Interview

Xiaohe Rui* and James F. English

Literary Awards as Humanities Data: Measuring Prestige in the Digital Age – an Interview with James English

https://doi.org/10.1515/dsll-2025-0011
Received June 24, 2025; accepted August 3, 2025; published online August 29, 2025

Abstract: The utilisation of literary prizes within the domain of Digital Humanities signifies the emergence of a novel metric for evaluating prestige. In this interview, Xiaohe Rui and James English discussed the development of the field, as well as the challenges in bridging the disciplines, ethics and accessibility of prize data used in Digital Humanities. English introduced his research team's work on literary data, using tools such as Python, Tableau, Gephi and t-SNE to generate insights and foster a dynamic, peer-driven learning environment that may offer greater pedagogical value than traditional English classes. He also emphasized that in order to comprehend literary prestige in the present age, it is necessary to examine the manner in which cultural value is disseminated across a variety of media platforms. This interdisciplinary, data-driven approach, English suggested, will persist as a vital and engaging niche within the domain of literary studies.

Keywords: Digital Humanities; prize studies; data; prestige; computational models

James English is the John Welsh Centennial Professor of English and founding Faculty Director of the Price Lab for Digital Humanities at the University of Pennsylvania. He served as Chair of Judges for the National Book Award for Fiction in 2016. His book, *The Economy of Prestige* (Harvard UP), a study of the history, functions, and effects of prizes in literature and arts, was named Best Academic Book of 2005 by *New York* Magazine. His economic theory of cultural prestige has exerted extensive influence in the fields of literary theory, cultural sociology and cultural economics. In February 2025, Xiaohe Rui, Professor of English at Xi'an International Studies University, and co-translator of

***Corresponding author: Prof. Xiaohe Rui**, Professor of English, School of English Studies, Xi'an International Studies University, South Chang'an Road 437, Xi'an, 710061, China,
E-mail: ruixiaohe@xisu.edu.cn
Prof. James F. English, John Welsh Centennial Professor of English, Department of English, University of Pennsylvania, 3340 Walnut Street, Philadelphia, PA, 19104, USA, E-mail: jenglish@english.upenn.edu

the Chinese version of *The Economy of Prestige*, conducted an interview with English on the use of literary prizes in the field of Digital Humanities.

Xiaohe Rui (Rui for short hereafter): *Literary prizes have been increasingly used in digital humanities research as proxies for critically esteemed fiction. You, Andrew Piper, Richard Jean So, and a number of others have used digital approaches to analyze literary prize trends. What do you see as the key strengths and weaknesses of this approach?*

James English (English for short hereafter): Yes, literary scholars have compiled sets of novels longlisted or shortlisted for major awards and used those as a proxy for critically-esteemed contemporary fiction in general. A digitized corpus of prizelisted novels – or, more often, a set of metadata for those novels – is compared algorithmically to corresponding data of a set of random or garden-variety novels from the same years in order to discover patterns of change (or continuity) in the relationship between prestigious and unprestigious fiction. Or the prize novels are compared to bestsellers, revealing differences between critically successful and commercially successful fiction. Or the comparison might be to a set of romances or fantasy novels, revealing differences between literary fiction (conceived now as a kind of genre) and other genres. In studies of that latter kind, the shortlists of genre-specific prizes such as the Hugo and Nebula for science fiction or the Edgars for crime fiction might also serve as proxies for high critical reputation in those genres. There is no shortage of prizes, so you can find a set of novels corresponding to almost any piece of the contemporary literary landscape by searching the prize lists.

This kind of work can lead to methodological innovations and provocative findings, for sure. But it also has some pretty obvious limitations. It operates with very thin concepts of both prizes and prestige. With respect to prizes as such, this work teaches us very little. You can't understand much about how prizes work or what they do without looking under the hood at their founding documents and guiding principles, the internal mechanisms of their administration and finance, the demographics and professional histories of their personnel, the dramaturgy of their award ceremonies, the language of their commendation and acceptance speeches, the medals and statuettes awarded to their winners, and all the bickering and hype that often surround them, and which their administrators frequently cultivate. When I was working on *The Economy of Prestige*, these were some of the primary areas of my research. But when I use prize lists in computational studies, I mostly put all that aside and just tag any book or author that appears on a prize list as *prestigious* or *esteemed*.

And this points to the other and more compromising limitation of such an approach; it typically relies on too simple an equation between literary awards and critical esteem. Much of what makes prizes so interesting, at least to me, is that they function as a kind of "scandalous currency"; they convey to the recipient (and others

involved) a curious amalgam of both status and stigma. To say that someone writes "the kind of books that win prizes" is to pay them a rather backhanded compliment. When we decide to take a prize like the Pulitzer or the Booker at face value, as a sound proxy for critical status, we are ignoring its many vocal detractors, its history of scandal and derision, its obvious errors of judgment and defects of design: all the fundamentally ambiguous, even paradoxical dimensions that give a prize its particular cultural power. Prizes do convey prestige. But one of the principles that guided my work in the *Economy of Prestige* is that prestige in a particular field of culture is not quite the same thing as so-called "specific" cultural capital, that is, one's reputation among critics, academic experts, and other artists on the field. That kind of endowment weighs in, but so do celebrity and public profile (journalistic capital), ethnicity and nationality (identity capital), even commercial success (economic capital). As its etymology suggests, prestige (from the Latin prestigiae, meaning illusion or sleight-of-hand) is a tricky and treacherous thing. Prizes have spread so far and wide because they are the best devices we've come up with for managing the complex transactions and interconversions involved in an economy of prestige.

Rui: *Through use of algorithmic comparisons you've uncovered patterns that might not be evident through traditional literary analysis. What are the most revealing patterns or notable findings that you and these scholars have uncovered so far?*

English: The first work of this kind that I did was back in 2016, in a piece called "Now, Not Now." That study found a dramatic shift on the field of contemporary Anglophone fiction starting around 1980. The shift involved the time-period of the novels' settings. In the 1960s and 1970s, if you picked up a new work of fiction, whether it was a top bestseller or a novel shortlisted for a major award, it very likely depicted events and actions taking place in the present moment. There were comparatively few novels on the bestseller lists or the prize shortlists depicting events of the historical past. But suddenly, beginning in the late 1970s, novels set in the past began to show up as a larger and larger fraction of the books contending for leading novel-of-the-year awards, while at the same time becoming scarcer and scarcer among the bestsellers. And this divergence was decisive. It has remained the case since the 1980s that critically successful fiction is mostly set in the past and commercially successful fiction is almost always set in the present.

Now this particular finding was based on statistical analysis, but it was not heavily computational. In another study published that same year, "How Cultural Capital Works: Prize-winning Novels, Bestsellers, and the Time of Reading" (2016), Andrew Piper and Eva Portelance used machine-learning algorithms to compare the lexicons of different sets of novels arrayed in a hierarchy of high to low prestige. This method was capable of discerning subtler relationships between contemporary novels' temporal orientations and their prestige than what I discovered. Piper and

Portelance found what they describe as a correlation between contemporary novels' cultural status and their reliance on "nostalgic narrativity." Their algorithm found 50 % more passages of "nostalgic" language in prizewinning novels (at the top of the status hierarchy) than in bestsellers, and more in the bestsellers than in lowest-status genres of science fiction and romance. Prestige, according to this study, does not just gravitate toward novels whose action is set in the past but toward novels shaped by a "nostalgic mentality," which Piper and Portelance critiqued as a predominantly patriarchal, father-oriented discursive orientation.

Rui: *DH often emphasizes these kinds of large-scale patterns. How do scholars balance this with the need for close, qualitative reading of individual prize-winning works?*

English: Well, I don't think individual works of scholarship necessarily need always to be balanced in that way, every large-scale finding augmented or supported by a little demonstration of close reading. I'm ok with studies that dispense entirely with close reading, just as I'm happy to learn from a close reading that is not balanced by some kind of large-scale quantitative evidence. But it is often the case that when we use computational methods to discover a broad statistical pattern or trend, that finding leads to further insights gleaned through qualitative analysis of individual works. To stick with the examples we were just discussing, in his recent book *Writing Backwards: Historical Fiction and the Reshaping of the American Canon*, Alexander Manshel argues that the turn to the past in high-status literary fiction has been even more dramatic than my study suggested, since there are many novels of the prizelisted sort that, while not actually set in a past time period, nonetheless deploy the resources and perspectives of historical fiction or historical allegory. And he very persuasively links this reorientation of high-status fiction toward history to the symbolic elevation of minority writers in American literature. The agents of canon formation, both individual and institutional, have in effect imposed the burden of writing "historically" as a condition of entry for writers of color. But, in a series of fine close readings, Manshel shows that some of those writers (his prime example is Colson Whitehead) have deployed historical narrative in ways that undermine rather than reinforcing the "nostalgic mentality" critiqued by Piper and Portelance. Manshel's work is a great example of how a broad-brush and relatively crude statistical pattern of the kind discovered by me or by Piper and Portelance in the data of prize and bestseller lists can be tested, fleshed out, and refined not just through additional data gathering and statistical analysis but through traditional close reading. As DHers have been insisting all along, the one kind of work is not a replacement for the other; the methods are complementary, and Manshel is adept at both.

Rui: *How does DH use literary prizes to challenge or complement traditional approaches to the study of literary prestige and canon formation?*

English: This is a good way to frame the question – better I think than the terms you used when first proposing this conversation. At that point, the focus was to be on "Interdisciplinary Collaborations between Literary Prize Studies and Digital Humanities." The term "interdisciplinary" is constantly invoked in discussions of DH; it's a term the field loves to use to describe itself. But it can be misleading. When we speak about the analysis of literary prizes by scholars such as Piper, Manshel, J. D. Porter, Richard Jean So, Laura McGrath, or myself, we are talking about data-driven, computational work in literary studies by scholars of contemporary literature, housed in English departments. Occasionally, some of the coding is done by a grad student in linguistics or computer science, but generally speaking this work does not involve a collaboration between literary scholars and scholars in some other discipline. It is rarely interdisciplinary in any meaningful sense. Almost no one I can think of in literary studies devotes themselves wholly to the study of prizes or declares "prize studies" to be their field. Even scholars whose work is most concerted around prizes, such as Claire Squires and Stevie Marsden in the UK, would describe their field, I think, as book history, or, to use the keywords of the book-history field group SHARP, "the history of authorship, readers, and publishing." Prizes are an important phenomenon in the literary world and a subject of interest for scholars who work in this field, book history, studying the history and theory of literary canon formation, or the history of literary institutions. Among the various methods used by scholars in all these areas are some forms of statistical analysis. Often the datasets are small and hand-built and the statistical analysis is very light in computational terms. When the data get bigger and the methods more technical, we like to call the work digital humanities. But it's not as though there are two separate disciplinary camps struggling to find a little common "interdisciplinary" ground. The leading figures in literary DH, from Alan Liu and Franco Moretti to Ted Underwood, Andrew Piper, Meredith Martin, Richard Jean So, and Whitney Trettien, are all literary historians and historians of the book, first and last. Within that field, they have championed more computationally advanced, technology-intensive methods. At times, to be sure, some of these methods – mainly methods of algorithmic or "distant" reading of large literary corpora – have provoked great controversy. But it is controversy within the one discipline. And after all, book historians have long been interested in counting things: pirated editions of works by a particular author, for example, or the number of volumes of a certain kind in various city libraries. They have long been assembling and digitizing corpora and building statistical support for empirical arguments about a rise or fall or tendency of some kind in the publishing business or among library patrons or women's reading groups or what have you. Over the years, these kinds of tasks have become more reliant on computing. At this stage, I think nearly everyone engaged in such work is well past the grand controversies of quantitative vs qualitative, and simply trying to gather better data and develop better methods for

uncovering these kinds of trends or patterns of literary activity. That is happening within literary studies, not in a distinct interdisciplinary zone of academe.

Now, to turn to your question, one strain of literary historical scholarship is concerned with the allocation of reward and esteem on the literary field – often specifically with the distribution of literary value and the construction of a canon of masterpieces. Historical data having to do with prizes is only useful for work in this area going back to about the mid-20th century. Prior to that there just aren't enough prizes, awards, and honors to support much quantitative analysis. But even for studying the so-called "contemporary canon," the shifting set of late-20th and early-21st century works that are held to be the masterpieces of this period, prizes are not necessarily the best proxy. As John Guillory described in *Cultural Capital*, the canon is largely the province of the school, a matter of specifically *academic* status. Using prizes as a metric of canonicity tends to underestimate the power of the professoriate to canonize works that prize juries ignored. This is why J. D. Porter, in his influential studies of the popularity versus the prestige of canonical literature, used not prizes and awards but the number of academic articles in the *MLA International Bibliography* as his main proxy for prestige. More recently, Laura McGrath has used the frequency of a work's appearance on college syllabi – based on data from the Open Syllabus Project – as a metric of its canonicity. Other proxies for the canon are textbook anthologies such as the *Norton Anthology of American Literature*, or best-of lists – typically compilations of critics' picks – such as the Modern Library list of "Top 100 Best Novels," or the Guardian's "100 Greatest Novels of All Time."

Given all these other options, should we stop using prizes as data for studying literary canon formation? I wouldn't go that far. Manshel, McGrath, and Porter collaborated a few years ago on a digital humanities project that speaks to just this question. They gathered several hundred works from four different lists of the greatest or most important novels of the early 21st first century, as well as one list of the bestsellers from this period, and looked for patterns in popularity, measured by sales and by the number of ratings on Goodreads, and prestige, measured by articles in the MLA bibliography, appearances on syllabi in the Open Syllabus Project, and shortlists of five major fiction awards. What they found was that correlations between prize lists and specifically academic metrics of value (academic articles and syllabi) are actually quite strong. Prizes, they concluded, are not the be-all and end-all of canonicity, but they do matter to the process of canon formation. Methodologically, I think their approach in this study is a great model. Rather than accepting any one list or set of lists as the proxy for critical esteem, we should design our projects in digital book history around multiple metrics and indicators that capture a work's value from different angles, different positions on what is after all a turbulent and contested field.

Rui: *What specific digital tools or methods would you say are best suited for this kind of analysis of prize data?*

English: Well, I can tell you about our own experience with this kind of work at the Price Lab for Digital Humanities at Penn. J. D. Porter and I lead a small, ongoing research team at the lab that studies contemporary readers and reading practices. We usually have two or three undergraduates and two or three graduate students in the group. We meet for an hour or two every week to take stock of the work everyone has accomplished since the last meeting, and to decide our next steps. Most of what we do is data gathering and exploratory data analysis (EDA). We gather and organize literary data, most often from websites for readers and reviewers, like Goodreads, StoryGraph, and BookMarks, but also data about publishers and sales figures, library collections, academic syllabi, literary prizes: all the kinds of data you and I have been discussing. Our standard approach is to write code for webscraping and so on in Python, save data in. csv format, and use data visualization tools to look at the data together and develop hypotheses. Our favorite tool is Tableau, a powerful but relatively easy to use analytics platform that we can connect to our data files and generate beautiful interactive charts and graphs that guide our thinking and often spark new ideas and insights. J. D. is an expert visualizer and a wizard with Tableau, but others on the team can interact with his visualizations independently or in smaller groups. Unfortunately, Tableau is not open source. It's an expensive program marketed by a large corporation called Salesforce, and for that reason it's not widely used in digital humanities. But it's available free to teachers and students, and we use it more than any other single tool for EDA. It's not good for every kind of exploration, of course. When we want to view data in the form a network, especially if the dataset is pretty large, involving for example thousands of readers, tens of thousands of books, and millions of book ratings, we use Gephi. And when we just want to look at simple linear regressions or capture outliers, we often just use Excel or GoogleSheets. Those are actually pretty powerful tools for DH, if you can tap into their more advanced features.

When, after exploring our data this way, through visualizations, we arrive at a hypothesis that we want to pursue more deeply, we reach for other, task-specific tools. For example, some of our data sheets are quite high-dimensional: hundreds or thousands of columns in every row. In order to make sense of that kind of high-dimensional data, you often need to project it onto just two dimensions. For that task, we've used a stochastic neighbor embedding technique called t-SNE.

Other DH teams and individual scholars work very differently than we do, and prefer different tools. I don't think that our mainly exploratory, visual, team-based approach is necessarily the most efficient or productive. But it works a kind of pedagogical magic, with everyone learning from everyone else. Our technical

backgrounds and abilities vary widely, as do our literary tastes, but we are all obsessed with contemporary literature and culture. In the course of conducting our research projects we are constantly sharing our knowledge with each other, comparing notes. Everyone is a teacher as well as a student. And everyone is there because they want to be, because it's so much fun to be part of. I honestly believe that more and deeper learning takes place in a semester of participation on this kind of EDA team than in a typical college English class. We are accustomed to appraising the value of literary DH in terms of potential breakthroughs in research, but its real value to our struggling discipline may have more to do with breakthroughs in pedagogy, with its potential to inject new joy and energy into our teaching and learning.

Rui: *A major area of concern in digital humanities has been algorithmic ethics. Are there inherent biases in the kinds of algorithmic analysis that you and other scholars use to study patterns of literary prestige and canons of taste?*

English: Oh my god, there are biases on all sides! I think sometimes the algorithms are the least of it. Every mechanism involved in the selection and elevation of certain literary works, certain authors, instead of others, is subject to bias. There are biases in the education system that provides the skills needed to be an author or a reader; there are biases in the publishing industry that chooses which authors are published; biases in the process of selecting which books get reviewed in the newspapers and literary press, the way that critics review those books, and the books critics then select when asked to judge a prize or to help construct a list of the "Greatest Novels of All Time." And then there's the general bias of all these mechanisms, the bias shared by us academic scholars and teachers, in favor of what we call literary fiction, which is basically a bias toward fiction read almost exclusively by an educated elite. We are dealing here with broad patterns of social bias that minoritize and stigmatize certain populations of people, their experiences, their cultural norms and tastes.

Given that the whole social system is rife with bias, and the literary sphere in particular is avowedly hierarchical and undemocratic, scholars need to study bias in relational terms. The question is not whether the demographics of American literary prize winners perfectly match the national population as a whole, for example, but how they compare to relevant subsets: to the demographics of prize judges and prize administrators, of the publishing industry, of MFA programs. Where in the whole chain of selecting and sorting mechanisms are the decisive differences made? Where do minoritized, disadvantaged groups lose the most ground on the literary field, or gain the most, relative to other groups?

Using algorithmic methods to analyze prizelist metadata alongside data on education, publishing, and so on, strikes me as less of an ethical risk than not using such methods and relying on lore, anecdote, and desire. Statistical analyses of some

kind are absolutely necessary to understand how biases on the literary field are maintained and what changes would be most effective to correct them. Some of the best work I know of in this area has been done by Juliana Spahr and Stephanie Young, who have looked at the high-prestige end of the literary field as a small-world, gendered network structured by exclusions and inequities closely tied to certain academic institutions and programs. With Claire Grossman and Jordan Pruett they compiled a fantastically useful dataset called the "Index of Major Literary Prizes in the US."[1] Housed at *Post45*, that dataset contains lists of the winners and judges of more than 50 literary awards (for poetry as well as prose) dating back to 1918, with data on their gender and education (degree and institution), as well as detailed metadata on most of the winning books. Their own analyses have focused on forms of bias operating through poetry prizes, but other scholars have used their data to see how social biases of other kinds manifest in fiction prizes. Manshel and Melanie Walsh, for example, have supplemented the Index with additional data on the demographics (including racial identity) of fiction prize judges, winners, and finalists since 1988. In a piece called "What 35 Years of Data Can Tell Us About Who Will Win the National Book Award" in *Public Books*, they show among other things how strongly the racial diversity of prize juries influences the diversity of the authors they select for the shortlist.

My point, I guess, is that his whole area of study – the history and function of literary prizes, canons of taste, hierarchies of value – is focused, precisely, on bias. Prizes are not especially interesting in themselves; I'd rather read a good novel than a transcript of an awards ceremony any day! I study them in order to understand more clearly how literature serves as a structure of belonging and exclusion, a system of aesthetic distinctions that are also always social distinctions. With literature as with other social practices, bias is more deeply embedded, more layered, more extensive, and more complex than we imagine. The thoughtful gathering and analysis of quantitative data seems to me a necessary step toward gaining critical purchase on such multi-dimensional forms of inequity.

***Rui**: To sum up, then, do you think this kind of DH work will continue to flourish? Will DH methods be key to our better understanding not just of literary awards but, as you once phrased it, of the circulation of cultural value?*

English: Hmm. Responding to this question now, from my office at one of the US universities that has been under constant assault from anti-science, anti-intellectual, anti-academic forces on the extreme right, I feel reluctant to make

1 Grossman, Claire, Juliana Spahr, and Stephanie Young. 2022. "The Index of Major Literary Prizes in the US." Edited by Dan Sinykin and Melanie Walsh. *Post45 Data Collective*, December. https://doi.org/10.18737/CNJV1733p4520221212.

optimistic predictions about any stream of ongoing academic research. But assuming our higher education system fights back and survives this dark winter, then yes, I would expect statistical research into systems of cultural reward and esteem not only to continue, but to expand from literary studies into the study of other media. One of my favorite works of cultural criticism since the 2000s is Simone Murray's *Adaptation Industry*, which shows how elaborately interwoven the various media of art and entertainment have become. Reading Murray's book made me realize that my habitual ways of thinking about literary value were becoming obsolete. The value of intellectual property these days is parcelled out across multiple platforms, fractured and reallocated, leveraged and enhanced, in an ongoing churn of remediations. Murray emphasizes the commercial aspects of this adaptational system, but it affects the distribution of symbolic value, as well. We are not understanding much about the canonicity of Jane Austen, for example, if our analysis depends entirely on data connected with the literary field. You and I began this conversation talking about the way literary prestige became sutured to historical settings and discourses in the late 20th century. To gain a clearer picture of how and why that happened, and with what effects, we would need to explore additional data about the value attached to historical tropes and settings in cinema, television, theater, tourism and virtual tourism, gaming, theme parks, and more. Austen, of course, circulates through all of these, all the time. But so, to various extents, do many of the names and titles we think of as properly belonging to the literary sphere. There is, in short, plenty of data analytical work remaining to be done by scholars interested in the economics of cultural prestige. Those scholars will probably never amount to more than a small niche in literary studies, one of the many, many subfields that constitute our discipline. But I do think it will continue to be a lively and productive subfield, and one of more interest than most to students and even to non-academics.

Rui: *Thanks so much for your time!*
English: *It's a great pleasure!*

Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Use of Large Language Models, AI and Machine Learning Tools: None declared.
Conflict of interest: The authors state no conflict of interest.
Research funding: None declared.
Data availability: Not applicable.