The Standardization of Fast Fiction: Methods and Preliminary Results

Dennis Yi Tenen[1], Yingyi Shu, and Narrative Intelligence Lab*

[1] Department of English and Comparative Literature, Columbia University, New York, USA

Abstract

In this paper, we isolate and describe a particular sense of "standardization" in literature, associated with a major change in authorial practice at the turn of the twentieth century. Does fast fiction, like fast fashion, bear the marks of its manufacture? We measure standardization at four levels of textual redundancy: vocabulary, phrasing, syntax, and plot. We compare a random sample of American "super-producers," fast writers who published more than a book a year between 1870 and 1929, with a random sample of "average" professionals, who published a book every two to four years. Further, the spread between parallel anchored corpora made up of prestige works and highly formulaic syndicate writing allows us to place our findings on a readily interpretable scale at every level of analysis.

Keywords: computational humanities, literary history, sociology of literature, narratology, stylistics, cultural production, labor

DRAFT

1 Introduction

In this paper, we attempt to isolate and describe a particular sense of standardization in literature, associated with a major change in authorial practice at the turn of the twentieth century.

By this new regime of creative production, we mean that the writing of prose, and more broadly the creation of all intellectual goods, was affected by changes in manufacturing associated with the late ("second") industrial period—changes usually understood in relation to material goods alone. These included the division of composition into repeatable operations, the introduction of formulas and templates, and the increasing use of prefabricated textual components.

But does it matter how a thing is made? At the time, critics worried about the declining quality of American letters, evidencing an emerging anxiety about industrial manufacture. They believed this new style of fast fiction to be less diverse, more formulaic, full of cliché and hackneyed prose, advancing along tired familiar plot lines. Were such intuitions correct? Our concern here lies less in the quality of the product than in the mode of production. Namely, does writing quickly leave a detectable mark on the prose itself? Did mass-produced fiction become more regular, by some detectable measure, within and across works, in ways that correspond to what contemporary critics believed they were seeing? Did fast fiction, like fast fashion, bear the marks of its manufacture? And if it did, how do we make sense of such measures in relation to some recognizable baselines, against which our observations can be interpreted?

Such questions about standardization are complicated by the fact that measures of redundancy or surprise are sensitive to many other sources of regularity. Literature naturally repeats itself because authors share language, genre, theme, plot, setting, literary convention, or historical sources. Prose may also become more predictable through changes in taste or literary fashion. Any historical comparison must therefore distinguish these regular old regularities from those associated with a shared regime of production. The comparison then has to separate redundancies within individual works and writers from those without.

© 2026 by the authors. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).

* First and second authors writing alongside members of Columbia's Narrative Intelligence Lab: Leyee Dai, Linda Qu, and Alisa Gulyansky (conceptualization, writing, review and editing).

We measure redundancy at four levels of textual organization: vocabulary, phrasing, syntax, and plot. Our method involves first isolating a representative sample of 110 American "super-producers"—fast writers who published more than a book a year in the period between 1870 and 1929, typically some forty books over a career, drawn at random from bibliographies of the period. Subsequently, we compare them with 110 "average" professionals, who published a book every two to four years.

To establish a sense of a meaningful difference, we contrast these metrics against a pool of 51 "ordinary" American novels reviewed in the *New York Times* between 1907 and 1930. Their scores provide the zero point of every scale. On another axis of comparison, we run the same metrics against a canon of "prestige" writers: novels by authors with an entry in *The Norton Anthology of American Literature* who also have a volume in the Library of America's main series. Finally, we set two benchmarks of highly "formulaic" writing from the Stratemeyer Syndicate: its series books of 1906–1930, and *Nancy Drew* (1930–1978) specifically, both convincing examples of highly formulaic genre fiction, written under a production regime where standardization is well documented.

The impact of writing fast or slow can finally be read against these benchmarks, on a scale where ordinary, formula, and prestige fiction are also visible.

2 Historical Background

The term mass production is traditionally associated with the manufacture of physical goods, and particularly with the systematized, large-scale fabrication of interchangeable parts in the United States during the interwar period. Writing for the *New York Times* in 1926, Henry Ford insisted that mass production described a method rather than "merely quantity production," defining it instead as "the focussing upon a manufacturing project of the principles of power, accuracy, economy, system, continuity and speed" [7].

Historians of industrialization have long shown that mass production, in this sense, emerged gradually rather than all at once. In his influential work on the topic, David Hounshell traces its central features back into the nineteenth century. At the core of this regime lay the logic of the assembly line: break manufacture into discrete operations, regularize the parts, formalize procedure, automate where possible. What emerged finally wasn't simply a means of making more goods, but a different way of making them, signifying a reorganization of manufacture around a governing philosophy of standardization [4; 8; 18].

Less obvious are the impacts of industrial standardization on textual production. According to previous estimates, authorial productivity grew dramatically in this period [24]. Where an author of ten to twelve novels was considered prolific in the 19th century, an author of pulp fiction routinely finished the same volume in a year. Our median super-producer published 1.5 books of fiction a year, which is about four times as many as the median professional author of the time, measured at 0.4 novels per year. In the extremes, writers reached ten times Dickens's pace: Gilbert Patten, writing as Burt L. Standish, published some seventeen books of fiction a year, 547 in all; Prentiss Ingraham twelve a year; and Edward L. Wheeler ten (Figure 1).

What made such a drastic increase in productivity possible?

Metaphysical appeals to the spirit of the age aside, we propose a simpler and more materialist explanation: literary labor was reorganized along many of the same lines that Ford and Hounshell describe in relation to American industrialism. Just as it was with sewing machines or bicycles, greater literary output depended on changing how writing was done. A number of new authorial techniques emerged to meet the demands of sustained, high-volume composition. The author's handcraft, romanticized by the Victorians and high modernists alike, was thereby systematically split into modular operations, facilitating standardization and reuse—mirroring the shift from artisanal to industrial manufacture in other domains.

DRAFT

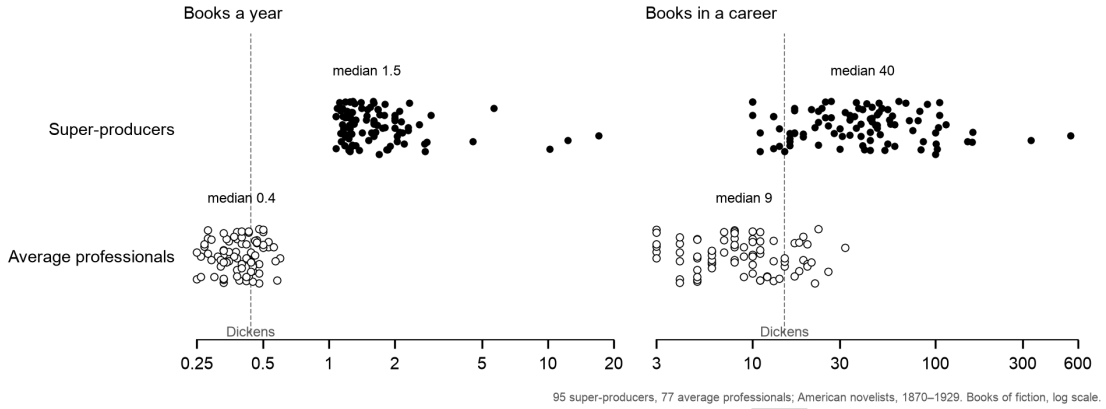

Figure 1: Books of fiction per year and per career, super-producers and average professionals.

In retrospect, we may reduce the numerous emerging techniques of literary mass production to several related practices. Just like the auto workers, writers broke the creative process into smaller, repeatable operations. To write better (in some sense), they used style guides and proven plot formulas. To write faster, they used templates and plot skeletons, augmented by "stock lists" of characters, plot devices, and "situations" meant to be shared and reused across different works. These writing practices were finally coordinated into teachable writing systems, with the explicit aim of increased authorial productivity. Seen in this light, the extraordinary increase in creative output followed from changes in the organization of literary labor itself. However, the impact of these changes is scarcely understood in empirical terms [cf. 1; 11].

3 Main Corpus

Both advocates and critics of literary mass production assumed that its effects were visible in the prose.

We test that assumption by comparing two distinct groups of writers, fast (or "super-producers") and slow (our "average professional" authors). Further, parallel "anchor" corpora calibrate our findings against other populations, such as "ordinary reviewed," "prestige," and "syndicate" pools.

Fast and slow authors. We calculate literary throughput in terms of books of fiction per active year, from a writer's first book to their last. We anchor this scale on Dickens, already prolific by professional standards at about fifteen novels in thirty-four years (0.44 a year) [20, ch. 8]. Super-producers write at twice his rate or more, representing the top 2.5% of sampled authors by rate. Average professionals publish 0.25 to 0.6 a year, representing authors between the 83rd and 94th percentiles, within a frame dominated by the long tail of authors who published one or two books total in their career and not considered in this study. The comparison isn't therefore one between exceptionally slow and exceptionally fast authors, but rather between those who are exceptionally productive and those who are average but still fairly active.

Frame and draw. Candidates are American writers with a readable novel in HathiTrust, 1870-1929, their output counted by cross-referencing publicly available catalogs. The difficulty of this sampling strategy comes in verifying low or average output. Where positive evidence of publication can count definitively toward the author's totals, evidence of low or average output can be the result of catalog gaps.

For this reason, each writer was placed on the count least favorable to their group. A super-producer was counted on books alone, a lower bound. An average professional is counted on books plus magazine novels and short stories (ten stories to a book), an upper bound. An average

DRAFT

candidate whose upper count passed 0.6 a year was dropped.

Further, in our main comparison groups, we drew one book per writer at random from the middle third of the writer's novelistic career. Copies come from public digital collections. Differences in digitization were controlled throughout, with additional steps taken to avoid the potential confounders resulting from copy source.

4 Scale Calibration

Any metrics derived from the main corpus of fast and slow authors must be understood in reference to some other body of work, in search of "a difference which makes a difference" [2, p. 459].

Consequently, we compare our metrics against four reference corpora, where our literary-historical intuitions are strongest.

Ordinary novels. The first of these we call "ordinary reviewed works," represented by 51 American novels reviewed in *The New York Times*, 1907–1930. Following the logic of Underwood and Sellers's study of literary prestige, which treats a review in a selective periodical as a mark of distinction [22], we choose the *Times* for the opposite reason: it was a broad venue. It reviewed about half of all fiction listed in the *Book Review Digest* (52.5%), while a selective weekly such as the *Dial* reviewed about a tenth (9.7% a year). A review in the *Times* thus marks visibility rather than distinction.

The prestige canon offers us a baseline for works recognizably belonging to the "high literary" canon. Most of its authors were not especially fast, with notable exceptions such as Henry James and Jack London. To avoid personal bias, we set a strict inclusion standard, selecting novels by authors with an entry in *The Norton Anthology of American Literature* who also appear in the Library of America's main series. Thus, whatever their productivity, these authors are canonical by external standards, in a list that includes Edith Wharton, Frank Norris, Theodore Dreiser, William Faulkner, and John Dos Passos among others. Random draws give up to four novels per writer, 62 books by 20 writers, 1890–1930.

At the other end of the prestige spectrum, we find a corpus of **Syndicate fiction**, produced by the Stratemeyer Syndicate in long-running series for young readers. The Syndicate's own correspondence documents a fiction factory, in which several writers expanded the Syndicate's outlines to exacting specifications, often to deadlines of a few months [19; see also 9]. These books were undoubtedly fast and recognizably formulaic in their critical reception. Our other work on the Syndicate shows the *Nancy Drew* series to have the highest rate of convergence across its ghostwriters [16].

The spread between prestige works and the Syndicate allows us to place our findings on a readily interpretable scale, in accordance with expert intuition, at every level of analysis.

Sample size. We expected super-producers to sit about halfway between canon and Syndicate, some 0.4 standard deviations from average professionals; detecting that gap takes about 100 writers per group, and we targeted 110.

5 Instrument

Vocabulary, phrasing and syntax inside the book. To find repetitive, formulaic, and otherwise subjectively redundant works, we borrow from the literature on information theory to develop a method which might be best called "conditional self-compression" [3; 5; 14].

It involves first, taking two random non-overlapping text segments of 7,500 words each from a single work. We then compress each separately and together using the LZMA2 compressor [17; 25]. Once the compressor "sees" the first segment, it is able to compress the second segment more tightly—giving us a measure of predictability from early to late segments, expressed in bits saved.

DRAFT

Conditional self-compression has the advantage of measuring multiple levels of textual redundancy using a single instrument, allowing us to highlight some patterns of redundancy while suppressing others. For example, by initially shuffling a novel's words, we obtain a measure of pure **lexicon** or vocabulary, independent of word order.

When we take the initial ordered measurements and subtract the shuffled compression, we can be sure the remainder is due primarily to the effects of the word order alone. The subtraction (after Montemurro and Zanette [15]) discovers **shared phrasing** of arbitrary length, mapping well to what the reader would perceive as stock description or cliché: "he said with a grin" or "as a matter of fact."

In our third compression measurement, we replace each word with its part-of-speech tag, according to the Penn Treebank taxonomy. Compression and shuffled subtraction on part-of-speech tags give us a coarse sense of **syntax** effects, which a reader would perceive as uniformity in sentence structure.

Plot repetition requires a different instrument, which we call the "event-verb profile," validated against a corpus of known plot doppelgangers. The profile works as follows:

1. Summarize and reduce. A language model summarizes each book's passages, concentrating on events, with character names replaced by pseudonyms. From evenly spaced summaries we keep only the verbs, prune common commentary verbs (*highlight*, *symbolize*), and weight the rest by their rarity across all books. As with conditional self-compression, this step masks confounders such as theme or style, reducing each story to the event verbs that carry the plot.

2. Compare. The initial sample is compared, by mean cosine similarity, with the same writer's neighboring novels, published in close succession before and after it.

3. Validate. Across 110 pairs of texts and their versions (translations, abridgments, close rewrites and retellings) hidden among 960 books, the instrument ranked a text's partner above unrelated books 96% of the time (AUC 0.96), and first in 68% of cases.

In addition, we gated each instrument against the Syndicate's *Nancy Drew* series, where we expect the results of extreme standardization, at the opposite end of the scale from the canon. Each instrument had to pass the threshold of Cohen's "large" effect, $d = 0.8$, a difference of 0.8 standard deviations, a demanding and conventional measure [6, p. 157].

All four instruments pass, the plot instrument with the least certainty, since each writer contributes only a few neighboring novels.

6 Preliminary Results

When does a difference become meaningful? Before any book was scored, we estimated that samples of this size could detect a difference as small as a third of a spread, where one spread is the standard deviation among ordinary novels. In phrasing, the Syndicate sits more than a full spread from ordinary novels, three times that threshold. We report the chance of a meaningful difference (the probability of practical significance) beside each interval, indicating the share of bootstrap resamples more than a third of a spread from zero, in either direction. Under flat priors, it approximates the posterior probability that the difference lies outside the region of practical equivalence [10, p. 336; 13, p. 4].

A careful sampling strategy, set against a well-validated instrument array, produces calibration. We can now report the difference between fast and slow authors, not in arbitrary numerical terms, but in relation to known landmarks (Figure 2).

DRAFT

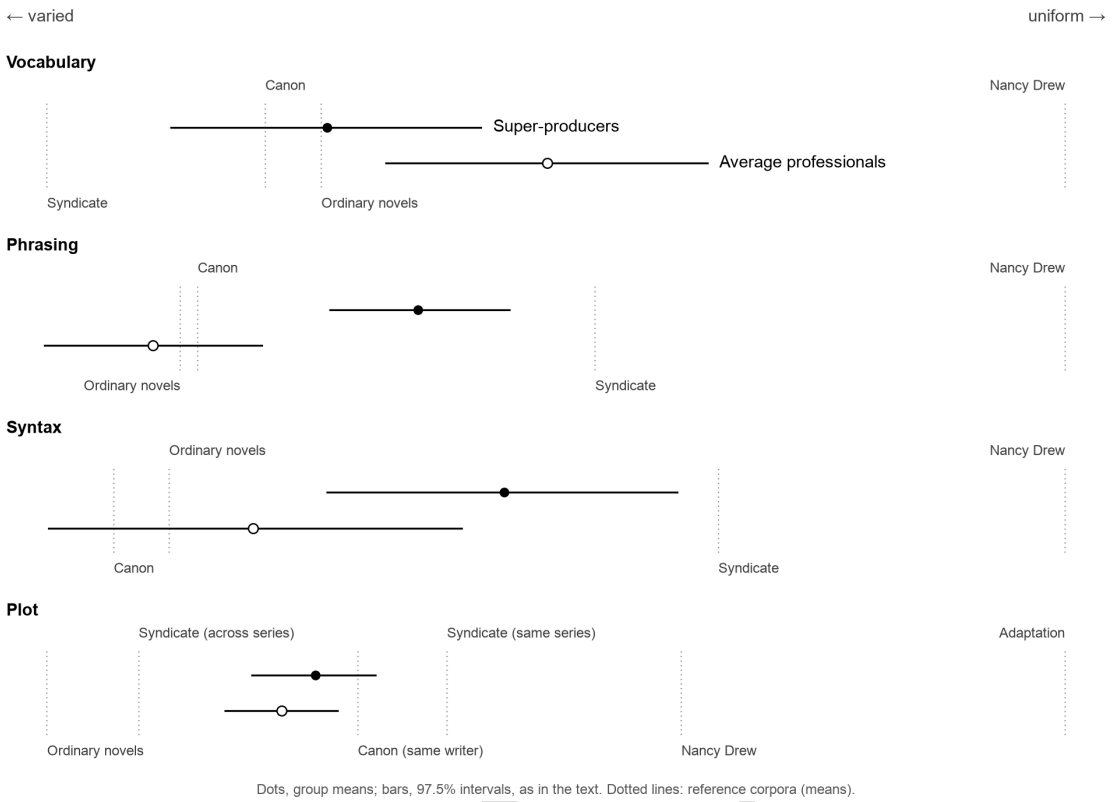

Figure 2: Super-producers and average professionals among the reference corpora, at four levels.

A clear picture emerges from the data, however incomplete at this stage. Our super-producers exhibit a slightly more varied lexicon than average professionals, and roughly on par with ordinary authors reviewed in *The New York Times* (super-producers +0.01 spreads, average professionals +0.32; a difference of −0.27 [−0.58, +0.03]). Here and below, differences are adjusted for the decade of publication, and the brackets give their 97.5% interval, the range our samples leave open.

Part of that move may be explained by the fact that at least 12 of our authors are themselves syndicated, representing umbrella pen names with multiple authors contributing to lexical variation. Separating syndicate from individual super-producers rules this out (solo super-producers +0.01, house names −0.01). Though more work remains to be done on interpreting these results, we suspect the conventions of genre to drive the effect. Fast stories are likely to move fast and cover a lot of ground, evidencing a wider range of settings and situations.

The super-producers drift farthest toward syndicate fiction in terms of shared phrasing and syntax, as expected. Their phrasing sits +0.78 spreads above the average professionals' [+0.35, +1.18], more than halfway to the Syndicate's series; syntax leans the same way (+0.33 [−0.04, +0.68]), though mostly among juvenile books.

Contrary to our intuition, plot shows little difference. Both groups retell their own stories about as much as a canonical novelist does across two books (average +0.23, super-producers +0.26, canon +0.31; a difference of +0.03 [−0.05, +0.11]), well below Nancy Drew's neighboring volumes (+0.62).

DRAFT

	Canon	Syndicate	Nancy Drew	Average	Super	Super − average	Chance meaningful
Vocabulary	−0.08	−0.39	+1.05	+0.32	+0.01	−0.27 [−0.58, +0.03]	31%
Phrasing	+0.05	+1.20	+2.55	−0.08	+0.69	+0.78 [+0.35, +1.18]	99%
Syntax	−0.08	+0.77	+1.26	+0.12	+0.47	+0.33 [−0.04, +0.68]	47%
Plot	+0.31	+0.09	+0.62	+0.23	+0.26	+0.03 [−0.05, +0.11]	9%

Table 1: Preliminary results. Ordinary novels are 0 at every level; plot runs to 1, a text and its adaptation. Chance meaningful: that the difference exceeds a third of a spread.

Measured against the canon, phrasing separates writers by speed, not by prestige: super-producers repeat their phrasing +0.62 spreads more than canonical writers [+0.22, +1.03], average professionals no more (−0.10 [−0.52, +0.35]).

7 Further work

What traces does writing fast finally leave on the prose? Given the calibrated results, we observe the clearest signal of redundancy not in lexicon or plot, but in the "middle" layer of language, among the stock phrases. Super-producers repeat their own phrases far more than average professionals, more than halfway to the Syndicate's series and beyond canonical writers. Their syntax leans the same way, but mostly in books for young readers. Their vocabulary is, if anything, slightly more varied, and their plots repeat no more than a canonical novelist's. An obvious next step for our study is therefore to control for juvenile fiction.

A bigger mystery remains buried in the texture of redundancy. A first look at the recurring runs of four or more words reveals speech formulas, phrases of time and motion, and above all narrative connectives and idioms ("in front of the", "rose to his feet", "as a matter of course"). Yet no phrase marks the group as a whole: each writer repeats their own, where we expected to find a common stock. The closest model for such repetition comes from the study of oral epic. Albert Lord showed that singers composing in performance, at speed, drew on formulas developed over generations and shared among singers [12, p. 22]; Alison Wray describes the same economy of prefabricated phrases in ordinary speech [23, pp. 17–18]. Our fast writers seem to keep a personal stock instead. The full paper will test whether a coarser match, in which "rose to his feet" and "rises to her feet" count as one formula, reveals more that they share.

Finally, our measures register how much writers repeat themselves, inside a book and across neighboring books. Standardization in the industrial sense means more [8; 21]: many producers converging on a common pattern and using prefabricated, interchangeable literary components. Our other work finds such convergence among the ghostwriters of the *Nancy Drew* series [16]. In the full version of this study, we plan to test whether independent fast writers drew on shared components, and to inventory them more thoroughly, which would require a more complete picture of genre conventions, beyond individual works or authors.

8 AI Usage Statement

This work was written under "human first, human last" principles, meaning that the authors take responsibility for all information going in and out of the project. We take credit for primary ideation, study design, literature review, archival research, analysis, authorship, interpretation, and study conclusions. The usage of AI tools is limited to routine preprocessing, research assistance, code automation, scoring, rapid prototyping, document preparation and editorial assistance.

DRAFT

References

[1] Algee-Hewitt, Mark, Allison, Sarah, Gemma, Marissa, Heuser, Ryan, Moretti, Franco, and Walser, Hannah. “Canon/Archive: Large-scale Dynamics in the Literary Field”. Tech. rep. 11. Stanford Literary Lab, 2016.

[2] Bateson, Gregory. *Steps to an Ecology of Mind*. First published 1972. Chicago: University of Chicago Press, 2000.

[3] Benedetto, Dario, Caglioti, Emanuele, and Loreto, Vittorio. “Language Trees and Zipping”. In: *Physical Review Letters 88*, no. 4 (2002), p. 048702. DOI: 10.1103/PhysRevLett.88.048702.

[4] Chandler Jr., Alfred. *The Visible Hand: The Managerial Revolution in American Business*. Cambridge, MA: Belknap Press of Harvard University Press, 1977.

[5] Cilibrasi, Rudi and Vitányi, Paul M. B. “Clustering by Compression”. In: *IEEE Transactions on Information Theory 51*, no. 4 (2005), pp. 1523–1545. DOI: 10.1109/TIT.2005.844059.

[6] Cohen, Jacob. “A Power Primer”. In: *Psychological Bulletin 112*, no. 1 (1992), pp. 155–159. DOI: 10.1037/0033-2909.112.1.155.

[7] Ford, Henry. “Henry Ford Expounds Mass Production”. In: *The New York Times* (Sept. 19, 1926), p. XX1.

[8] Hounshell, David. *From the American System to Mass Production, 1800–1932: The Development of Manufacturing Technology in the United States*. Baltimore: Johns Hopkins University Press, 1984.

[9] Johnson, Deidre. *Edward Stratemeyer and the Stratemeyer Syndicate*. Twayne’s United States Authors Series 627. New York: Twayne, 1993.

[10] Kruschke, John K. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. 2nd ed. Amsterdam: Academic Press, 2015.

[11] Liddle, Dallas. “Could Fiction Have an Information History? Statistical Probability and the Rise of the Novel”. In: *Journal of Cultural Analytics 4*, no. 2 (2019). DOI: 10.22148/16.033.

[12] Lord, Albert B. *The Singer of Tales*. Harvard Studies in Comparative Literature 24. New York: Atheneum, 1965.

[13] Makowski, Dominique, Ben-Shachar, Mattan S., Chen, S. H. Annabel, and Lüdecke, Daniel. “Indices of Effect Existence and Significance in the Bayesian Framework”. In: *Frontiers in Psychology 10* (2019), p. 2767. DOI: 10.3389/fpsyg.2019.02767.

[14] Malyutov, Mikhail B., Wickramasinghe, Chammi Irosha, and Li, Sufeng. “Conditional Complexity of Compression for Authorship Attribution”. SFB 649 Discussion Paper 2007-057. Humboldt University Berlin, Collaborative Research Center 649: Economic Risk, 2007.

[15] Montemurro, Marcelo A. and Zanette, Damián H. “Universal Entropy of Word Ordering Across Linguistic Families”. In: *PLoS ONE 6*, no. 5 (2011), e19875. DOI: 10.1371/journal.pone.0019875.

[16] Narrative Intelligence Lab. “Authorial Style in Syndicate Fiction”. Under review at *Poetics*. 2026.

[17] Pavlov, Igor. “LZMA SDK”. Public domain; LZMA, Lempel-Ziv-Markov chain algorithm.

DRAFT

[18] Scranton, Philip. *Endless Novelty: Specialty Production and American Industrialization, 1865–1925*. Princeton, NJ: Princeton University Press, 1997.

[19] Stratemeyer Syndicate. “Stratemeyer Syndicate Records”. MssCol 2903, Manuscripts and Archives Division, The New York Public Library. 1832–1984.

[20] Sutherland, John. “The Victorian Novelists: Who Were They?” In: *Victorian Fiction: Writers, Publishers, Readers*. Also New York: St. Martin’s Press. London: Macmillan, 1995, pp. 151–164. DOI: 10.1007/978-1-349-23937-5_8.

[21] Timmermans, Stefan and Epstein, Steven. “A World of Standards but not a Standard World: Toward a Sociology of Standards and Standardization”. In: *Annual Review of Sociology 36*, no. 1 (2010), pp. 69–89. DOI: 10.1146/annurev.soc.012809.102629.

[22] Underwood, Ted and Sellers, Jordan. “The Longue Durée of Literary Prestige”. In: *Modern Language Quarterly 77*, no. 3 (2016), pp. 321–344. DOI: 10.1215/00267929-3570634.

[23] Wray, Alison. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press, 2002.

[24] Yi Tenen, Dennis. “The Emergence of American Formalism”. In: *Modern Philology 117*, no. 2 (2019), pp. 257–283. DOI: 10.1086/705672.

[25] Ziv, Jacob and Lempel, Abraham. “A Universal Algorithm for Sequential Data Compression”. In: *IEEE Transactions on Information Theory 23*, no. 3 (1977), pp. 337–343. DOI: 10.1109/TIT.1977.1055714.

DRAFT