technical details on the training of the model and the calculation of this measure, see https://digitalnk.com/blog/2020/10/01/language-models-literary-cliches-analyzing-north-korean-poetry-with-bert/.

[25]Pairwise differences in means are all significant at the 0.001 level except for the regular novels and canonical novels pair (significant at the 0.05 level). Pairwise differences in medians are all significant at the 0.001 level except between poetry and prestigious novels ($p = .5$). I use a t-test to assess the significance of pairwise differences in means and Mood's median test for differences in medians. The data are highly skewed and violate the t-test's normality assumption. However, Lumley et al. (2002) show that the t-test is "valid for any distribution" when the sample size is large, as in the present case.

https://doi.org/10.1017/S0021911821002278 Published online by Cambridge University Press

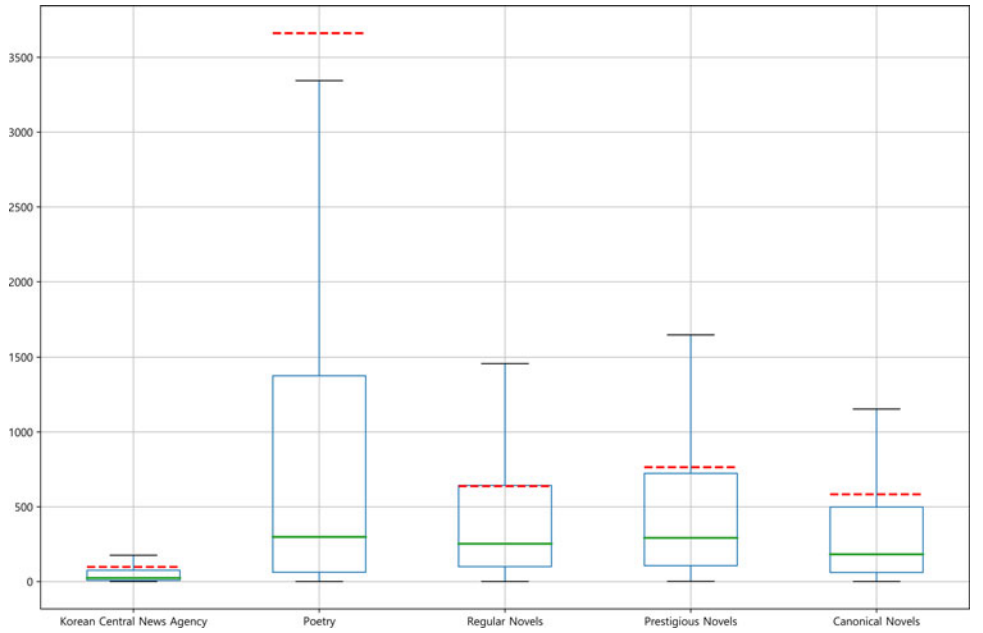


Figure 7. Boxplots of the originality scores for 10,000 sentences from each of the five corpora. The green solid lines are the medians and the red dotted lines the means.

indicates that while North Korean poetry does have some very original verses, it also contains a large amount of highly predictable ones.

Looking at the fiction corpora, one can see that the writing style of novels from the *Immortal* collections (canonical novels) is, on average, less original compared with other types of novels. The mean score is only slightly lower than that for regular novels, but the median is much lower and the overall distribution skews more heavily to the left, indicating that the corpus has fewer original sentences and more clichéd expressions. The corpus of prestigious novels, on the other hand, displays a significantly higher level of originality than the two other corpora of fiction. The writers are all members of the April 15 LPU and have published titles in the *Immortal* collections, but when freed from the stylistic and thematic constraints imposed by the collection, their writing style becomes more creative. These writers are also more creative than regular writers. This could mean that admission to the April 15 LPU is predicated upon literary creativity rather than a propensity to write on the themes favored by the Party. If this were true, the structure of the North Korean literary field might not be so different from that of many other literary cultures where creative avant-gardes eventually petrify into rigid academies (Bourdieu 1992). Another possibility, however, is that acceptance into the April 15 LPU gives writers more leeway to experiment, a phenomenon that has previously been observed in the realm of visual arts (De Ceuster 2011).

As almost all of the novels available in the corpus were published after their author's admission to the April 15 LPU, I am unable to test which of these hypotheses might be correct. Nam Taehyŏn's *Hymn of Youth* (청춘송가) is the only novel in the sample to be published before its author's institutional consecration, but Nam's "bright and light" (Choi 2005), plain writing style did not change much after he achieved recognition. Scoring the book individually and comparing it with the score of his later novel *Unification Love Song* (통일련가), for instance, shows that both works fare low on the originality scale.

Hymn of Youth's low score is a reminder that the present estimation of literary originality is limited to writing style and does not extend to plot, themes, or even context. The book's romantic storyline was considered radically innovative at the time of its publication—but that novelty was about *what* was being said rather than *how* it was being said. Likewise, Nam's *Unification Love Song* scores low despite

https://doi.org/10.1017/S0021911821002278 Published online by Cambridge University Press

Table 2. Descriptive statistics for the five corpora

	KCNA	Poetry	Regular Novels	Prestigious Novels	Canonical Novels
Count	10,000	10,000	10,000	10,000	10,000
Mean	98.0	3,661.4	638.5	764.9	583.7
Standard deviation	354.1	29,200.8	1,877.7	2,363.0	2,064.0
Min	1.1	1.0	1.2	1.9	1.2
25%	8.5	62.8	100.4	106.9	61.5
50%	24.2	298.8	252.9	292.6	182.8
75%	75.8	1,375.8	643.2	723.9	498.7
Max	13,164.3	1,808,681.5	122,895.7	126,058.1	98,007.0

dealing with the same subject—unconverted long-term North Korean prisoners in the South—as Ri Chongnyŏl's *Sons of the Motherland* (조국의 아들), the novel with the highest individual originality score in the corpus. Relying on a single feature (style) to judge originality is certainly a limitation but could be augmented as new methods are developed to gauge the predictability of plot (Underwood 2020; Wilmot and Keller 2020). There are, of course, other limitations to the present approach. For one, other forms of literary distinctions in the DPRK, such as literary prizes, critical acclaim, or institutional appointments were disregarded. Kim Taesŏng, the author with the highest originality ratings among the writers of "regular novels," for instance, was chair of the North Korean Writers' Union—which would suggest that other forms of literary recognition might also correlate with literary originality. Scarcity of data also currently makes deeper analysis at the individual and historical levels impossible, but as digitized North Korean texts become easier to acquire, such limitations could soon be overcome.

Conclusion

The thematic analysis conducted using topic modeling has shown the importance of genre distinctions for the study of the relationship between politics and literary value in North Korea. Writers who received the highest level of literary distinction usually had careers as authors of military or industrial production novels. The relative value attached to each theme shifted over time, depending upon which social class the Party sees at its base. The lower proportion of agricultural short stories in the production of writers who achieved literary consecration might also indicate that the genre does not carry much prestige. There was no other statistically significant relation to literary value for other topics in the sample, but this does not necessarily imply that such other relations are nonexistent. Finally, while politics directly influences the level of prestige achieved by individual authors, it does not preclude the existence of other factors of literary evaluation.

The second part of the article analyzed one such factor, stylistic originality. Using the probabilities returned by a masked language model to estimate how original a sentence's writing is, I find that novels of the *Immortal* collections are stylistically less creative than other types of fiction. A more surprising finding, however, is that when the authors of these novels write on other topics, they exhibit higher literary originality than their peers. This suggests that originality has a dual status in North Korea. Where the pinnacle of literary value is concerned, conformity is more valued than creativity. But one step below in the literary hierarchy, stylistic originality does correlate with literary recognition. Either because consecrated writers are able to write more freely than their peers or because the state is selecting the most creative writers to produce its cult of personality.

Regardless of the causal direction, this nonetheless shows that creativity is valued in the North Korean literary system, although not necessarily in the same proportions as in other literary systems and in negotiation with political imperatives. Whereas conformity is more important than originality for the production of the most highly valued literary texts that deal with the personality cult of the

https://doi.org/10.1017/S0021911821002278 Published online by Cambridge University Press

Leaders, stylistic creativity does act as a distinguishing factor between writers of different prestige levels. The relationship between politics and creativity in the North Korean literary field can thus not simply be reduced to one of opposition. Whether writers accumulate political capital to gain a right to creativity or their literary capital gives them the possibility to accumulate political capital by gaining a monopoly on the production of a specific part of political discourse, what is at stake is a process of arbitrage, negotiation and trade between different forms of capital. It is therefore this fungibility of political and symbolic capitals rather than the antagonism between literature and politics—or between ideological and autonomous writing—that constitutes the core of literary activity in North Korea.

Acknowledgments. The author would like to thank Immanuel Kim, Henry Lydecker, Alex Shaw, Clara Sitbon, Chao Sun, the three anonymous reviewers, and the journal's editors for their help, comments, and corrections on earlier versions of the article. The author also acknowledges the Sydney Informatics Hub and the University of Sydney's high-performance computing cluster, Artemis, for providing the computing resources that contributed to the results reported here.

Author Bio. Benoit Berthelier is Lecturer in the Department of Korean Studies, Faculty of Arts and Social Sciences, at the University of Sydney.

List of References

An Mak. 1946. "Sin chŏngse wa minjujuŭi munhwa yesul chŏnsŏn kanghwa ŭi immu" 신정세와 민주주의문화예술전선 강화의 의무 [The new conjuncture and the duty of reinforcing the democratic fronts of the arts and letters]. *Munhwa Chŏnsŏn* 2:2–12.

Antoniak, Maria, Melanie Walsh, and David Mimno. 2021. "Tags, Borders, and Catalogs: Social Re-Working of Genre on LibraryThing." *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW1): Article 29.

Barthes, Roland. 1964. "Éléments de sémiologie" [Elements of semiology]. *Communications* 4:91–135.

Beecroft, Alexander. 2014. "World Literature without a Hyphen." In *World Literature in Theory*, edited by David Damrosch, 180–90. Oxford: Wiley Blackwell.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21),* 610–23. New York: Association for Computing Machinery.

Blei, David. 2012. "Topic Modeling and Digital Humanities." *Journal of Digital Humanities* 2 (1). http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/ (accessed September 14, 2020).

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. "Language (Technology) Is Power: A Critical Survey of 'Bias' in NLP." https://arxiv.org/abs/2005.14050 (accessed December 12, 2020).

Bourdieu, Pierre. 1979. *La Distinction* [Distinction]. Paris: Minuit.

Bourdieu, Pierre. 1991. "Le champ littéraire" [The literary field]. *Actes de la Recherche en Sciences Sociales* [Proceedings of Research in the Social Sciences] 89:3–46.

Bourdieu, Pierre. 1992. *Les Règles de l'art* [The rules of art]. Paris: Seuil.

Bradley, Efron, and R. J. Tibshirani. 1994. *An Introduction to the Bootstrap*. New York: Chapman and Hall/CRC.

Broadwell, Peter, Jack Chen, and David Shepard. 2019. "Reading the Quan Tang shi: Literary History, Topic Modeling, Divergence Measures." *Digital Humanities Quarterly* 13 (4). http://www.digitalhumanities.org/dhq/vol/13/4/000434/000434.html (accessed September 1, 2020).

Buchholz, Larissa. 2017. "What Is a Global Field? Theorizing Fields beyond the Nation-State." *Sociological Review* 64 (2): 31–60.

Carnap, Rudolf. 1964. *The Logical Syntax of Language.* London: Routledge and Kegan Paul.

Casanova, Pascale. 1999. *La République mondiale des lettres* [The world republic of letters]. Paris: Seuil.

Casanova, Pascale. 2005. "Literature as a World." *New Left Review* 31:71–90.

Cho, Heekyoung. 2018. "Rethinking World Literature through the Relations between Russian and East Asian Literatures." *Cross-Currents: East Asian History and Culture Review* 28:7–26. https://cross-currents.berkeley.edu/e-journal/issue-28/cho (accessed December 11, 2020).

Choi Chae-bong. 2005. "Kyŏngk'wae han munch'e ro namnyŏ sarang kŭmgi kkaeda" 경쾌한 문체로 남녀사랑 '금기' 깨다 [Breaking the taboo of girl-boy love with a light-hearted style]. *Hankyoreh*, August 4.

Choi, Yearn Hong. 1994. "North Korean literature in the Chosun Munhak." *Global Economic Review* 23 (1): 75–98.

Clark, Katerina. 1981. *The Soviet Novel: History as Ritual.* Bloomington: Indiana University Press.

Clark, Kevin, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. "What Does BERT Look At? An Analysis of BERT's Attention." BlackBoxNLP@ACL. https://arxiv.org/abs/1906.04341 (accessed November 1, 2020).

Dällenbach, Lucien, and Annette Tomarken. 1980. "Reflexivity and Reading." *New Literary History* 11 (3): 435–49.

David-West, Alzo. 2007. "An Inquiry of Intentions in Kim Hye-yŏng's First Meeting: A North Korean Short Story in Korea Today." *Transnational Literature* 6 (1). http://dspace.flinders.edu.au/jspui/bitstream/2328/27124/1/Inquiry_of_Intentions.pdf (accessed January 14, 2022).

De Ceuster, Koen. 2011. "To be an Artist in North Korea: Talent and Then Some More." In *Exploring North Korean Arts*, edited by Rüdiger Frank, 51–71. Nürnberg: Verlag für moderne Kunst.

https://doi.org/10.1017/S0021911821002278 Published online by Cambridge University Press

Devlin, Jacob, Ming-Wai Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, edited by Jill Burstein, Christy Doran, and Thamar Solorio, 4171–86. Minneapolis, MN: NAACL.

Djagalov, Rossen. 2018. "Literary Monopolists and the Forging of the Post-WWII People's Republic of Letters." In *Socialist Realism in Central and Eastern European Literatures: Institutions, Dynamics, Discourses*, edited by Natalia Skradol and Evgeny Dobrenko, 25–37. London: Anthem Press.

Epstein, Stephen. 2002. "North Korean Short Stories on the Cusp of the New Millennium." *Acta Koreana* 5 (1): 33–55.

Fricke, Harald. 2008. "How Scientific Can Literary Evaluation Be?" In *The Quality of Literature: Linguistic Studies in Literary Evaluation*, edited by Willie van Peer, 191–207. Amsterdam: John Benjamins Publishing Company.

Friedman, Susan Stanford. 2013. "World Modernism, World Literature, and Comparaticity." In *The Oxford Handbook of Global Modernisms*, edited by Mark Wollaeger and Matt Eatough, 499–525. Oxford: Oxford University Press.

Gardner, John. 1983. *The Art of Fiction*. New York: Random House.

Goldstone, Andrew. 2013. *Fictions of Autonomy: Modernism from Wilde to de Man*. New York: Oxford University Press.

Goldstone, Andrew. 2015. "Distant Reading: More Work to Be Done." August 8 (blog). https://andrewgoldstone.com/blog/2015/08/08 /distant/ (accessed December 6, 2020).

Günther, Hans, ed. 1990. *The Culture of the Stalin Period*. New York: Palgrave Macmillan.

Jawahar, Ganesh, Benoît Sagot, and Djamé Seddah. 2019. "What Does BERT Learn about the Structure of Language?" In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, edited by Preslav Nakov and Alexis Palmer, 3651–57. Stroudsburg, PA: Association for Computational Linguistics.

Herrnstein Smith, Barbara. 1983. "Contingencies of Value." *Critical Inquiry* 10 (1): 1–35.

Jakobson, Roman. 1960. "Linguistics and Poetics." In *Style in Language*, edited by Thomas Sebeok, 351–77. Cambridge, Mass.: MIT Press.

Jockers, Matthew, and David Mimno. 2013. "Significant Themes in 19th-Century Literature." *Poetics* 41:750–69.

Jurafsky, Dan, and James H. Martin. 2019. "Speech and Language Processing." https://web.stanford.edu/~jurafsky/slp3/ (accessed September 21, 2020).

Kang Chinho and Imha Yu. 2009. *Ch'ongsŏ 'pulmyŏl ŭi yŏksa' yŏn'gu* 총서 '불멸의 력사' 연구 [Research on the collection "Immortal History"]. Seoul: Somyŏng Ch'ulp'ansa.

Kang Sangho. 1993. "Namkigo sip'ŭn iyagi tŭl—naega ch'irŭn pukhan sukch'ŏng" 남기고 싶은 이야기—내가 치른 북한 숙청 [Stories I want to leave behind—The purge I underwent in North Korea]. *Chungang Ilbo*, January 11–October 12.

Kang Yongt'aek. 2003. *Kim Sowŏl kwa Cho Kich'ŏn ŭi siŏ sayong yangsang pigyo yŏn'gu* 김소월과 조기천의 시어 사용 양상 비교 연구 [Comparative study of the use of poetic words by Kim Sowŏl and Cho Kich'ŏn]. Seoul: Yŏngnak.

Kim Chaeung. 2012. "Pukhan ŭi kyegŭp chŏngch'aek kwa kyegŭp wigye kujo ŭi hyŏngsŏng (1945–1950)" 북한의 계급정책과 계급위계구조의 형성(1945–1950) [Class policies and formation of hierarchical class structure in North Korea (1945–1950)]. *Yŏksawa hyŏnsil* 85:393–433.

Kim Chaeyong. 1994. *Pukhan munhak ŭi yŏksajŏk ihae* 북한 문학의 역사적 이해 [A historical understanding of North Korean literature]. Seoul: Munhak kwa chisŏngsa.

Kim Chinhwan. 2016. "Pukhan sahoe kyegŭp yŏn'gu" 북한사회 계급 연구 [Research on classes in North Korean society]. In *2016-nyŏn t'ongil pukhan yŏn'gu ch'oejong pogosŏ* 2016 년 통일·북한 연구 최종보고서 [Final research report on North Korea and Unification], 235–77. Seoul: National Institute for Unification Education.

Kim Immanuel. 2012. *"North Korean Literature: Margins of Writing Memory, Gender, and Sexuality."* PhD diss., University of California, Riverside.

Kim Immanuel. 2016. "Problems with Institutionalizing the April 15 Literary Production Unit." *Korea Journal* 56 (1): 140–64.

Kim Jong Il. [1967a] 1987. "Munhak yesul pumun esŏ dang ŭi yuil sasang ch'egye rŭl t'ŭnt'ŭnhi seulde taehayŏ" 문학예술부문에서 당의 유일사상체계를 튼튼히 세울데 대하여 [On strongly building the party's monolithic ideology in the domain of arts and letters]. In *Chuch'e hyŏngmyŏng wiŏp ŭi wansŏng ŭl wihayŏ* 주체 혁명 위업 의 완성 을 위하여 [For the completion of the great task of the Juch'e revolution], 38–52. Pyongyang: Chosŏn rodongdang ch'ulp'ansa.

Kim Jong Il. [1967b] 1994. "4.15 munhak ch'angjaktan ŭl naeolte taehayŏ" 4.15 문학창작단을 내올데 대하여 [On unveiling the 4.15 literary production unit]. In *Kim Jong Il sŏnjip* 김정일선집 [Selected works of Kim Jong Il], vol. 1, 242–84. Pyongyang: Chosŏn rodongdang ch'ulp'ansa.

Kim Jong Il. 1992a. *Chuch'e munhangnon* 주체문학론 [Theory of Juche literature]. Pyongyang: Chosŏn rodongdang ch'ulp'ansa.

Kim Jong Il. 1992b. *On Juche Literature*. Pyongyang: Foreign Languages Publishing House.

Kim Jong Il. 2007. *The Songun-Based Revolutionary Line Is a Great Revolutionary Line of Our Era and an Ever Victorious Banner of Our Revolution*. Pyongyang: Foreign Language Publishing House.

Lankov, Andrei. 2002. "Kim Takes Control: The 'Great Purge' in North Korea, 1956–1960." *Korean Studies* 26 (1): 87–119.

Lankov, Andrei. 2014. *The Real North Korea: Life and Politics in the Failed Stalinist Utopia*. Oxford: Oxford University Press.

Latynina, Alla. 1996. "The Stalin Prizes in Literature as the Quintessence of Socialist Realism" In *In the Party Spirit: Socialist Realism and Literary Practice in the Soviet Union, East Germany and China*, edited by Hillary Chung, 106–28. Amsterdam: Rodopi.

Lee Ch'ŏnggu. 1950. "Ssobet'ŭ si munhak kwa uri siin tŭl" 쏘베트 시문학과 우리 시인들 [Soviet poetry and our poets]. *Munhak yesul* 2:55–66.

Lee Kwangsu. 1916. "Munhak iran hao" 文學이란 何오 [What is literature?]. *Maeil Shinbo*, November 10–23.

https://doi.org/10.1017/S0021911821002278 Published online by Cambridge University Press

Lee Sangsuk. 2013. "Kim Jong Un sidae ŭi ch'ulbal kwa pukhan si ŭi ch'ui" 김정은 시대의 출발과 북한시의 추이 [Study on North Korean poetry in the era of Kim Jeongun]. *Han'guk Sihak Yŏn'gu* 38:181–212.
Louwerse, Max, Nick Benesh, and Bin Zhang. 2008. "Computationally Discriminating Literary from Non-Literary Texts." In *Directions in Empirical Literary Studies*, edited by S. Zyngier *et al*., 175–91. Amsterdam: John Benjamins Publishing Company.
Lumley, Thomas, Paula Diehr, Scott Emerson, and Lu Chen. 2002. "The Importance of the Normality Assumption in Large Public Health Data Sets." *Annual Review of Public Health* 23:151–69.
Merriman, Ben. 2015. "A Science of Literature." *Boston Review*, June 23. https://bostonreview.net/articles/ben-merriman-moretti-jockers-digital-humanities/ (accessed January 14, 2022).
Milner, Andrew. 2005. *Literature, Culture and Society*. New York: Routledge.
Min Pyŏnggyun. [1949] 1993. "Ssobet'ŭ si munhak ŭl sŏpch'wi ham e isŏsŏ kyŏngŏm kwa kyohun" 쏘베트 시문학을 섭취함에 있어서 경험과 교훈 [Experiences and lessons in the assimilation of Soviet poetry]. *Rodong Sinmun*, April 29. In *Hyŏndae Munhak pip'yŏng charyojip. Ibukp'yŏn* 현대문학 비평 자료집 [Anthology of contemporary literary criticism: Northern volume], edited by Kim Chaeyong *et al*., 383–91. Seoul: Taehaksa.
Moretti, Franco. 2013. *Distant Reading*. London: Verso.
Moretti, Franco. 2015. "A Response." *PMLA* 132 (3): 686–89.
Mun, Yŏngho. 1993. *Chosŏnŏ pindo susajŏn* 조선어빈도수사전 [Dictionary of frequencies in Korean]. Pyongyang: Kwahak paekkwa sajŏn chonghap ch'ulp'ansa.
Myers, Brian. 1994. *Han Sŏrya and North Korean Literature*. Ithaca, N.Y.: Cornell University Press.
Neumann, Birgit, and Gabriele Rippl. 2017. "Anglophone World Literatures: Introduction." *Anglia* 135 (1): 1–20.
Paek Sŏnggŭn. 2017. "Hyŏngmyŏng sosŏl ch'angjak ŭi midŭmjik'an kiji" [The trustworthy base of the creation of revolutionary novels]. *Rodong Sinmun*, June 18.
Pan Sangsŏ. 2013. "Hŭimang ŭi ŏndŏk" 희망의 언덕 [Hills of hope]. *Ch'ŏngnyŏn Munhak* 5 (8): 53–62.
Park Namsu. 1999. *Chŏkch'i 6yŏn ŭi pukhan mundan* 적치 6 년의 북한문단 [Red Flag: Six years of the North Korean literary world]. Seoul: Pogosa.
Park, Sunyoung. 2015. *The Proletarian Wave: Literature and Leftist Culture in Colonial Korea, 1910–1945*. Cambridge, Mass.: Harvard University Press.
Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why*. New York: Basic Books.
Person, James F. 2013. "The 1967 Purge of the Gapsan Faction and Establishment of the Monolithic Ideological System." North Korea International Documentation Project, Wilson Center. https://www.wilsoncenter.org/publication/the-1967-purge-the-gapsan-faction-and-establishment-the-monolithic-ideological-system (accessed December 6, 2020).
Pihl, Marshall. 1977. "Engineers of the Human Soul: North Korean Literature Today." *Korean Studies* 1:63–110.
Quine, Willard V. 1953. "On What There Is." In *From a Logical Point of View*, 1–19. Cambridge, Mass.: Harvard University Press.
Ri Dongwŏn. 1996. *Chuch'e munhak riron* 주체문학리론 [Theory of Juche literature]. Pyongyang: Kim Il Sung Chonghap Taehakkyo Ch'ulp'ansa.
Ryu Ch'anyŏl. 1998. "90-yŏndae ŭi pukhansi" 90 년대의 북한 시 [North Korean poetry of the 1990s]. In *Pukhan munhak ŭi inyŏm kwa silch'e* 북한문학의 이념과 실체 [Actual state and concept of North Korean literature], edited by Myŏngjae Lee, 228–47. Seoul: Kuk'hak charyowŏn.
Salazar, Julian, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhof. 2020. "Masked Language Model Scoring." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetrault. Stroudsburg, PA: Association for Computational Linguistics.
Schmid, Andre. 2018. "Historicizing North Korea: State Socialism, Population Mobility, and Cold War Historiography." *American Historical Review* 123 (2): 439–62.
Shin Hyŏnggi. 1988. *Haebang chikhu ŭi munhak undongnon* 해방직후의 문학운동론 [Theory of literary movements in the immediate post-Liberation period]. Seoul: Hwada.
Shin Hyŏnggi and Sŏngho Oh. 2000. *Pukhanmunhaksa* 북한문학사 [North Korean literary history]. Seoul: P'yŏngminsa.
Smith, Hazel. 2015. *North Korea: Markets and Military Rule*. Cambridge: Cambridge University Press.
So, Richard Jean. 2017. "All Models Are Wrong." *PMLA* 132 (3): 668–73.
So, Richard Jean. 2020. *Redlining Culture: A Data History of Racial Inequality and Postwar Fiction*. New York: Columbia University Press.
Spivak, Gayatri. 2003. *Death of a Discipline*. New York: Columbia University Press.
Suleiman, Susan Rubin. 1983. *Authoritarian Fictions: The Ideological Novel as a Literary Genre*. New York: Columbia University Press.
Underwood, Ted. 2017. "A Genealogy of Distant Reading," *Digital Humanities Quarterly* 11 (2). http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html (accessed August 31, 2020).
Underwood, Ted. 2020. "How Predictable Is Fiction?" July 5 (blog). https://tedunderwood.com/2020/07/05/how-predictable-is-fiction/ (accessed September 20, 2020).
Underwood, Ted, and Jordan Sellers. 2016. "The Longue Durée of Literary Prestige." *Modern Language Quarterly* 77 (3): 321–44.
Van Cranenburgh, Andreas, Karina van Dalen-Oskam, and Joris van Zundert. 2019. "Vector Space Explorations of Literary Language." *Language Resources and Evaluation* 53:625–50.
Van Rooden, Aukje. 2015. "Reconsidering Literary Autonomy: From an Individual towards a Relational Paradigm." *Journal of the History of Ideas* 76 (2): 167–90.

https://doi.org/10.1017/S0021911821002278 Published online by Cambridge University Press

Wilmot, David, and Frank Keller. 2020. "Modelling Suspense in Short Stories as Uncertainty Reduction over Neural Representation." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetrault, 1763–88. Stroudsburg, PA: Association for Computational Linguistics.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Oxford: Basil Blackwell.

Appendices

Appendix 1. The 24 writers from the April 15 LPU used for analyzing the output of canonized writers before and after they achieved recognition

Name	Admission Date	Texts in Corpus	Name	Admission Date	Texts in Corpus
Kwŏn Chŏngung 권정웅	1967	4	An Tongch'un 안동춘	1990	11
Kim Pyŏnghun 김병훈	1967	4	Yun Kyŏngch'an 윤경찬	2013	8
Kim Sambok 김삼복	1998	4	Yun Chŏnggil 윤정길	2014	1
Kim Sugyŏng 김수경	1989	1	Chŏn Hŭngsik 전흥식	2012	5
Nam Taehyŏn 남대현	1992	4	Chŏng Kijong 정기종	1992	7
Ri Tonggu 리동구	2002	2	Cho Kwŏnil 조권일	2018	1
Ri Sinhyŏn 리신현	1998	5	Chin Chaehwan 진재환	1984	2
Ri Chongnyŏl 리종렬	1981	2	Choi Yongjo 최영조	2011	2
Rim Pongch'ŏl 림봉철	2015	7	Choi Haksu 최학수	1976	3
Pak T'aesu 박태수	2000	3	Tak Sukbon 탁숙본	2013	10
Paek Namnyong 백남룡	1985	10	Hŏ Ch'unsik 허춘식	1997	1
Paek Pohŭm 백보흠	1989	8	Hyŏn Sŭnggŏl 현승걸	1976	3

https://doi.org/10.1017/S0021911821002278 Published online by Cambridge University Press

Appendix 2. Proportion of the publications in the corpus used for the training of a North Korean BERT. The final corpus corpus is 1.2 gigabytes, containing 4,428,438 sentences and 208,638,834 words. The data were scraped from the web or extracted from electronic books obtained in the DPRK

Magazines	25.1%
– *Kim Il Sung chonghap taehakkyo hakpo* (김일성종합대학교학보)	– 16.9%
– *Chosŏn Munhak* (조선문학)	– 16.9%
– *Kyŏngje Yŏn'gu* (경제연구)	– 7.7%
– *Ryŏksa Kwahak* (력사과학)	– 11.2%
– *Adong Munhak* (아동문학)	– 9.6%
– *Chosŏn Yesul* (조선예술)	– 11.1%
– *Chŏllima* (천리마)	– 16.3%
– *Ch'ŏngnyŏn Munhak* (청년문학)	– 10.2%
Fiction	3.0%
Complete works of Kim Il Sung and Kim Jong Il	4.5%
Nonfiction books	1.1%
Korean Central News Agency articles (2001–2018)	25.3%
Rodong Sinmun newspaper (2000–2018)	41.1%

Cite this article: Berthelier, Benoit (2022). Quantifying Quality: A Computational Approach to Literary Value in North Korea. *The Journal of Asian Studies* **81**, 267–288. https://doi.org/10.1017/S0021911821002278

https://doi.org/10.1017/S0021911821002278 Published online by Cambridge University Press