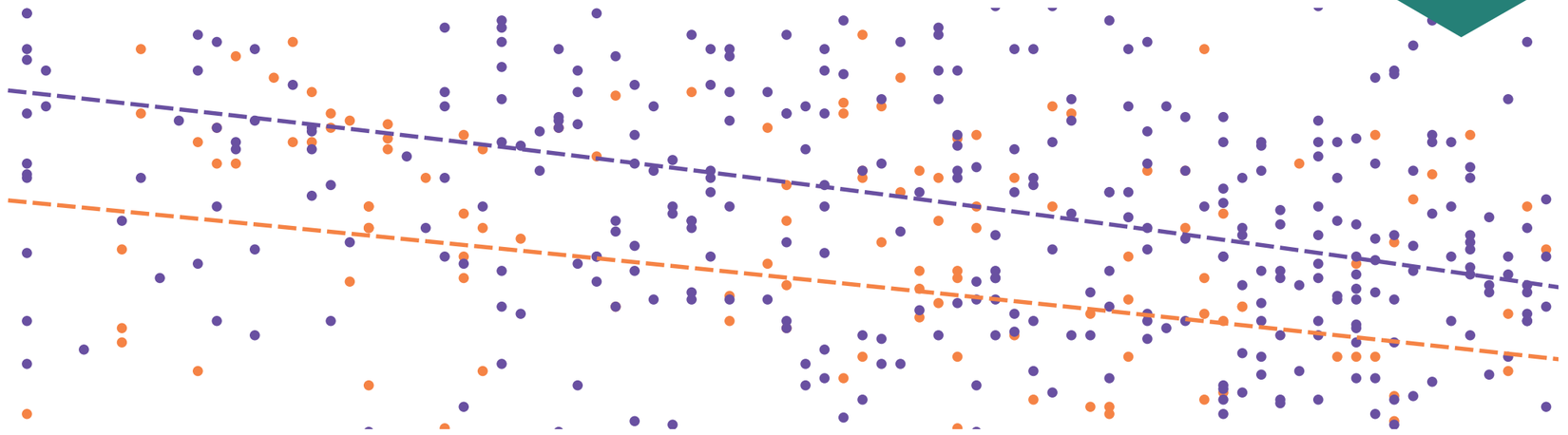
Canon/Archive. Large-scale Dynamics in the Literary Field

Mark Algee-Hewitt
Sarah Allison
Marissa Gemma
Ryan Heuser
Franco Moretti
Hannah Walser



Literary **Lab** **Pamphlet 11** January 2016

Pamphlets of the Stanford Literary **Lab** ISSN 2164-1757 (online version)

Mark Algee-Hewitt
Sarah Allison
Marissa Gemma
Ryan Heuser
Franco Moretti
Hannah Walser

Canon/Archive. Large-scale Dynamics in the Literary Field[1]

I. Sociological Metrics

1. Dowry and vegetables

Of the novelties introduced by digitization in the study of literature, the size of the archive is probably the most dramatic: we used to work on a couple of hundred nineteenth-century novels, and now we can analyze thousands of them, tens of thousands, tomorrow hundreds of thousands. It's a moment of euphoria, for quantitative literary history: like having a telescope that makes you see entirely new galaxies. And it's a moment of truth: so, have the digital skies revealed anything that changes our knowledge of literature?

This is not a rhetorical question. In the famous 1958 essay in which he hailed "the advent of a quantitative history" that would "break with the traditional form of nineteenth-century history", Fernand Braudel mentioned as its typical materials "demographic progressions, the movement of wages, the variations in interest rates [...] productivity [...] money supply and demand."[2] These were all quantifiable entities, clearly enough; but they were also *completely new objects* compared to the study of legislation, military campaigns, political cabinets, diplomacy, and so on. It was this *double* shift that changed the practice of history; not quantification alone. In our case, though, there is no shift in materials: we may end up studying 200,000 novels instead of 200; but, they're all still novels. Where exactly is the novelty?

199,000 books that no one has ever studied – runs the typical answer – how could there *not* be novelties? It's a whole new dimension of literary history.

"We know more about people exchanging goods for reasons of prestige than about the kinds of exchanges that go on every day", wrote André Leroi-Gourhan in *Gesture and Speech*, a few years after Braudel; "more about the circulation of dowry money than about the selling of vegetables..."[3] Dowry and vegetables: perfect antithesis. Both are important, but for opposite reasons: dowry, because it happens once in a lifetime; vegetables, because we eat them every day. And at first sight, it seems like the perfect parallel for the 200 and the 200,000 novels. But as soon as we start looking deeper into the matter, complications arise. Take two historical novels published in the same year of 1814: Walter Scott's *Waverley*, and James Brewer's *Sir Ferdinand of England*. Intuitively, one would associate *Waverley* with the prestige of the dowry, and *Sir Ferdinand* with the humble role of chicory. In fact, though, Scott's novel was both a great formal breakthrough, *and* the book everybody was reading all over Europe: dowry and vegetables, rolled into one. But if that is the case, what difference can all the *Sir Ferdinands* of the digital archive make? We used to know nothing about them, and now we know something. Good. Does this knowledge also make a difference?[4]

Let us illustrate the problem with one of the findings from our own research: the decline of the semantic field of "abstract values" – words like "modesty", "respect", "virtue" and so on – described by Ryan Heuser and Long Le-Khac in "A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method" (**Figure 1.1**). As that punctilious 2,958 makes clear, Heuser and Le-Khac saw the width of the archive as a crucial aspect of their research. Had they studied the old, narrower canon instead, would their results have changed? **Figure 1.2** provides the answer: no. The canon

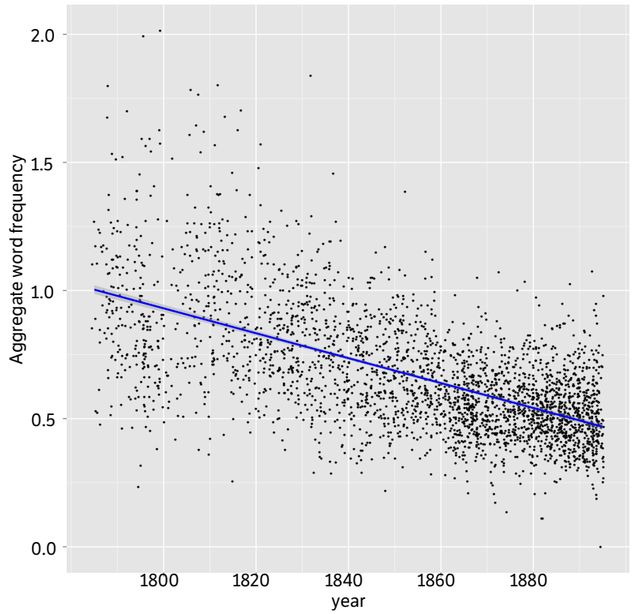


Figure 1.1 Abstract values in British novels, 1785-1900

Ryan Heuser and Long Le-Khac, "A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method", Literary Lab Pamphlet 4, 2012, p. 18.

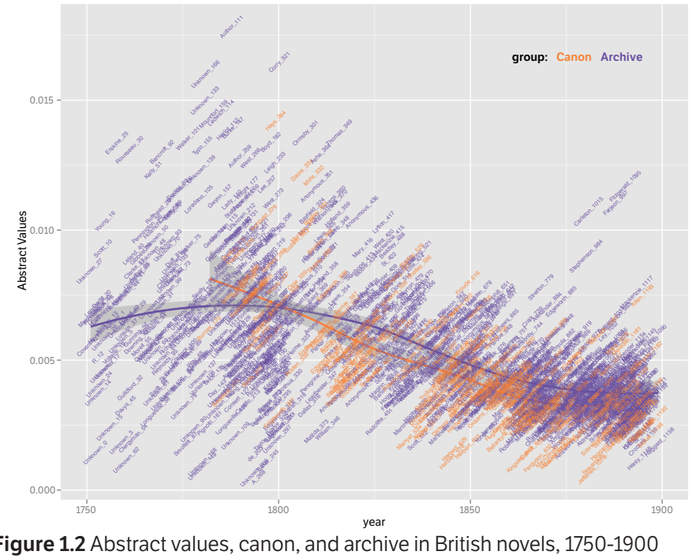


Figure 1.2 Abstract values, canon, and archive in British novels, 1750-1900

In this figure, the canon consists of the 250 novels originally included in the Chadwyck-Healey Nineteenth-Century Fiction Collection. We explain the choice of Chadwyck-Healey in section 3 below.

1 This project has been supported by a grant from the Fondation Maison Sciences de l'Homme of Paris and the Mellon Foundation; the research was conducted in collaboration with a group working at the Sorbonne, in the Labex OBVIL.

2 Fernand Braudel, "History and the Social Sciences: The *Longue Durée*", in *On History*, Chicago 1980, p. 29.

3 André Leroi-Gourham, *Gesture and Speech*, 1965, Cambridge 1993, p. 148.

4 It might not. In a piece forthcoming in a special issue of *MLQ* on "Scale and Value", James English has convincingly argued that a "a sample gathered on the principle that every individual work of new fiction must hold equal value in the analysis" – that is to say, a sample very similar to our "archive" – is actually not very "suitable for a sociology of literary production, where 'production' is understood to mean not merely (or even primarily) the production of certain kinds of texts by authors but the production of certain kinds of value by a social system, whose agents include readers, reviewers, editors and booksellers, professors and teachers, and all the many moving pieces of literature's institutional apparatus." The fact that, when the present pamphlet turned to the study of the archive, it ended up focusing almost exclusively on the "production of certain kinds of texts" seems clearly to corroborate English's thesis. On the other hand, in so far as a "social system" creates "value" not only by assigning it to certain authors or texts, but also by *denying* it to others ("In matters of taste, more than anywhere else, all determination is negation; and tastes are perhaps first and foremost distastes": Bourdieu, *Distinction*), readers and the rest of "literature's institutional apparatus" *are* present in our narrative – but always and only with a destructive role.

precedes the archive by about 15-20 years; but the historical trajectory is the same.

This does not mean that the new archive contains no new information; it means, however, that we must still learn to ask the right type of questions. But before doing so, something needs to be clarified. Canon and archive: what do we mean, by these two words?

2. Bias in the Archive

Let's begin with three preliminary notions: the published, the archive, and the corpus. The first is simple: it's the totality of the books that have been published (the plays that have been acted, the poems that have been recited, and so on). This literature that has become "public" is the fundamental horizon of all quantitative work (though of course its borders are fuzzy, and may be expanded to include books written but kept in a drawer, or rejected by publishers, etc.) The archive is for its part that portion of published literature that has been preserved – in libraries and elsewhere – and that is now being increasingly digitized. The corpus, finally, is that portion of the archive that is selected, for one reason or another, in order to pursue a specific research project. The corpus is thus smaller than the archive, which is smaller than the published: like three Russian dolls, fitting neatly into one another. But with digital technology, the relationship between the three layers has changed: the corpus of a project can now easily be (almost) as large as the archive, while the archive is itself becoming – at least for modern times – (almost) as large as all of published literature. When we use the term "archive", what we have in mind is precisely this potential convergence of the three layers into one; into that "total history of literature", to borrow an expression from the *Annales*, that used to be a mirage, and may soon be reality.

This, in theory. In practice, things are not so simple. Take the present project. Its initial corpus consisted of about 4,000 English novels from 1750 to 1880; for the eighteenth century, they came from ECCO; for the nineteenth, from the Chadwyck-Healey Nineteenth-Century Fiction corpus and the Internet Archive of the University of Illinois.[5] By the old standards of literary history, 4,000 novels were a very large corpus; but its actual coverage turned out to be quite uneven. For the period 1770-1830, for instance, we had about one third of the titles listed in the Raven-Garside-Schöwerling bibliography; for the later nineteenth century, however, the percentage was much lower, around 10%. The same for specific genres: we held 96% of Adburgham's silver-fork bibliography, but only 77% of Gallagher's industrial novels, 53% of Stevens' historical novels before Scott, and 35% of Perazzini's gothic bibliography.[6]

Clearly, these were slippery statistical grounds. Compared to the handful of texts usually considered canonical, our 190 gothic novels were a very large number, and it was tempting to identify them with the archive *tout court*; but were they truly representative of the "population" of the English gothic as a whole? Almost certainly not; simplifying somewhat, a sample is representative when it has been randomly chosen from a given population; but our 190 novels had definitely *not* been chosen that way. Ultimately, they all came from a few great libraries – and libraries don't buy books in order to have representative samples; they want books they consider worth preserving. *Good* books; good, according to principles that are likely to be similar to those that lead to the formation of canons. Though our corpus was twenty times larger than the traditional canon, then, it was perfectly possible that its principle of selection *would make it resemble the canon much more than the archive as a whole*. That was the problem.[7]

We wanted our results to be reliable, hence we generated a random sample of the field to be studied: 507 novels *tout court* for the period 1750-1836, 82 gothic novels, and 85 historical novels before Scott.[8] All in all, 674 novels. In the digital age, this wouldn't take long.

We generated the sample at the end of the school year, in June 2014. Then we turned to our own database, where we found 35 of the 82 gothic novels, 35 of the 85 historical novels, and 145 of the 507 novels from the Raven-Garside bibliographies. In early July, we passed the list of the titles we had not found – roughly 460 – to Glen Worthey and Rebecca Wingfield, at the Stanford Libraries, who promptly disentangled it into a few major bundles. Around 300 texts were held (in more or less equal parts) by the Hathi trust and by Gale (through NCCO and ECCO II).[9] Another 30 were in collected works, in alternate editions, concealed by slightly different titles, in microfiche or microfilm collections, etc.; about 100 existed only in print, and of 10 novels there were no known extant copies. In August, requests were sent to Hathi and Gale – with both of which Stanford has a long-standing financial agreement – for their 300 volumes. Of the 100 novels existing only in print, about half were held by the British Library, in London, which a few months earlier had kindly offered the Literary Lab a collection of 65,000 digitized volumes from its collections; unfortunately, none of the books we were looking for was there. The special collections at UCLA and Harvard, which held about 50 of the books, sent us a series of estimates that ranged (depending, quite reasonably, on the conditions of the original, and on photographic requirements which could be very labor-intensive) from $1,000 to $20,000 per novel; finally, six novels were part of larger collections held by Proquest, and would have cost us – despite Proquest's very generous 50% discount – $147,000, or $25,000 per title.[10]

Remember: this was a search involving many excellent librarians in London, Cambridge, Los Angeles, and of course at Stanford; a half dozen researchers at the Literary Lab; plus people at Hathi, Gale, and so on. The books we were looking for were only two centuries old; they had had print runs of at least 750-1,000 copies, and in a part of the world which, at the time, already possessed efficient libraries. The Literary Lab has some money for research (though, make no mistake, not *that* kind of money). In other words, one could hardly hope for better resources. And yet it took about six months to receive from Hathi and Gale the set of texts that should have allowed us to move from the initial 30%, to around 70-80% of the random sample:[11] a figure which

5 See https://archive.org/details/19thcennov. ECCO (Eighteenth Century Collections Online) is a two-part digital collection of 18th century materials, based on the English Short Title Catalogue (ESTC), and sourced from a number of libraries in the US and UK; part II of ECCO is an update, consisting of texts or editions that were not available when the original ECCO was released.

6 Alison Adburgham, *Silver Fork Society*, London 1983; Catherine Gallagher, *The Industrial Reformation of English Fiction*, Chicago 1985; Anne H. Stevens, *British Historical Fiction Before Scott*, London 2010; Federica Perazzini, *Il Gotico @ Distanza*, Roma 2013.

7 To complicate matters further, different genres have different canon-to-archive ratios: whereas epistolary and silver-fork novels have relatively large archives and small canons, the opposite is true of the industrial novel and the *Bildungsroman*, both of which attracted many major Victorian writers; while the two super-genres of gothic and historical novels lie somewhere in between the two extremes. On this – and much else – we need a lot more empirical evidence.

8 This last group was not a random sample: since Anne Stevens' bibliography included only 85 pre-Scott historical novels, we decided to look for all of them.

9 HathiTrust is a partnership of major research libraries, which serves as a repository for digital collections; these include volumes scanned as part of the Google project and the Internet Archive, as well as other smaller local projects. Gale's NCCO (Nineteenth Century Collections Online) is a digital collection of 19th century materials, usually sourced from major collections, and ranging across disciplines (literature, science and technology, photography, etc.) Thus far, there are twelve parts to NCCO, one of which consists of the Corvey novel collection; unlike ECCO, NCCO is not based on a standard bibliography in the field, so it's hard to predict what is being added. Gale is a large conglomerate of information and education services – run as a for-profit business – that sells content and services to libraries; it publishes both print works (reference and fiction) and electronic collections (ECCO, NCCO, and others). Its parent company is Cengage Learning, which defines itself as "a leading educational content, technology, and services company for the higher education and K-12, professional and library markets worldwide".

10 To these figures one should add what the Stanford libraries have paid for ECCO, ECCOII, and NCCO to begin with: with the usual generous discounts, something like one million dollars for the three collections. ProQuest is another for-profit education service whose products include the Historical Newspapers series, Literature Online, Dissertation Abstracts, and others. Its parent company is Cambridge Information Group.

11 "*Should* have allowed", because receiving a text from these collections is not the same as being able to work on it. Much of the data from Chadwyck-Healey and ECCO I used to be delivered on tape, in formats requiring drives that are both hard to find and difficult to use; more "convenient" data deliveries (such as network data transfer, or on external hard drive) have their own problems, ranging from the vagaries of mail systems to bizarre firewall incompatibilities and odd documentary requirements of usage agreements. (Most of Stanford libraries' licensing agreements, for instance, used to be quite vague on the subject of text-mining, or sharing outside the Library preservation structures; over the past five years libraries have explicitly insisted on

would probably make many of our findings questionable, as the missing 20-30% would be, almost by definition, furthest from all conceivable forms of canonization.

Clearly, the idea that digitization has made everything available and cheap – let alone "free" – is a myth. As we became slowly aware of this fact, we decided to start working with a selection from the corpus we had: a database of 1,117 works, 263 from Chadwyck-Healey, and 854 from various archival sources. Initial results took us quickly in one direction; new findings added further momentum; and, by the time the (near-)random sample was (almost-) available, we were too involved in the work to re-start from zero. We don't present this as an ideal model of research, and are aware that our results are weaker as a consequence of our decision. But collective work, especially when conducted in a sort of "interstitial" institutional space – as ours still is – has its own temporality: waiting months and months before research can begin would kill any project. Maybe in the future we will send out a scout, a year in advance, in search of the sample. Or maybe we will keep working with what we have, acknowledging the limits and flaws of our data. Dirty hands are better than empty.

3. From the Canon to the Literary Field

If the selection of our archive was determined by historical library practices (which novels were on the shelves? which were easy to digitize?), that of our canon was a matter of critical judgment – though not our own. The first canon we turned to in this project, the Chadwyck-Healey Nineteenth-Century Fiction Collection, was designed by an editorial board of two, Danny Karlin and Tom Keymer.[12] It is a set of about 250 novels chosen for being so very worth preserving, and so valuable to scholars, that libraries would pay for digital access to the set.

Compiled in the late 1990s, with new novels added subsequently, the marketing materials of the Nineteenth-Century Fiction Collection claim that it "represents the great achievements of the Victorian canon and reflects the landmarks of the period," while also covering "many neglected or little-known works, most of them out of print or difficult to find." From 1794, for example, the collection includes Ann Radcliffe's *Mysteries of Udolpho* and William Godwin's *Caleb Williams*, but also Jane Austen's *Lady Susan* (a very short novel probably written around then, but published posthumously in 1871), and Thomas Holcroft's radical *Adventures of Hugh Trevor*. The first two are obvious choices; the other two less so. It seems that selecting 250 texts makes room for lesser-known novels of critical or historical importance: not only the six major Austen novels, but also *Lady Susan*; not only Godwin, but also Holcroft. In so far as we understand a "canon" to signal a relatively small number of texts selected and consecrated for close study, Chadwyck-Healey – a major searchable collection immediately available to researchers today –[13] is not a bad proxy.

Still, a proxy it is; and we realized that relying on a single source was the wrong way to think about such a many-sided and elusive concept as that of the canon. In "Between Canon and Corpus: Six Perspectives on 20th-Century Novels" (Literary Lab Pamphlet 8, 2015), Mark Algee-Hewitt and Mark McGurl had addressed a similar problem by presenting several lists of "best twentieth-century novels" selected by very different groups, and then analyzing their varying degrees of proximity. We followed a different path, which led us from Chadwyck-Healey's short catalogue of books to two long lists of authors: those mentioned by the *Dictionary of National Biography*, and those listed as "primary subject author" for twentieth-century academic articles indexed by the MLA Bibliography; in a lateral project, we also added the texts included in the Stanford Ph.D. exam lists of the last 30 years. In doing so, we were neither looking for the "right" definition of the canon (which none of them was), nor hoping that the *DNB*, MLA, and Stanford would agree with each other (which they didn't).[14] Rather, these different measurements were meant to replicate the multiple aspects of the idea of the canon: the fact that the national culture (*DNB*) defines it in one way, and international scholarship (MLA) in a somewhat different one; that it may be conceived of as a series of personalities (*DNB* and MLA), or as a collection of texts (Ph.D. lists). The specific choices remained questionable – of course! – but the criteria that we had followed would be multiple, explicit, and measurable. That was the novelty.

Then, we realized that there were other features of the novelistic field that could enter the equation. In their bibliographies, Raven and Garside had for instance identified the novels which had been reprinted in the British isles, or translated into French and German between 1770 and 1830; and one could envisage similar data for future research – from print runs to presence in circulating libraries and more. In these cases, too, the criteria would be multiple, explicit, and measurable; but with a major difference from the *DNB* and MLA. Reprints and translations measure the appeal of novels for a "general" audience, and through the institutions of the literary market; *DNB* and MLA focus on "specialized" readers, and institutions of higher education. One measures the "popularity" of novels; the other, their "prestige".[15]

Popularity and prestige. With this conceptual pair, our research found itself on the same terrain as Bourdieu's path-breaking chart of the French literary

---

the inclusion of text-mining rights in current licenses, but previous agreements remain in a gray area).
Finally, extracting data from an ocean of tape or hard drive, with insufficient or incorrect metadata and no database to assist, is a truly Byzantine process. The Libraries would search the ECCO database – for instance – using Gale's search interface, and citing its URL as that interface instructs. But for the Libraries to get a raw file to the Lab, they need to go through a couple of hard drives (or tapes) containing hundreds of thousands of directories named only with series of random numbers; the metadata "manifest" that Gale delivers with these raw files is contained in about ten Microsoft Word files formatted as if for print: two columns, authors in bold, very basic catalog data, a document ID, and ESTC ID, and a directory path. These documents are immense: ECCO II, Literature and Language module, Authors L-Z – which represents about 1/10th of the ECCO II delivery – is a 2,750-page document. Second, the ID numbers included are *not* the ones that you see in the Gale interface; they are internal, invisible numbers. So, despite all the Lab's work in identifying ECCO sources using the database and noting the official Gale ID number, the Libraries have had to re-search each item by author or title in order to find the name of a file to copy: that Gale ID number is not included *at all* in the file manifest. "My lesson", concluded a research librarian who assisted through the whole process, "is this: even when we've found the file you need, we still haven't really *found* the file".

12 Personal communication with Steven Hall confirmed that the editors were unconstrained in their choice of texts.

13 Provided, that is, that said researchers belong to an institution with the necessary resources. According to one university's ProQuest representative, in the entire world there are only "over 600" universities which subscribe to the Literature Online (LION) database.

14 Even leaving aside the representativeness of the Stanford Ph.D. exams, the author-centered approach of the *DNB* and MLA places Scott's *Castle Dangerous*, or Thackeray's *Catherine*, on the same plane as *Waverley* and *Vanity Fair*, which cannot be right. But alternative criteria have similar flaws, or are impossibly time-consuming.

15 That popularity is measured on nineteenth-century data, and prestige is derived from twentieth-century sources, is of course a problem. Twentieth-century studies have it better in this respect: in "Becoming Yourself: the Afterlife of Reception" (Literary Lab Pamphlet 3, 2011), for instance, Ed Finn charted the position of contemporary authors in the American literary field by using two categories – "consumption" and "conversation" – that belonged to the same chronological frame: "consumption" derived by Amazon.com "also bought" data, and "conversation" by contemporary reviews. Interestingly, "consumption" and "conversation" align rather well with our "popularity" and "prestige"; while the six "canons" discussed by Algee-Hewitt and McGurl also gravitate around market success on one side, and more "qualified" cultural selection on the other.
In an attempt to correct the discrepancy between nineteenth- and twentieth-century data, follow-up studies may enlarge the prestige metrics by taking into account textbooks and anthologies for the school (as Martine Jey is doing for France), prizes (James English, *The Economy of Prestige*), reviews from eighteenth- and nineteenth-century periodicals, or early collections of novels such as Barbauld's, Ballantyne's, and Bentley's. It is by no means certain, however, that collections and reviews should be seen as indicators of prestige, rather than as mere cogs in the developing novelistic market; in an interesting recent essay, Michael Gamer has made a case for both possibilities, by presenting them as having canonical ambitions, while also competing in the commercial market. (See "A Select Collection: Barbauld, Scott, and the Rise of the (Reprinted) Novel", in Jillian Heydt-Stevenson and Charlotte Sussman, eds, *Recognizing the Romantic Novel*, Liverpool 2008.) William St Clair, for his part, has expressed unambiguous skepticism about the role of reviews ("in general, the influence of the reviews appears to have been greatly exaggerated both at the time and by subsequent writers [...] I can discern no correlation between reviews, reputations, and sales"), and about the concept of novelistic prestige in the early 19th century: "As far as the prose fiction of the romantic period is concerned, there was no recognized contemporaneous canon. Indeed, the whole notion of a canon made little sense when most novels were published anonymously. One author dominated the age, 'the author of *Waverley*', not publicly acknowledged to be the famous poet Sir Walter Scott until the mid-1820s." See William St Clair, *The Reading Nation in the*

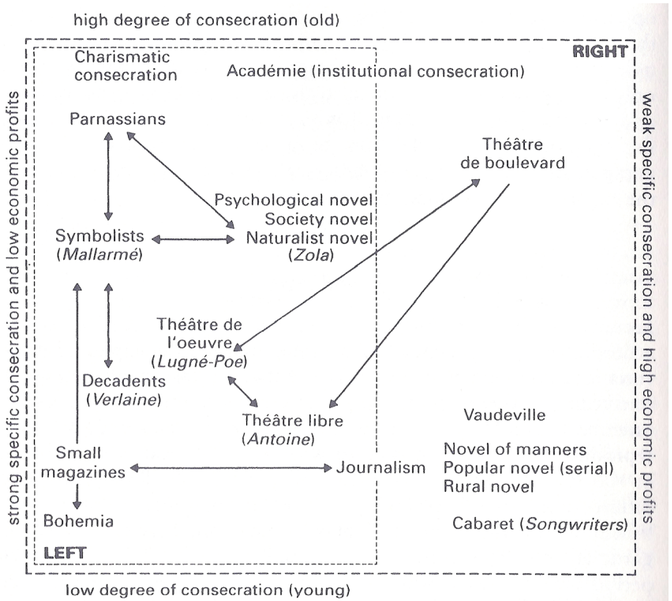


Figure 3.1 The French literary field at the end of the nineteenth century

Bourdieu's diagram of the literary field, though wonderfully suggestive, offers no empirical evidence for the specific position of the various genres and movements. The absence of explicit and measurable criteria is probably the reason why – despite its elegance, and its wide influence – Bourdieu's chart has never become a genuine research tool, replicated and adapted by other scholars. The hard-to-believe regularity of the distribution, so unlike those of **Figures 3.2** and **3.3**, and of Bourdieu's own diagrams in Distinction, is itself probably a consequence of the speculative foundation of the diagram. Pierre Bourdieu, *The Rules of Art: Genesis and Structure of the Literary Field,* 1992, Stanford 1996, p 122.

field (**Figure 3.1**). By placing popularity data on the horizontal ("high/low economic profits") axis, and prestige ones along the vertical ("high/low consecration") one, we could provide a "British" version of Bourdieu's chart. For now, this covered only a single genre, and a handful of decades; but at this

Romantic Period, Cambridge 2004, p. 189.
On the other hand, the existence of a relationship between reviews and reputation has been recently – and convincingly – proposed by Ted Underwood and Jordan Sellers in "How Quickly Do Literary Standards Change?" http://figshare.com/articles/How_Quickly_Do_Literary_Standards_Change_/1418394. Underwood and Sellers study poetry instead of novels, and start their investigation in 1820, when St Clair's book and our own corpus more or less end; too much of a mis-match in object and time frame for a direct comparison. But we are slowly approaching the moment when evidence from independent studies may be successfully compared and integrated.

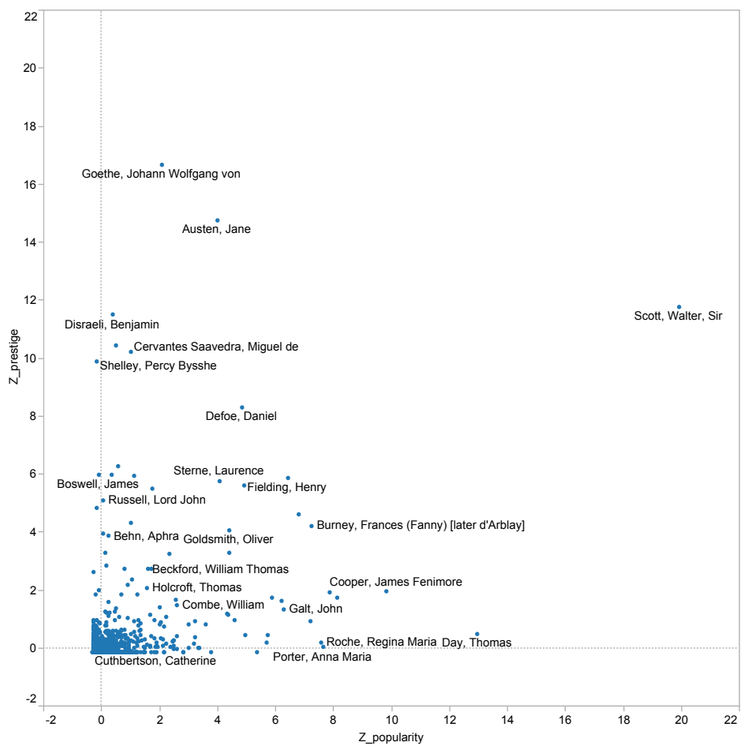


Figure 3.2. The British novelistic field, 1770–1830

Results for the popularity axis are based on the number of reprints (in the British isles) and of translations (into French and German); for the prestige axis, they are based on the number of mentions as "primary subject author" in the MLA Bibliography, and on the length of DNB entries.

The position of writers is determined by the number of standard deviations above the mean of the field; John Galt, for instance, is 7.5 standard deviations above the mean on the popularity axis, and 1 above the mean on the prestige axis; at the opposite extreme, Percy Shelley is 10 standard deviations above the mean in terms of prestige, but slightly below the field's mean in terms of popularity.

point, an empirical cartography of the literary field was no longer a daydream (**Figure 3.2**).

In **Figure 3.2**, all data are dwarfed by Walter Scott's incredible scores: only two novelists are slightly higher than him on the prestige axis (Goethe and Austen), and no one is even close in terms of popularity: the next author along that axis – Thomas Day, author of the Rousseauian bestseller *The History of Sandford and Merton* (1789) – is seven standard deviations below Scott.[16] Once the out-of-scale results of "the author of *Waverley*" are removed

16 Since we are not measuring print runs, the chart actually *understates* Scott's popularity: whereas most contemporary novels had a first run of 1,000 copies, the first three *Waverley* novels had opening runs of 6,000, 8,000, and 10,000 respectively.

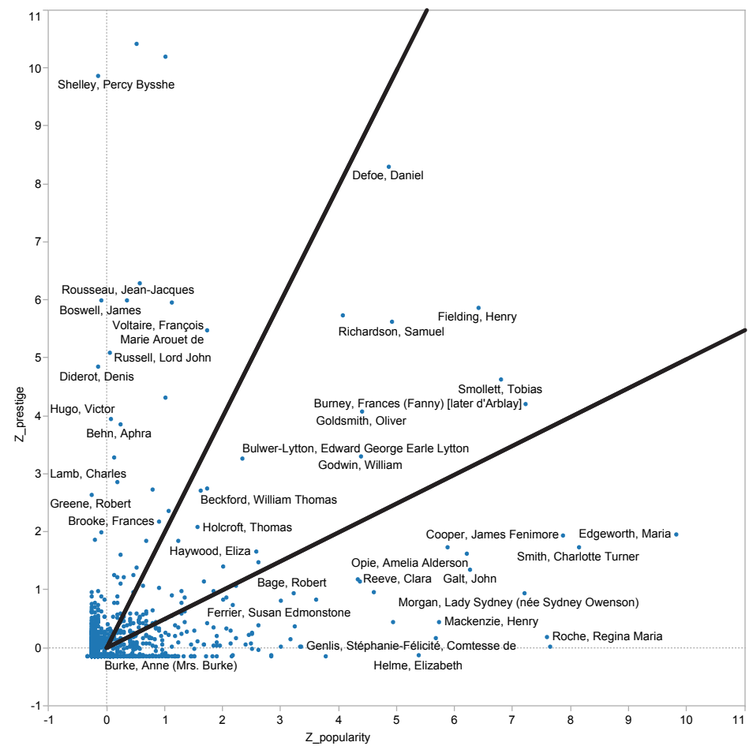


Figure 3.3. The three regions of the British novelistic field, 1770–1830

The three regions of this diagram express variable relationships between popularity and prestige. The area near the vertical axis has prestige scores at least twice as high as the scores for popularity; the area near the horizontal axis is its mirror image, with popularity at least twice as high as prestige; while in the central area the two sets of measurements tend to balance each other.

A study of popularity and prestige on a much larger time-scale is currently in progress at the Literary Lab, directed by J.D. Porter, with data collected both algorithmically and by a team of undergraduate researchers led by Micah Siegel.

from the picture, however, a tri-partition of the British novel becomes clearly visible (**Figure 3.3**).

Let's begin with the group near the horizontal axis: writers with high popularity scores – 5, 8, 10, 13, standard deviations above the average – but quite low on prestige; at most a couple of standard deviation, but often just one, or less. Here we find MacKenzie's sentimental *Man of Feeling* and Day's educational best-seller; the gothic cohort, with their frequent sentimental overtones (Radcliffe, Reeve, Roche, Helme, Maturin), Jacobin and anti-Jacobin novels (Charlotte Smith, Opie), national tales (Edgeworth, Morgan), and the new hegemonic form of the historical novel (Galt, Genlis, Horace Smith, Porter, Cooper). We could call this the space of genre, in the sense of *all* genres:

"the" novel unfolding as a family of distinct forms, whose easily recognizable conventions pave the way to market success. *Waverley*'s opening chapter, entirely devoted to generic allusions in titles, is the perfect symptom of this state of affairs.

Moving "up" from this region to the central part of the diagram takes us into very different territory. If one is ever justified in simply saying, "Here is the canon", this must be the case: Defoe, Richardson, Fielding, Sterne, Goldsmith, Smollett, Burney, Godwin … All of them, clustered in a perfectly balanced space (4-to-7 standard deviations above the popularity mean, and 3-to-8 above the prestige one), where the wide audience of formula fiction blends seamlessly with high cultural recognition. Looking at this central region makes you "see" the process of canonization as the combination of two simultaneous processes: popularity slowly shrinking with the passing years along the horizontal axis – in that respect, most eighteenth-century giants are well below Roche, Porter, Charlotte Smith, and Opie – while prestige increases along the vertical one.[17] Though there is clearly more than one way of becoming a canonical writer,[18] the main lesson of this image is that the canon is not the "the economic world reversed" of Bourdieu's formula for the autonomous literary field; the canon – or at least *this* canon – is made of authors from whom commercial publishers are still expecting to make profits two or three generations after their initial success. And prestige, for its part, is not necessarily *in antithesis* to popularity; here, it seems rather to grow out of it, "distilling" economic returns into something more impalpable, but also more durable. [19]

Things are different in the "high-prestige" region of **Figure 3.3**, which is clearly dominated by foreign writers (Cervantes, Voltaire, Diderot, Rousseau, Goethe, Schiller, Hugo…), or by those British authors who, though they did write at least one novel, or even a few, can hardly be seen as "professional" novelists. Among them are the encyclopedic figure of Samuel Johnson, and the almost equally versatile Horace Walpole; poets like Percy Shelley (and, lower down, Thomas "Anacreon" Moore and James Hogg); the novelist-politician Disraeli and the politician-politician Lord Russell (who published an improbable *Nun of Arrouca* in 1822); essayists like James Boswell and Charles Lamb; at lower prestige levels, the, musician and playwright Charles Dibdin, the playwright and actress Charlotte Cibber Chalke, the economist and travel writer Arthur Young. Among the few novelists-novelists, politics plays an unusually strong role: aside from Russell and Disraeli, we encounter the bluestocking Sarah Scott (*Millennium Hall* and *Desmond*), Mary Shelley, and Hannah More – whose *Coelebs in Search of a Wife*, legend has it, was the only novel Queen Victoria entirely approved of.

With the prestige/popularity diagrams, a first arc of our project had found its natural conclusion. Although, against our original intentions, we had ended up quite far from the archive,[20] our operationalization of the concept of the canon had been both surprising and satisfying: it had brought the notion down to earth, resolving it into the simpler elements of popularity and prestige – or, in plainer words: of the market and the school. Within these new coordinates, the canon remains as visible as ever, *but it loses its conceptual autonomy*, becoming the contingent outcome of the encounter between opposite forces. It is *these forces*, then, that deserve to be further investigated, if one wants to know more about the canon;[21] and future research might easily add print runs and presence in the circulating libraries to the popularity metrics, and excerpts from textbooks, or mentions in the non-fiction archive, to the prestige ones.[22] With each new addition, we will acquire a better sense of the composite nature of the canon – and of its *historical* nature, too: the canon of 1770-1830 (and, we suspect, of the following 70-80 years) was the product of the happy age of the European bourgeoisie, when the imperatives of success and education could be seen as compatible with each other, as was appropriate for a ruling class which, for the first time in history, felt at home in the market as well as the school. To have made the dual nature of the nineteenth-century canon intuitively "visible" – such is the achievement of these initial sections.[23]

II. Morphological Features

4. Measuring redundancy

Though different from Bourdieu's in many respects, the charts presented in the previous section shared his main methodological premise: they had a social rather than a literary foundation.[24] To make **Figure 3.3**, you don't need to open a single novel. As literary historians, however, we *wanted* to open the novels, and find out whether their social destiny – popular, prestigious, both, neither… – had any connection to their morphological features. So, while working at the diagrams of the literary field, we were also focusing on the internal composition of Chadwyck-Healey and of the sample from the larger archive. Here, the first step consisted in measuring the amount of redundancy and information present in the corpus. That readers prefer informative texts to redundant ones – thus keeping the former in print, while dooming the latter to extinction – is a widespread received idea, and we wanted to test it. Taking a cue from information theory, Mark Algee-Hewitt measured what is called "second order redundancy" (predictability at the level of individual words), using a modification of Shannon's measure of information load which determines the information content of each text by assessing how predictable each word-to-word transition is, given the range of possible transitions. Since "of" is much more often followed by "the" than by "no", for instance, the word

17 In terms of shrinking popularity, Austen and her contemporaries would provide a perfect case study: as **Figure 3.2** shows, about 25 authors (one third of them from the eighteenth century) were more popular than Austen in the sixty years covered by the diagram. As nineteenth-century novelistic bibliographies become more reliable, we will know how many of them were still more popular than her a generation or two later (initial results from the 1830s and 1840s suggest: Scott, and no one else).

18 Scott's immediate fame *and* acclaim are different from Austen's significantly slower pace, or from the ambiguous status of authors long confined to specific niches because of their initial audience (Carroll) or genre (Radcliffe, Doyle). And then, of course, there is the nemesis of any general theory of the canon – *Moby-Dick*.

19 Although our findings are completely different from Bourdieu's idea of the French literary field, they don't necessarily falsify his thesis, as we are working only on novels (to the exclusion of poetry, drama, magazines, and so on), and on a different country and period. Truth be told, we need many empirical maps of literary fields (plural), from different cultures and epochs, for the "literary field" (singular) to become a solid historical concept.

20 In **Figures 3.2–3.3**, which have as their cut-off point two or three standard deviations above the mean of the field, all authors in the high prestige and in the middle area, and about half of those in the high popularity area, can be considered canonical. As one descends "lower", the field's tri-partition remains visible a little longer, then disappears. What happens *then* it's a fascinating question – for another study.

21 Or more precisely: *if one wants to de-compose the concept of the canon into the two underlying elements of popularity and prestige*. Here, it's worth comparing the initial epistemological choice of this project with that of Algee-Hewitt's and McGurl's "Between Canon and Corpus". The main difference is not that between texts ("Between Canon and Corpus") and authors ("Canon/Archive") – which could be easily ironed out – but between an analysis based on networks, and one based on a Cartesian diagrams. Networks are much better *at investigating the relationships among individual nodes* (the hyper-canonical cluster identified in Figure 3 of the study, the singular centrality of *Grapes of Wrath*, the disconnect between bestsellers and the other groups), but *cannot connect the nodes to anything outside the network itself*. Cartesian diagrams, for their part, *embed the "outside" into their very axes* (like here popularity and prestige), but inevitably *loosen the relationships among individual data points* (in a diagram, there is no equivalent to network edges and clustering measures). Clearly, this is not a case of one strategy being "better" than the other, but of research projects that aim at investigating different properties of the system, and choose their means of analysis accordingly.

22 Needless to add, some of these measurements may be discontinuous and hard to come by (like print runs), while others (like textbooks) may start at a significantly later date. But if the notion of the literary field must help us understand different epochs and countries, having recourse to disparate historical indexes will be inevitable; rather than hoping for a – chimerical – homogeneity of the sources, we should learn to make heterogeneous data conceptually comparable.

23 "Between Canon and Corpus" shows how much things have changed since then: in the twentieth century, canon(s) are all characterized by a "systematic differentiation, if not contradiction, between artistic and commercial value". It is precisely this differentiation/contradiction that is absent from the "canonical" region of **Figure 3.3**.

24 "I propose that the problem of what is called canon formation", writes John Guillory, in a similar vein, "is best understood as a problem in the constitution and distribution of cultural capital, or more specifically, a problem of access to the means of literary production and consumption." John Guillory, *Cultural Capital: The Problem of Literary Canon Formation*, Chicago 1995, p. ix.

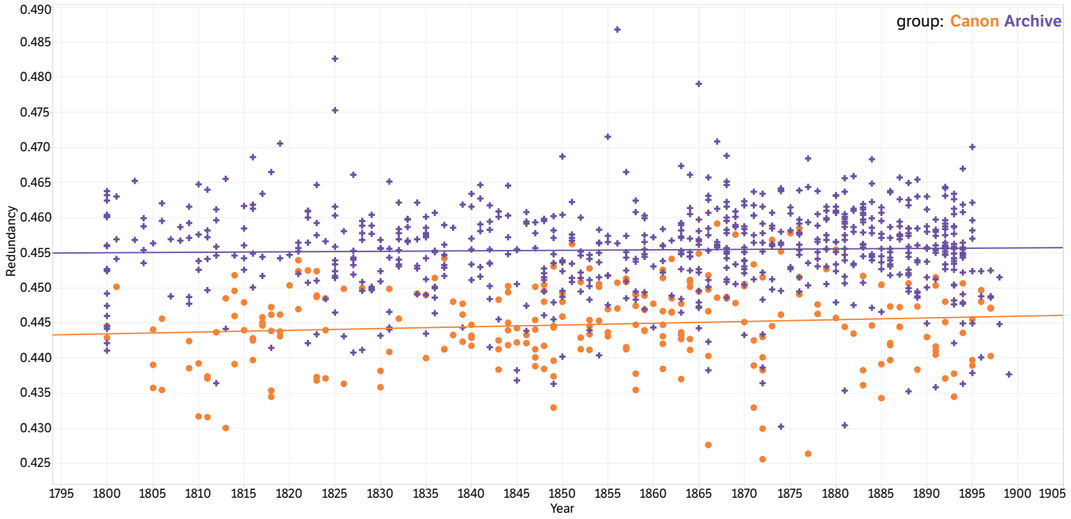


Figure 4.1. Measuring redundancy, 1800-1900

Purple crosses indicate archival novels, orange circles canonical ones

pair "of no" is far less predictable – hence more informative – than the bigram "of the".[25] **Figures 4.1** and **4.2** summarize Algee-Hewitt's investigation.

Figure 4.2 was particularly striking: that three-fourths of the Chadwyck-Healey collection would be less redundant than three-fourths of the archive was a *much* stronger separation than we had expected to find. And yet, we weren't completely happy. The clarity of the contrast had simply confirmed a received idea: forgotten authors used language in a redundant fashion; if they had remained unread, it was because they weren't really *worth* reading. And vice-versa: we still enjoy reading Austen because she is a paragon of information, as the close-up of **Figure 4.3** makes perfectly clear.

Not exciting, corroborating a received idea.[26] And then, there was a second problem. Though Algee-Hewitt had operationalized the concept of redundancy, and produced striking quantitative findings, it wasn't clear how we could dis-aggregate the overall score and look *at* the results, determining which specific word pairs returned all the time – or never did so. We had successfully measured redundancy, but couldn't really *analyze* it: an unsettling

25 Throughout this pamphlet, we will use "redundancy" and "repetition" almost interchangeably, placing them in antithesis to "information" and "variety"; though this is a simplification, we don't think it affects the level at which we are working, nor the type of results we have found. On a similar note, the relationship between information and redundancy is often referred to as "entropy"; we have opted for different definitions in order to make the various aspects of this research as comparable as possible.

26 And it was already the second time: in **Figure 1.2**, the fact that the canon regularly preceded the archive by 15-20 years seemed to "prove" that other received idea according to which great writers open the way, and the rest follow.

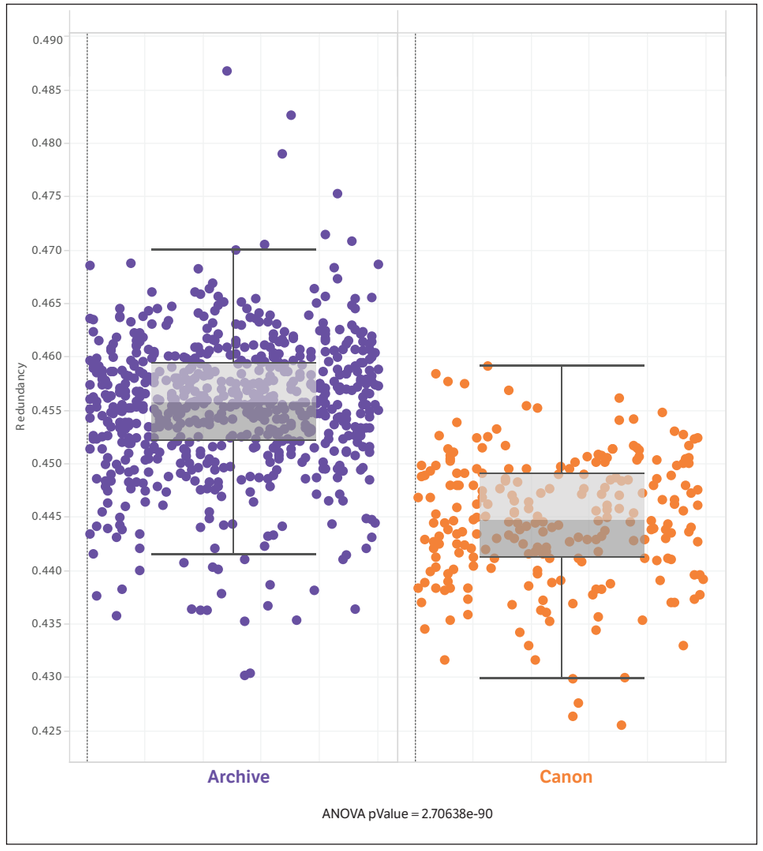


Figure 4.2. Redundancy in the nineteenth century: a synthetic diagram

This figure aggregates the data of **Figure 4.1** into the two sub-corpora of canon and archive. Each "box" includes the two central quartiles of the group, separated by a line which indicates the group's median value; the "whiskers" emerging from the box represent the two extreme quartiles, while outliers are indicated by individual dots.

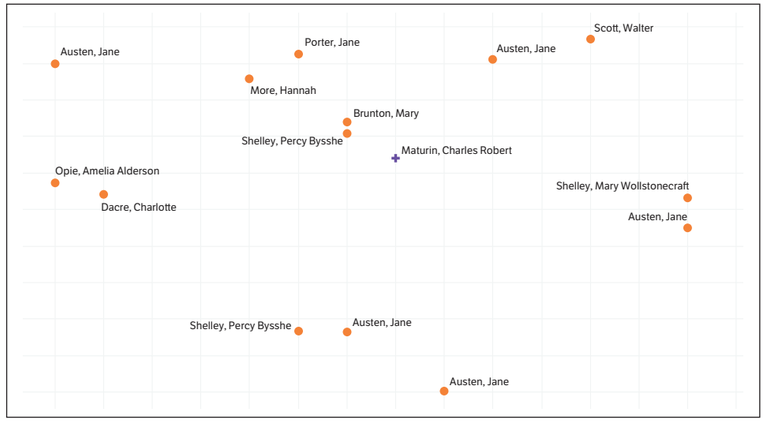


Figure 4.3. Very low redundancy in the early nineteenth century

A novel that never repeated a single word would have zero redundancy and 100% information – but this "information" would have no value, because it would rapidly become incomprehensible. Meaning always depends on a mix of repetition and novelty: that's why the scores in these figures oscillate in a rather narrow range. Differences *within* this range are however both consistent and significant, as is illustrated by this enlargement of the bottom left area of **Figure 4.1.**

departure from that interplay of quantitative measurement and qualitative interpretation which had been a constant of our work since the beginning. Here, statistical significance seemed impervious to critical meaningfulness: the "text" created by extracting the 100 most frequent bigrams from each novel in the corpus was a spreadsheet with over 100,000 cells: "reading" them was out of the question (**Figure 4.4**). A more technical approach – following the decay curve of the most frequent constructions – turned out to be equally inconclusive: very frequent bigrams ("there is", "I am", "to the") had very similar frequencies in all the texts, and variation occurred only in minute traces far down the curve. Plus, there were *so many* bigrams, in each novel, that their effects manifested themselves through an immense number of extremely small changes: in a relatively short text of 66,500 words, for instance, there were 66,499 bigrams, about 40,000 of which never repeated themselves. And whereas the number of shared words between two texts was substantial – at least 3-4,000 – the shared bigrams were usually less than 1000; too few for a solid comparative analysis.

We seemed to have created for ourselves a home-grown version of the uncertainty principle: the more precisely we measured redundancy, the harder it became to determine "where" it actually was. Redundancy operated at a scale that was all-pervasive, and apparently decisive in shaping the destiny of books; but the whole process took place so far below the level of conscious reading as to be practically invisible. In the future, perhaps even the near future, such a problem might be addressed by experimental psychology; in the meantime, we turned to a standard linguistic measure of lexical variety known as type-token ratio.[27] The lower a text's redundancy, we reasoned, the higher must its variety be: convex to concave. We would get an image that would be the exact reverse of **Figure 4.2**. So we did our calculations, and the result was **Figure 4.5.**

Placing **Figures 4.2** and **4.5** next to each other produced the following paradox: the canon was far less repetitive than the archive (hence much more

f_the_1441	in_the_672	to_the_634	of_his_341	of_a_333	a
f_the_1148	in_the_578	to_the_521	of_his_309	of_a_308	t
f_the_266	in_the_99	to_the_95	on_the_86	of_his_69	b
f_the_942	in_the_404	to_the_365	of_his_245	to_be_197	a
f_the_1486	in_the_781	to_the_633	of_his_365	of_a_364	a
f_the_746	to_be_616	in_the_574	it_was_389	she_had_365	c
f_the_679	in_the_401	sir_ulick_348	to_the_330	to_be_298	h
f_the_702	in_the_494	of_her_440	to_the_404	to_be_387	I
f_the_359	said_i_236	in_the_212	to_the_191	i_am_181	c
f_the_389	in_the_295	to_be_271	of_her_194	i_am_183	t
f_the_459	to_be_428	in_the_382	i_am_297	of_her_264	t
f_the_226	in_the_143	mr_glowry_84	and_the_71	of_a_69	t
f_the_161	in_the_91	to_be_85	to_the_52	i_am_40	c
f_the_342	to_the_246	the_marquis_215	in_the_214	of_his_212	a
f_the_245	to_the_196	in_the_184	mrs_villars_126	of_her_115	s
f_the_648	in_the_419	to_the_338	i_have_280	it_is_262	i
f_the_607	in_the_471	to_the_332	to_be_247	of_her_225	h
f_the_603	to_the_321	in_the_290	of_his_234	to_be_208	h
f_the_425	to_the_304	in_the_289	said_i_264	of_a_210	a
f_the_383	in_the_194	on_the_153	and_the_135	to_the_125	c
f_the_1627	in_the_1325	of_her_1315	to_the_1286	to_her_975	c
f_the_1004	in_the_626	to_the_573	he_had_441	he_was_389	t
f_the_751	in_the_452	to_the_412	he_had_399	she_had_350	t
f_the_4794	in_the_3327	to_the_2510	to_be_2195	of_her_2146	s
f_the_1169	to_the_724	in_the_676	to_her_512	of_her_490	t
f_the_1681	to_the_703	in_the_653	said_the_525	and_the_403	c
f_the_1302	in_the_589	to_the_586	of_his_369	and_the_318	c
f_the_1632	in_the_688	to_the_649	the_earl_408	of_his_385	a
f_the_1465	to_the_555	in_the_525	said_the_379	of_his_358	s

Figure 4.4. Reading bigrams: 0.00003% of the data

A section of the spreadsheet used for the calculations behind **Figures 4.1–4.2**. Though the bigrams themselves are perfectly identified, it's nearly impossible to "interpret" what they mean other than in statistical fashion. In this respect, Walser and Algee-Hewitt observed, bigrams were comparable to Braudel's "demographic progressions" and "variations in interest rates": all phenomena that could not be perceived at the passage-by-passage level on which we typically conduct our readings.

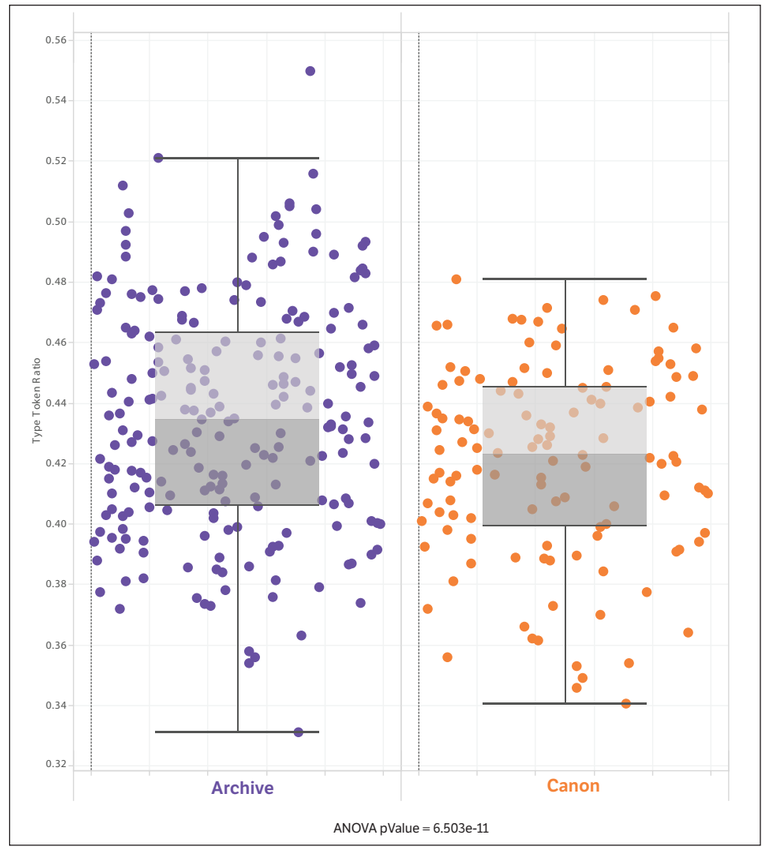


Figure 4.5. Measuring variety: a synthetic diagram of type-token ratio

Though the distinction between the two sub-corpora is here much less sharp than in **Figure 4.2**, the result is actually more dramatic: 4.2 had fully confirmed our expectations about canon and archive, whereas this chart completely contradicted them: the lexicon of the canon was not more variedthan that of the archive, but significantly less so. (The procedure followed to determine type-token ratio is described in footnote 28, at the beginning of the next section).

varied) from the perspective of word pairs, and at the scale of the entire text; and less varied (hence more repetitive) from the perspective of single words, and at the scale of a thousand. In itself, the fact that different textual scales would behave differently was not a surprise: two previous pamphlets ("Style at the Scale of the Sentence" and "On Paragraphs") had focused exactly on that question. But in those cases, different scales had been associated *with completely different features*: sentences with style, paragraphs with themes, and so on. Here, the features measured were very closely related. How could results reverse themselves from two words to a thousand? And we mean that "how" literally, not as a cry of despair: concretely, what textual mechanism could transform the first result into the second?

Algee-Hewitt addressed the question by "translating" all words into parts-of-speech, thus re-formulating redundancy via *categories* of bigrams rather than individual units; "clever little" and "first cruel", for instance, both became "adjective-adjective"; "a condition" and "the kitchen" became "determiner-noun", etc. Re-calculating everything in terms of "grammatical redundancy"

27 This is how the *Longman Grammar of Written and Spoken English* defines type-token ratio: "The relationship between the number of different word forms, or *types*, and the number of running words, or *tokens*, is called the *type-token ratio (or TTR)*. As a percentage, type-token ratio is equal to (types/tokens) x 100." See Biber, Johansson, Leech, Conrad, Finegan, *Longman Grammar of Spoken and Written English*, Harlow 1999, pp. 52-3.
The *Longman Grammar* follows the variations of type-token ratio across four registers (Conversation, Academic prose, Fiction, and News), and three sample lengths (100, 1,000, and 10,000 words). For 100-word segments the results are as follows: Conversation 63; Academic prose 70; Fiction 73; News 75. For 1,000-word segments: Conversation 30; Academic prose 40; Fiction 46; News 50. And for 10,000-word segments: Conversation 13; Academic prose 19; Fiction 22; News 28. Notice how the difference between the registers increases dramatically with the length of the segment: at 10,000 words, the type-token ratio of News is more than double that of Conversation, whereas it was only 16% higher at 100 words. We opted for 1,000 words segments, which seemed to be long enough to capture a good amount of variety, and short enough to allow direct analysis.

made it possible to identify which kinds of bigrams were most distinctive of the canon, and which of the archive (**Figures 4.6–4.7**).[28]

This time, the two sub-corpora revealed to have very different centers of gravity: the archive was dominated by nouns, while the canon had a very large presence of function words (conjunctions, determiners, prepositions). The archive's delight in titles (count Goldstein, uncle Gerard), punctiliousness about places and people (in Ireland; to Shirley), and liberality with proper nouns in general (Hector's lodgings, Shelburne upon) finally gave us a clue to its high redundancy: "count Goldstein" and "Shelburne upon" may not appear very often in a novel – but when they do, the two words are likely to re-occur together, increasing the text's redundancy; and the same for constructions like the adjunct nouns "iron will" and "autumn tints". It wasn't an answer to all our questions, but it was a beginning. And then, in order to address the other side of the paradox, we turned back to type-token ratio.

Preposition-proper noun (IN_NNP): to Shirley; in Ireland
Adjective-adjective (JJ_JJ): young happy; first cruel
Noun-adjective (NN_JJ): child incapable; nomenclature peculiar
Noun-noun (NN_NN): iron will; evening sky
Noun- proper noun (NN_NNP): count Goldstein; uncle Gerard
Noun-plural noun (NN_NNS): iron bars; autumn tints
Proper noun-preposition (NNP_IN): Alps of; Shelburne upon
Proper noun-noun (NNP_NN): Agnes' wedding; Manchester cotton
Proper noun-plural noun (NNP_NNS): Cumberland coasts, Hector's lodgings
Noun-pronoun (NN_PRP): tail itself, driver himself.

Figure 4.6 Most distinctive grammatical bigrams: archive

Conjunction-gerund (CC_VBG): and walking; and taking
Determiner-adjective (DT_JJ): the silly; an eventful
Determiner-noun (DT_NN): a condition; the kitchen
Determiner-plural noun (DT_NNS): the environs; the travelers
Preposition-determiner (IN_DT): at the; in a
Adjective-plural noun (JJ_NNS): folded arms; harsh features
Noun-preposition (NN_IN): account of; sense of
Plural noun-preposition (NNS_IN): grains of; years of
Possessive pronoun-plural noun (PRP$_NNS): their excursions; our girls

Figure 4.7 Most distinctive grammatical bigrams: canon

28 For this part of the work, Algee-Hewitt used the Stanford Parts-of-Speech Tagger; the abbreviations enclosed in parentheses (IN_NNP etc.) are however those used by the Treebank project (https://www.cis.upenn.edu/~treebank/) of the University of Pennsylvania.

5. "But I couldn't go away"

In the case of type-token ratio, the first thing that needed to be done was to come up with a mode of analysis appropriate to a corpus where most novels had not been reprinted for a century or two, making optical recognition difficult, and hence potentially invalidating all subsequent calculations. Ryan Heuser, who had first directed our attention to type-token ratio in the early phases of the project, found a way to measure it equally reliably across texts of very different quality.[29] Once the results were in, we started by looking at low type-token ratio, to see how its specific kind of repetitiveness compared to the redundancy calculated by Algee-Hewitt. We knew from **Figure 4.6** that low lexical variety would often correlate with canonical texts, and indeed the frequency of the Chadwyck-Healey collection, which amounted to around 20% of the corpus overall, rose to 50% among the 500 segments with the lowest type-token ratio (whereas it was a mere 3.2 in the top 500). Among the 50 texts with the lowest scores, about half were from Chadwyck-Healey: several children books (*Alice, Through the Looking-Glass, The Water Babies, Black Beauty, Little Lord Fauntleroy, Island's Night's Entertainments…*), ten of Trollope's novels (*The Last Chronicle*

29 Heuser began by creating a very large dictionary of novelistic English – 232,845 distinct words – and slicing all texts into segments of 1,000 "dictionary-words". (Actual segments would be anywhere from 1,000 to ~1,500 words long, depending on how many "non-dictionary" words – OCR errors, *hapax legomena*, etc – they had.) Since the number of tokens was fixed at 1,000, dividing the number of types in each segment by 1,000 produced segment-based scores whose average gave us the type-token ratio for the text.
The function was written with two parameters: "slice_len" [the length of the segment (set at 1000)] and "force_english" [whether to include words not in a very large English dictionary (set at False)]. The reasoning behind the "force English" parameter, which excluded all non-"English" words, was that, without it, the archive would have a higher type-token ratio simply by virtue of its bad OCR. Conversely, the concern *with* forcing English was that the same bad OCR would produce a *lower* type-token ratio: if the segment had to expand over ~1,500 "real" words in order to find 1,000 "English" ones, then it might privilege shorter, easier-to-spell-and-OCR words, which are also the most frequent in the language, thus driving type-token ratio downwards. In the event, these two undesirable outcomes seemed to balance each other out.

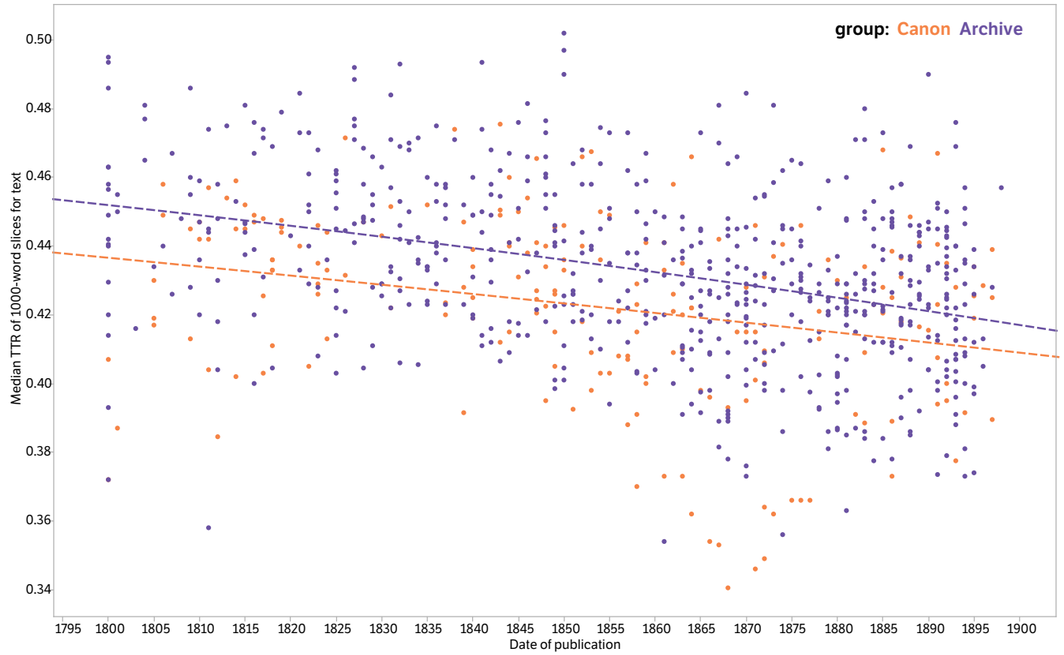


Figure 5.1 Type-token ratio, 1800-1900

The "pull" of children's stories towards a low type-token ratio is visible between 1860 and 1880; in general, though, the type-token ratios of both canon and archive remain rather stable across the nineteenth century.

of Barset, *Phineas Finn the Irish Member, Can You Forgive Her?*, *The Eustace Diamonds…*), plus two Irish novels (Edgeworth's *Castle Rackrent* and Samuel Ferguson's *Father Tom and the Pope*, with *The Absentee* not very far). In itself, this mix was not particularly representative of the canon (whatever one may mean with that term); more significant seemed to be the fact that Chadwyck-Healey's scores remained low across the century (**Figure 5.1**), and that the trend involved some of the greatest nineteenth-century stylists: all of Austen was below the corpus mean (with *Persuasion, Sense and Sensibility,* and *Mansfield Park* in the bottom 20%); all of Dickens was below the mean (with *Little Dorrit, A Tale of two Cities, David Copperfield, Our Mutual Friend, Bleak House,* and *Great Expectations* in the bottom 20%); all of George Eliot was

below the mean – and *Adam Bede* contained the passage with the lowest type-token ratio of the entire century.

Now, *Adam Bede* is a strange novel for that kind of result, because it contains Eliot's famous reflections on Dutch painting: a manifesto for aesthetic precision and variety, *written* with extraordinary precision and variety (**Figure 5.2**).

The first 100 words of this passage have a type-token ratio of 79: higher than anything, in any register, discussed by the *Longman Grammar*. And yet, later in the novel, Eliot's style runs to the opposite extreme (**Figure 5.3**).

Eliot's passage includes the central moment of Hetty's confession to Dinah: the recollection of having abandoned her child in the woods, and of waiting for "its" death (to use the pronoun she herself uses). But "waiting" is the wrong word (**Figure 5.4**).

Grammatically, the most arresting feature of these sentences is the flood of inflected verb forms with Hetty as their subject: I made haste … I could hear … I got out … I was held fast … I couldn't go away … I wanted … I sat … I was … I had … I couldn't … In narrative analysis, verb forms are usually seen as indices of "action" – and comprehensibly so. But here, in a grating dissonance between grammar and semantics, they stand for paralysis instead: Hetty desperately wants to "go away" – and can't. And just as she cannot leave the physical setting of the episode, she cannot relinquish the *words* which describe it. She cannot *forget*: that's where the repetition comes from. Better: she can neither forget, *nor really say what has happened*. In a textbook instance of the opposition between "repeating" and "working through", she keeps saying the same things over and over again, because she cannot bring herself to utter the one thing that really matters: the word "death" is *never* repeated, and only appears in an oblique, misleading construction at the end of the passage.[30]

Why repetition? Because a trauma has occurred, and repetition is a great way to express it in language: an imprisonment in one's own words whose enigmatic force explains why Eliot, despite her love for analytical details, could write the most repetitive passage of the entire century. And then, Hetty's confession also brings to light the fundamentally *oral* component of type-token ratio. Next to Eliot's page, the two segments with the lowest lexical density are also confessions: of baby-changing in Edgeworth's *Ennui*,[31] and of love in Trollope's *Last Chronicle of Barset*.[32] In the same low range we find passages

30 "But it was morning, for it kept getting lighter, and I turned back the way I'd come. I couldn't help it, Dinah; it was the baby's crying made me go--and yet I was frightened to death. I thought that man in the smock-frock 'ud see me and know I put the baby there." Notice how "death" is referred to Hetty instead of her child.

31 "I thought, how happy he would be if he had such a fine babby as you; dear; and you was a fine babby to be sure; and then I thought, how happy it would be for you, if you was in the place of the little lord: and then it came into my head, just like a shot, where would be the harm to change you?"

32 "You are so good and so true, and so excellent,-- such a dear, dear, dear friend,

> It is for this rare, precious quality of truthfulness that I delight in many Dutch paintings, which lofty-minded people despise. I find a source of delicious sympathy in these faithful pictures of a monotonous homely existence, which has been the fate of so many more among my fellow-mortals than a life of pomp or of absolute indigence, of tragic suffering or of world-stirring actions. I turn, without shrinking, from cloud-borne angels, from prophets, sibyls, and heroic warriors, to an old woman bending over her flower-pot, or eating her solitary dinner, while the noonday light, softened perhaps by a screen of leaves, falls on her mob-cap, and just touches the rim of her spinning-wheel, and her stone jug, and all those cheap common things which are the precious necessaries of life to her—or I turn to that village wedding, kept between four brown walls, where an awkward bridegroom opens the dance with a high-shouldered, broad-faced bride, while elderly and middle-aged friends look on, with very irregular noses and lips, and probably with quart-pots in their hands, but with an expression of unmistakable contentment and goodwill. "Foh!" says my idealistic friend, "what vulgar details!"

Figure 5.2 "This rare, precious quality of truthfulness"

> came all of a sudden, as I was lying in the bed, and it got stronger and# stronger#… I# longed so to go back again… I# could n't* bear being so# lonely, and# coming to# beg for want. And# it# gave me strength and# resolution to# get up and# dress myself. I# felt I# must do it#… I# did n't* know how… I# thought I#'d find a# pool, if I# could#, like that other, in# the# corner of# the# field, in# the# dark. And# when the# woman went out, I# felt# as# if# I# was# strong enough to# do# anything… I# thought# I# should get# rid of# all# my misery, and# go# back# home, and# never let'em know# why I# ran away. I# put on my# bonnet and# shawl, and# went# out# into the# dark# street, with the# baby under my# cloak; and# I# walked fast till I# got# into# a# street# a# good way off, and# there was# a# public, and# I# got# some warm stuff to# drink and# some# bread. And# I# walked# on# and# on#, and# I# hardly felt# the# ground I# trod on#; and# it# got# lighter, for# there# came# the# moon-- O, Dinah, it# frightened me# when# it# first looked at me# out# o#' the# clouds-- it# never# looked# so# before; and# I# turned out# of# the# road into# the# fields, for# I# was# afraid o#' meeting anybody with# the# moon# shining on# me#. And# I# came# to# a# haystack, where I# thought# I# could# lie down and# keep myself# warm# all# night. There# was# a# place cut into# it#, where# I# could# make me# a# bed#; and# I# lay comfortable, and# the# baby# was# warm# against me#; and# I# must# have gone to# sleep for# a# good# while, for# when# I# woke it# was# morning, but not very light, and# the# baby# was# crying. And# I# saw a# wood a# little way# off#… I# thought# there#'d perhaps be a# ditch or a# pond there#… and# it# was# so# early I# thought# I# could# hide the# child there#, and# get# a# long way# off# before# folks was# up#. And# then I# thought# I#'d go# home#-- I#'d get# rides in# carts and# go# home#, and# tell'em I#'d been to# try and# see for# a# place#, and# could# n't* get# one. I# longed# so# for# it#, Dinah#-- I# longed# so# to# be# safe at# home#. I# do# n't* know# how# I# felt# about the# baby#. I# seemed to# hate it#-- it# was# like# a# heavy weight hanging round my# neck; and# yet its crying# went# through me#, and# I# dared n't* look at# its# little# hands and# face. But# I# went# on# to# the# wood#, and# I# walked# about#, but# there# was# no water"… Hetty shuddered. She was# silent for# some# moments, and# when# she# began again#, it# was# in# a# whisper.`` I# came# to# a# place# where# there# was# lots of# chips and# turf, and# I# sat down# on# the# trunk of# a# tree to# think what I# should# do#. And# all# of# a# sudden# I# saw# a# hole under# the# nut-tree*, like# a# little# grave. And# it# darted into# me# like# lightning-- I#'d lay# the# baby# there#, and# cover it# with# the# grass and# the# chips#. I# could# n't* kill it# any other# way#. And# I#'d done it# in# a# minute; and#, O#, it# cried so#, Dinah#-- I# could# n't* cover# it# quite up#-- I# thought# perhaps# somebody `ud* come and# take care of# it#, and# then# it# would n't* die. And# I# made haste out# of# the# wood#, but# I# could# hear it# crying# all# the# while#; and# when# I# got# out# into# the# fields#, it# was# as# if# I# was# held fast#-- I# could# n't* go# away#, for# all# I# wanted so# to# go#. And# I# sat# against# the# haystack# to# watch if# anybody# `ud* come#: I# was# very# hungry, and# I#'d only a# bit of# bread# left; but# I# could# n't* go# away#. And# after ever such a# while#-- hours and# hours#-- the# man came#-- him in# a# smock-frock*, and# he looked# at# me# so#, I# was# frightened#, and# I# made# haste# and# went# on#. I# thought# he# was# going to# the# wood#, and# would# perhaps# find# the# baby#. And# I# went# right on#, till# I# came# to# a# village, a# long# way# off# from the# wood#; and# I# was# very# sick, and# faint, and# hungry#. I# got# something to# eat there#, and# bought a# loaf. But# I# was# frightened# to# stay. I# heard the# baby# crying#, and# thought# the# other# folks# heard# it# too,-- and# I# went# on#. But# I# was# so# tried, and# it# was# getting towards dark#. And# at# last, by the# roadside there# was# a# barn-- ever# such# a# way# off# any# house-- like# the# barn# in# Abbot's Close; and# I# thought# I# could# go# in# there# and# hide# myself# among the# hay and# straw, and# nobody `ud* be# likely to# come#. I# went# in#, and# it# was# half full o#' trusses of# straw#, and# there# was# some# hay#, too#. And# I# made# myself# a# bed#, ever# so# far behind, where# nobody# could# find# me#; and# I# was# so# tired and# weak, I# went# to# sleep#…. But# oh, the# baby#'s crying# kept waking me#; and# I# thought# that# man# as# looked# at# me# so# was# come# and# laying hold of# me#. But# I# must# have# slept a# long#

Figure 5.3. The nineteenth-century's most repetitive passage: Hetty's confession in Adam Bede

The pound sign indicates that a word is being repeated within the given segment, while asterisks denote words that are not part of the "dictionary" used for the calculations. Some odd aspects of this and other passages are artifacts of the Stanford parser – which, for instance, considers negative contractions, such as "n't" at the end of "couldn't", as a separate word.

> And# I# made haste out# of# the# wood#, but# I# could# hear it# crying# all# the# while#; and# when# I# got# out# into# the# fields#, it# was# as# if# I# was# held fast#-- I# could# n't* go# away#, for# all# I# wanted so# to# go#. And# I# sat# against# the# haystack# to# watch if# anybody# `ud* come#: I# was# very# hungry, and# I#'d only a# bit of# bread# left; but# I# could# n't* go# away#.

Figure 5.4. "But I could hear it crying all the while"

from children stories (with their typically life-like narrators), Irish novels (which specialized in the imitation of speech), and countless instances of Trollope's petty-bourgeois stichomythia.[33] There are trial scenes (*The Ordeal of Richard Feverel, The Heart of Mid-Lothian*, William Scargill's *Tales of a Briefless Barrister*), ideological confrontations (*Marius the Epicurean*), an ecstatic vision of the "communism of happiness" (Mary Christie's *Lady Laura*),[34] and a great invective against money (Thomas Pemberton's *A Very Old Question*).[35] There are characters who talk too much because they are trying to be obliging (*Emma*), or because, like Van Helsing in *Dracula*, they need to rehearse the evidence over and over again. It could hardly be an accident, concluded Allison and Gemma, that our lowest-ranked (and largely canonical) 1,000-word segments were in *exactly* the same range as conversation in the *Longman Grammar*: a mean of 30 in their case, and a range of 27-33 for our bottom 500 segments.

We had turned to type-token ratio in the hope that it would lead us back to some kind of textual analysis – and we had not been disappointed: low scores captured crucial aspects of narrative structure, signaling trauma, intensity, and orality. And high scores?

6. "Embrasures bristling with wide-mouthed cannon"

Figure 6.1 shows the ten novels with the highest type-token ratio in the corpus; **6.2** the top-scoring passage, from Edward Hawker's *Arthur Montague, or, An Only Son at Sea*.

If the privileged social position of the canon were always correlated with linguistic privilege – Dario Fo, 1997 Nobel prize for literature, once wrote a play entitled *The worker knows 300 words, the boss 1,000; that's why he's the boss* – then canonical authors should have a much more varied language than forgotten ones. In terms of the type of lexical abundance measured by type-token ratio, however, the opposite is true. "The whole language of aesthetics is contained in a fundamental refusal of the *facile*", writes Bourdieu: "'vulgar' works [...] arouse distaste and disgust by the methods of seduction".[36] Facile, Hawker's language? Seductive? If anything, the opposite. A dichotomy such as vulgar/refined will never explain the connection between the archive and high type-token ratio. We must look elsewhere.

As often in this research, we found an answer in corpus linguistics. This time, it was the concept of "register": the "communicative purposes and situational contexts" of messages described by Douglas Biber and Susan Conrad in *Register, Genre, and Style*.[37] In the study of register, the fundamental opposition runs between oral and written, and it is a well-established fact that the latter has in English a much higher type-token ratio than the former. If the archive has a greater lexical variety than the canon, then, the reason is that *the archive inclines towards the "written" register much more than the canon* (while the latter, as we have seen in the previous section, is much more at ease with "oral" conventions). It's not that archival novels with high type-token ratio have fewer oral passages (dialogue, speech, exclamations, etc.); Gemma's work in progress on colloquial discourse suggests that they may even have more; it's that their "spoken" passages have a markedly "written" quality. Jane West's *Ringrove*, for instance, includes a lot of language typographically marked as "speech" – which however consists often of formal tirades that sound closer to a written disquisition than to an oral exchange. [38]

Linguistic conservatism is certainly one reason for the "written" quality of many archival works. A passage from William North's *The Impostor* – whose type-token ratio is near the top 1% of the corpus – expresses it well:

> There has of late years crept into our *belles lettres*, in addition to the *soi-disant* fashionable trash above mentioned, a violent predilection for low life, slang, and vulgarism of every kind. Dickens and Ainsworth led the way, and whole hosts became their followers ... Let us endeavor to reestablish pure classical taste.

Let us endeavor to reestablish ... In their study of prestige and style, Underwood and Sellers have found that many obscure books "at the very bottom

Edward Duros, *Otterbourne; A Story of the English Marches,* 1832
Edward Hawker, *Arthur Montague, or, An Only Son at sea,* 1850
Emma Robinson, *The Armourer's Daughter: or, The Border Riders,* 1850
William Lennox, *Compton Audley; or, Hands Not Hearts,* 1841
Mary Anne Cursham, *Norman Abbey: A Tale of Sherwood Forest,* 1832
William Maginn, *Whitehall; or, The Days of George IV,* 1827
Thomas Surr, *The Mask of Fashion; A Plain Tale, with Anecdotes Foreign and Domestic,* 1807
James Grant, *The Scottish Cavalier: An Historical Romance,* 1850
Cecil Clarke, *Love's Loyalty,* 1890
Jane West, *Ringrove, or Old Fashioned Notions,* 1827

Figure 6.1 High type-token ratio, or, the triumph of the archive

then cut through some acres of refreshing greensward, studded with the oak, walnut, and hawthorn, ascended a knoll, skirted an expansive sheet of# water; afterwards entering an# avenue of# noble elms, always tenanted* by a# countless host of# cawing* rooks, whose clamorous conclaves* interrupted the# stillness that reigned around, and# whose# visits to adjacent cornfields* of# inviting aspect raised the# ire and# outcry of# the# yelling VOL. I. C urchins employed to# guard them from depredation. Emerging from# this arched vista, a# near view was obtained of# the# mansion, approached through# a# thick luxuriant shrubbery of# full-grown* evergreens. It was# a# straggling stone structure of# considerable size and# doubtful architecture, having on either side an# ornamental wing, surmounted by# glazed cupolas*, and# indented below with# niches containing statues and# vases alternate. The# front face of# the# building displayed a# row of# fine Corinthian pillars-- their capitals screened by# wire-work* shields, to# defend them# from# the# injurious intrusions of# the# feathery tril*> e, who ever chirped* and# hovered about the# forbidden spots, coveting the# shelter denied them#. In the# vicinity of# the# house was# a# spacious flower-garden*, encompassed by# a# protecting plantation of# bay, holly, augustines*, arbutus, laburnum, yellow and# red Barbary, lilac, and# Guelder-rose*, ever# melodious with# the# shy, wary blackbird's whistle, the# sweet notes of# the# secreted thrush, and# the# varied carols of# their# fellow-choristers*, all conspiring to# give motion as well as# life to# their# leafy concealment. To# the# right, was# a# rich, park-like* prospect, sprinkled with# deer, grazing beneath clumps of# commingled oaks and# chestnuts or pulling acorns from# the# low, overhanging? branches of# some# solitary venerable stout-trinket* tree, whose# outspread limbs bent downwards to# the# earth from# whence their# life# was# drawn, as# if in# thankfulness for the# nourishment received. In# an# opposite direction stretched forth undulating woodland scenery, bordering on# an# open furzy down, which was# frequently occupied by# the# moveable* abodes* of# those houseless rovers-- the# hardy, spoliating*, mendacious tribe, whose# forefathers Selim*, on# (continued on page 13)

Figure 6.2 The nineteenth-century's least repetitive passage.

Arthur Hawker's landscape description has a type-token ratio of 60, well above the scores (46 for fiction and 50 for news) reported by the Longman Grammar for segments of equal length.

---

that I will tell you everything, so that you may read my heart. I will tell you as I tell mamma,-- you and her and no one else;-- for you are the choice friend of my heart. I can not be your wife because of the love I bear for another man".

33 "Do you think that I am in earnest?" "Yes, I think you are in earnest." "And do you believe that I love you with all my heart and all my strength and all my soul?" "Oh, John!" "But do you?" "I think you love me." "Think!"

34 "All are not equally happy; all can not be equally happy. But there is a sort of communism possible in happiness. The unhappy have a claim upon the happy; the happy have a debt towards the unhappy." "But how can one share one's happiness with others? It seems to me impossible. It is what I have most wished to do, but I see no way in which it can be done." "In one sense certainly you can not share your happiness, and you can not give it away. It is essentially your own, a development of your being, a part of yourself that you may not alienate."

35 "Money!" she cried derisively." Money! What is money to the trouble which has torn my heart ever since I have been married! What is money to those who thirst for love! I never wanted money; without money I was strong and happy; since I have had it I have been weak and miserable. Money broke down my poor father, and it was for money that Percy married, deceived, and has forsaken fine. Thank God that the wretched money has gone"

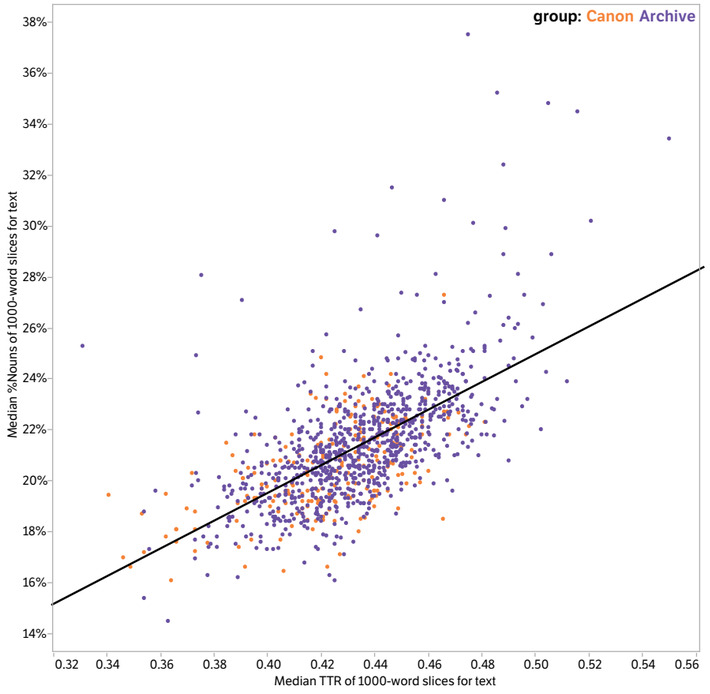
36 Pierre Bourdieu, *Distinction. A social critique of the judgment of taste*, 1979, Harvard UP, 1984, p. 486.

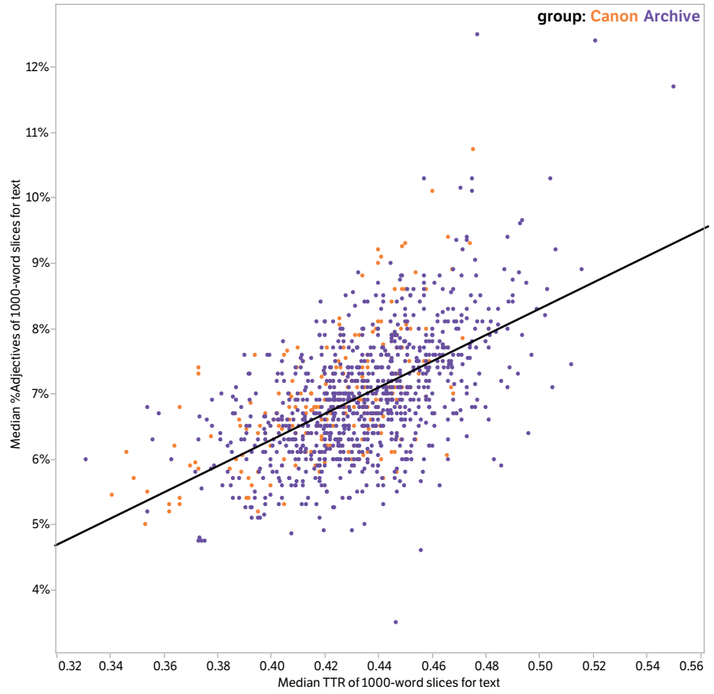
37 Douglas Biber and Susan Conrad, *Register, Genre, and Style*, Cambridge UP 2009, p. 2.

38 Here is one, on Byron's misuse of his poetic gifts: "There is a deep condensation of thought, an appropriateness of diction, an elegance of sentiment, and an original glow of poetical imagery; ever happy in illustrating objects, or deepening impressions;-- which so fascinate our fancy and bewilder our judgment, that we lose sight of the nature of the deeds he narrates, and the real character of the actors."

of [their model's] list [...] have some inspirational or hortatory purpose".[39] The same here: the "slang and vulgarisms" typical of oral registers offend "pure classical taste", and the cohort of **Figure 6.1** strike back, "elevating" the tone of discourse to the formal gravity of the written page: many nouns, many adjectives, and as few inflected verb forms as possible (**Figures 6.3-6.4-6.5**).[40]

So far, we have explained the affinity between high type-token ratio and the written register as the result of, loosely speaking, stylistic and ideological choices. But there is also a more neutral, "functional" reason for their correlation. In the findings of corpus linguistics, maximum lexical variety is consistently associated with news: a discourse which needs "an extremely high density of nominal elements", the *Longman Grammar* points out, in order to "refer to a diverse range of people, places, objects, events, etc." (53-54). There is a double source for lexical variety in news: the first is the necessary specificity internal to each distinct news item; the second, the utter discontinuity *between* one item and the next: as each article or correspondence begins, repetition is "reset" near zero, and type-token ratio can rise accordingly. This twofold logic returns in fictional texts with high type-token ratio: they include plenty of disparate materials, and further accentuate their diversity by using a plurality of generic forms. Jane West, six of whose novels are in the corpus' top 3% for type-token ratio, quotes poetry in 17 of her 24 top-ranked segments; in the absence of poetry, she turns to elaborate metaphors ("expect a fearful tempest to arise, which will clear the tree of its unsound branches"), and even pastiche.[41] William North's introduction to *The Impostor* – half literary criticism, half apologia – discusses a wide range of topics, and includes an excursus on...the wide range of topics he has decided to





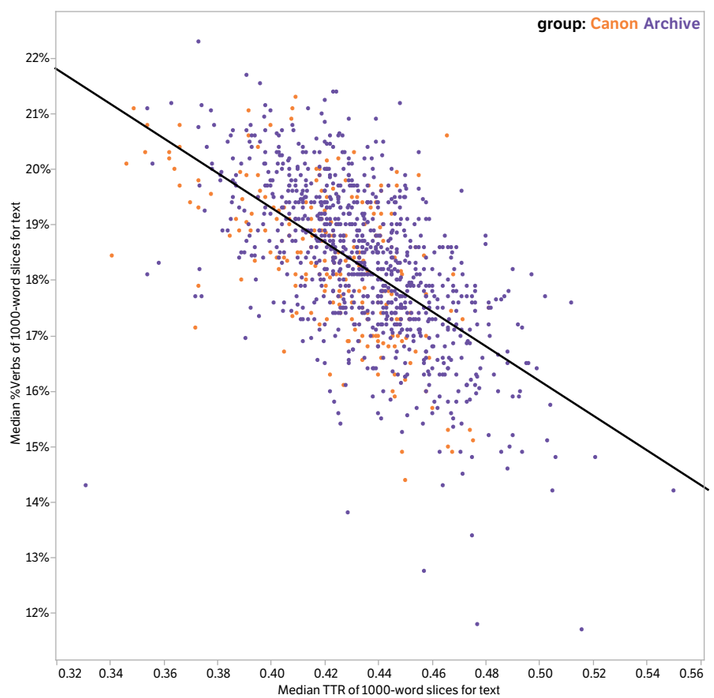


Figure 6.3 (top left)Type token ratio and nouns

Figure 6.4: (top right) Type token ratio and adjectives

Figure 6.5: (bottom right) Type token ratio and verbs

In Hawker's Gibraltar passage, in **Figure 6.2**, adjectives (and participles) are three times as frequent, and inflected verbs three-four times less frequent than the average in nineteenth-century fiction. By contrast, the Adam Bede passage in **Figure 5.4** contains only four nouns and one adjective – "hungry" – in 75 words.

39 Underwood and Sellers, p. 14.

40 The high frequency of nouns and adjectives takes us back to the "grammatical bigrams" discussed at the end of section 4: the "adjective-adjective", "proper noun-noun", "noun-adjective" word pairs. By combining those results with what has emerged in this section, we can finally solve the paradox of texts with high redundancy at the level of bigrams, and high variety at that of type-token ratio. The "labeling" function of bigrams like "count Goldstein" and "uncle Gerard", or the cliché-like loquacity of "iron will" and "clever little", can easily repeat themselves in the course of the novel, thus raising redundancy as measured *at that scale*; but even a mediocre writer is unlikely to repeat "clever little" within a 1,000-word window, thus leaving type-token ratio quite high. And the opposite will happen with the "determiner-noun" or "preposition-determiner" bigrams that are typical of canonical texts: as "the" is the most frequent word in English, it will inevitably repeat itself dozens of times in a 1,000-word segment, thus lowering its lexical variety; but since the noun next to the article can easily vary, redundancy at the level of bigrams will remain relatively low.

41 "First, Venus, queen of gentle devices! taught her prototype, lady Arabella, the use of feigned sighs, artificial tears, and Studied fainting: while Aesculapius descended from Olympus, and, assuming the form of a smart physician, stepped out of an elegant chariot, and on viewing the patient, after three sagacious nods, whispered to the trembling aunt, that the young lady's disorder, being purely mental, was beyond the power of the healing art. Reduced to the dire alternative of resigning the fair sufferer to a husband or to the grave, the relenting lady Madelina did not long hesitate." (Jane West, *A Tale of the Times,* 1799).

insert into his "romance."[42] Thomas Hope turns to political prophecy,[43] Lewis Wingfield to a half-parodic architectural digression,44 Edward Duros to erudite antiquarianism,[45] Edward Hawker to naturalistic instruction ...

But enough examples. It was time for some final reflections.

III. Large-Scale Dynamics in the Literary Field

It is not easy, "concluding" a project that had strayed so far from its original aim. We began with canon and archive as our objects of study, and with redundancy and type-token ratio as the means to investigate them; but then, the relationship between means and ends silently reversed itself: canon and archive moved to the periphery of our discussions, while redundancy and type-token ratio were increasingly occupying their center. There was nothing planned about this switch; for quite a while, we didn't even realize it had happened. But we *were* spending month after month wondering what bigrams actually "meant", and why on earth they managed to separate our texts as well as they did; later, once Allison and Gemma introduced the issue of oral and written registers, we spent even more time on type-token ratio, reading passages from unheard-of novels bristling with pound signs, asterisks, and words like "acclivities", "laburnum", and "commingling". Strange.

Why did we do that? Because we felt that working on type-token ratio would make us understand something about the "internal" forces – as distinct from the "external" ones discussed in section 3 – that shaped the literary field. It was another slippage in our object of study: the supposed line of demarcation between canon and archive – the diagonal slash still visible in our title – lost much of its interest, re-absorbed within a much larger landscape. With all due sense of proportion, there was a similarity with Bourdieu's trajectory of forty years earlier: when, starting from a study of *Sentimental Education*, and of Flaubert's position within nineteenth-century French literature, he developed a general framework where Flaubert was still present, but only as one element among many. The same here: canon and archive were still "in" the picture, with their differently colored markers; but now, the point of our diagrams consisted in throwing light on the literary field as a whole. A stylistic polarity exemplified by Eliot and Hawker no longer made us think of canon and archive, but of "oral" and "written" registers. The focus had shifted.

Still, a major difference persisted, between our work and Bourdieu's. For us, the sociology of the literary field *cannot rest on sociology alone*: it needs a strong morphological component. That's why redundancy and (especially) type-token ratio had become so important: their mix of the quantitative and the qualitative was perfect for the morpho-sociology of fiction that was our ultimate goal. Retrospectively, we must admit that the goal has remained out of reach – though it has moved a little closer. Out of reach, in the sense that, where the correlation between morphology and social fate was strongest – the case of redundancy – the elusive nature of the morphological unit of bigrams made a causal chain difficult to establish; whereas, by contrast, where the trait allowed for a rich and explicit analysis – the case of type-token ratio – the correlation was weaker, and became undisputable only for extreme cases. At the same time, two phenomena which had become visible near those extreme cases – the intensity of characters' voices near the lowest scores, and the topical miscellany of the narrator's prose at the opposite extreme – had opened a new line of inquiry, where the quantitative-qualitative continuum re-emerged very clearly, and led straight to two key concepts of Bakhtin's theory of the novel: polyphony, and heteroglossia (the "other languages" of consolidated extra-literary discourses, like politics, aesthetics, geography, architecture, etc.) Usually, these two notions are seen as closely related (and Bakhtin himself seemed to think so); but as Walser pointed out in our final round of discussions, our findings revealed that *they were actually localized in opposite regions of the novelistic field*: polyphony tendentially associated with canonical texts, and heteroglossia with forgotten novels. The proximity between heteroglossia and failure was especially arresting. For Bakhtin, when the novel comes into contact with other discourses, it creatively transforms them, appropriating their strength and reinforcing its own centrality within the cultural system. It's as if, with heteroglossia, nothing could ever go wrong. But that's exactly what happened with our small army of forgotten authors: the encounter with other discourses had a paralyzing effect, producing lifeless duplicates of non-fictional prose in lieu of dialogic vitality. As far as survival within the British literary system was concerned, it was a very bad choice.

Heteroglossia as a potential pathology of novelistic structure, then? "There is no fact which is [...] pathological in itself", writes Georges Canguilhem in his masterpiece on nineteenth-century conceptions of "normality": "an anomaly or a mutation is not in itself pathological, they just express other possible forms of life."[46] If this thesis is right, what doomed Hawker and North and Duros was less the choice of heteroglossia *in itself*, than the fact that it occurred in an age and country – in an ecosystem – *when the form of the novel was moving in the opposite direction*: tightening its internal narrative bolts, rather than looking for inspiration in external discourses (as was still happening in other countries). Even Dickens, for all his Parliamentarese, wrote novels with an outstanding measure of "orality". It was this specific historical conjuncture that made the "other languages" of heteroglossia bad for survival.

On this point, a longer historical view can be of help. Some time ago, the classicist Niklas Holzberg, wrote an essay whose key cognitive metaphor – "the Fringe" – has left a deep mark on the study of the ancient novel.[47] What Holzberg meant with his expression was that, around the extremely small cohort of Greek and Latin "novels proper", a much larger group of texts existed, where novelistic traits were mixed with elements from other discourses (historiography, travel reports, philosophy, political education, pornography...), thus expanding the scope of what the novel could do. In the twenty centuries that followed – as the novel "proper" increased its productivity, diversified its forms, and raised its status within the general culture – the role of the Fringe correspondingly contracted, and scholars of modern literature have hardly ever bothered with the idea. But in fact, the Fringe has never ceased to exist: the writers in **Figure 6.1** are its modern version, and their strange proliferation of topics is the typical sign of works situated on the border between the novel and other discourses. The real problem was that, in the meantime, the morphological function of the border – providing a favorable terrain for the encounter between the novel and other discourses – had become more uncertain. A century earlier, a novel engaging the nuances of spiritual au-

42 "By introducing literary criticism, satire of political and social evils, and popular illustrations of interesting facts in science, I have hoped to add to the interests of a romance, in which I trust no deficiency of adventure, plot, and carefully developed character will be found. But the day has gone by for mere fashionable novels. The age is utilitarian, and even novelists (the poets of present times) must conform to the mode"

43 "The time is at hand when all the tottering monuments of ignorance, credulity, and superstition, no longer protected by the foolish awe which they formerly inspired, shall strew the earth with their wrecks! Every where the young shoots of reason and liberty, starting from between the rents and crevices of the worn-out* fabrics of feudalism, are becoming too vigorous any longer to be checked: they soon will burst asunder the baseless edifices* of self-interest* and prejudice, which have so long impeded their growth. Religious inquisition, judicial torture, monastic seclusion, tyranny, oppression, fanaticism, and all the other relics of barbarism, are to be driven from the globe." (Thomas Hope, *Anastasius, or, Memoirs of a Greek*, 1819).

44 "a stately entrance hall in the most fashionable quarter of the metropolis, embellished with lofty Ionic columns of sham Sienna marble; in front of each a magnificent bust of sham bronze by Mr. NoUekins* on a pedestal of scagliola. From a heavily stuccoed* ceiling, wrought in the classic manner, depend six enormous lanterns in the Pagoda style, wreathed with gaping serpents. Along three sides there are rows of "em pire*" benches, covered with amber damask, on which are lolling a regiment of drowsy myrmidons in rich liveries*. Passing these glorious athletes, you enter an ante-room choked with chairs, sofas, settees*, whose florid gilding is heightened by scarlet cushions. Very beautiful. (Lewis Wingfield, *Abigel Rowe. A Chronicle of the Regency*, 1883).

45 "The shield, slung to his neck, bore no emblazonry, and his open baronet and pennon-less* lance argued him neither to have undergone the clapham, or knightly box on the ear (!); nor the osculum pads, which more gently signified the chivalric brotherhood. He was, however, well mounted and perfectly armed. Judging from his simple habergeon, and a silver crescent which he bore, more in the way of cognizance than as his own device, he might be pronounced a superior retainer in the service of some great feudatory." (Edward Duros, *Otterbourne; A Story of the English Marches*, 1832).

46 Georges Canguilhem, *The Normal and the Pathological*, 1966, New York 1989, p. 144.

47 Niklas Holzberg, "The Genre: Novels proper and the Fringe", 1996, in Gareth Schmeling, ed. *The Novel in the Ancient World*, revised ed., Brill, Boston-Leiden 2003.

tobiography, the mechanics of letter-writing, or the discontinuity of "sensation" could still grow into a masterpiece, and spawn a successful subgenre: *Pilgrim's Progress*, *Pamela*, *Tristram Shandy*, perhaps still even *Waverley*, had significant fringe-like traits. But in the course of the nineteenth century – probably as a consequence of the division of intellectual labor, which increased the distance between fiction and the social sciences, making their languages less and less translatable into each other – the role of heteroglossia within the development of novelistic form became problematic. It was this that decided the fate of those forgotten writers.[48]

Whether this also answers our initial question – on the archive changing our knowledge of literature – is not for us to say. What we can say is that, as the work proceeded, we found ourselves devoting more and more time to *Ringrove*, *The Impostor*, and *Arthur Montague*; and that, in a few lucky moments, we felt that these books were raising questions that, say, *Adam Bede* never would. A *few* lucky moments: it isn't easy, keeping your focus on the archive. In part, it is the pull of well-known writers – the pull of what you already know – that draws you back to the beaten track. In part, it is the troubling nature of what forgotten authors force you to face: a vast wreck of ambitious ideals, very unlike the landscape literary historians are used to study. Learning to look at the wreck without arrogance – but also without pieties – is what the new digital archive is asking us to do; in the long run, it might be an even greater change than quantification itself.

conquering Egypt, was# unable to# extirpate, but contrived to# expel, thereby entailing on# Europe their# lawless and# unpopular posterity, so obnoxious to# the# proprietors of# the# localities they select for# their# temporary residences. I see a# column of# slow rising smoke Overtop the# lofty wood that# skirts the# wild. c# 2 A# vagabond and# useless tribe# there eat Their# miserable meal. A# kettle slung Between two poles upon a# stick transverse Receives the# morsel****** of# cock purloined From# his accustomed perch. Hardening race! They# pick their# fuel out of# every hedge, Which#, kindled with# dry leaves, just saves unquench*'d The# spark of# life#. The# sportive wind blows wide Their# flutt* `ring rags, and# shows a# tawny skin--. The# vellum of# the# pedigree they# claim. Great skill have they# in# palmistry, and# more To# conjure clean away the# gold they# touch, Conveying worthless dross into its place: Loud when they# beg, dumb only when# they# steal." A# grove of# tall poplars* formed a# conspicuous object from# the# western look-out*; and# not far from# hence rose, up the# slope of# a# hill, a# dense extensive coppice, impervious to# the# eye, where the# lordly chief of# the# forest reared its# head proudly over its# arboreous companions, silently asserting its# supremacy; and# the# graceful beech, silvery ash, dark-green* spiral fir, Scotch larch, and# stunted hazel, were blended together, and# the# stream-wooing* willow dipped its# pensile shoots into# a# clear, gurgling stream, that# wound its# tortuous course along, its# sequestered, shady nooks pointing out# to# the# angler the# probable haunts of# the# hungry trout on# the# alert for# its# insect diet, and# snug spots# under the# gnarled roots of# undermined antique trees growing on# the# banks of# the# encroaching brook, hinting to# juvenile poachers, setters of# night-lines*, the# likely lurking-places* of# the# snake-like*, slimy eel. Situated in# a# dell, at no great# distance off, was# the# home-farm*, with# its# roomy barns, high granary, cow-sheds*, and# fowl-house*, on# entering# which#, perhaps the# cackling hen gave notice of# her sedentary occupation, or# the# outstretched neck of# the# hissing goose apprised you of# her# displeasure at# your approach. Without, probably the# clustering poultry, emitting their# various cries, surrounded you# without# alarm, expecting to# receive a# shower of# grain in# reward for# their# courage and# confidence; a# turkey or# two#, may be fearing to# be# late for# the# fare, running greedily up# the# yard to# join the# rest, and# the# gaily-dressed* peacock condescending to# associate with# his# inferiors* on# the# occasion. Roaming about#, you# doubtless encountered the# sheep-dog*, if# not# away#, attending on# his# fleecy charge in# adjoining pastures, the# nature of# his# bark denoting* delight oranger*, according to# his# knowledge or# ignorance of# your# countenance; and# in# passing the# sties*, you# naturally glanced at# the# swollen carcases* of# the# noxious inmates, lying in# their# miry beds surfeited* with# food, scarcely willing to# open# their# small eyes or# lift their# snouts* from# the# stone# troughs on# which# they# rested; a# short, low# grtint* perchance being the# only# acknowledgment of# their# consciousness of# the# presence of# a# visitor. The# farm-house* was# the# very picture of# rustic comfort-- a# model of# cleanliness and# neatness Within; and# the# brickwork* of# the# exterior almost totally hidden by# the# undying ivy that# clung tenaciously to# every# part, as# if# resolved not# to# separate from# a# pleasing acquaintance. A# primly clipped box-hedge* bounded it# on# one side#, and#, running# along# in# front#, was# a# wattled paling supporting a# mass of# white jessamine. The# dairy lay at# the# back, and# its# whitewashed walls were# always# well# garnished with# parallel tiers of# Stilton, sage, cream, and# other cheeses. Thus was# ensured a# certain supply of# creature comforts, contributing in# no# small# degree to# create that# full contentment that# pervaded the# household, where# food# was# abundant, beer and# cider plentiful, and# work light. A# few hundred yards from# the# farmstead was#" The# Retreat," where# I#, Arthur Montague, (for# it# is fitting I# should begin to# speak in# propria* persona) was# wont to# pass many an# hour in# listless idleness, looking on# the# blooming landscape, listening to# the# humming bees, or# teasing a# pet jackdaw, who# poked his# head# between# the# bars of# his# wicker cage when# confined there# for# misconduct. The# said Retreat# was# an# elegant little two-roomed* Gothic cottage, plastered with# sparkling sanded cement

Figure 6.2 continued.

48 By the same token, from that moment on the masterpieces of heteroglossia – like *Moby-Dick*, or *Ulysses* – had to move increasingly away from the main axis of novelistic development, appealing less and less to non-academic readers.