UNDERSTANDING GENRE IN A COLLECTION OF A MILLION VOLUMES (1800-1949):

enriching metadata in a large collection while acknowledging the mutability of genre

Table of Contents

Understanding Genre in a Collection of a Million Volumes

Level II Digital Humanities Start-Up Grant: Requesting $57,163 from NEH.

List of participants and collaborators:

Ted Underwood, primary investigator
Advisory board of scholars may include

Eleanor Courtemanche, Associate Prof., UIUC (19th century British fiction)
Stephanie Foote, Associate Prof., UIUC (19th/20th century American fiction)
Christopher Freeburg, Assistant Prof., UIUC (19th/20th century American literature)
Andrew Gaedtke, Assistant Prof., UIUC (20th century British literature)
Harriett Green, English and Digital Humanities Librarian, UIUC
Justine Murison, Associate Prof., UIUC (19th century American literature)
Timothy Newcomb, Prof., UIUC (20th century poetry)

Collaborators at HathiTrust Research Center, to include:

J. Stephen Downie, HTRC co-director; Associate Dean for Research, GSLIS
Loretta Auvil, senior research programmer, HTRC and Illinois Informatics Institute

Collaborators at NCSA may include:

Alan Craig, NCSA liason to XSEDE
Mark Straka, senior research programmer, NCSA

Understanding Genre in a Collection of a Million Volumes

Abstract: Large digital collections offer new avenues of exploration for literary scholars. But their potential has not yet been fully realized, because we don't have the metadata we would need to make literary arguments at scale. Subject classifications don't reveal, for instance, whether a given volume is poetry, drama, fiction, or criticism.

Working with a hand-classified collection of 4,275 English-language works, we have discovered new perspectives on the history of genre. But to flesh out those leads (and permit others to undertake similar projects) we need to move to a scale where manual classification would be impractical. We propose to develop software that can classify volumes by genre while allowing definitions of genre to change over time, and allowing works to belong to multiple genres. We will classify a million-volume collection (1800-1949), make our data, metadata, and software freely available through HathiTrust Research Center, and publish substantive literary findings.

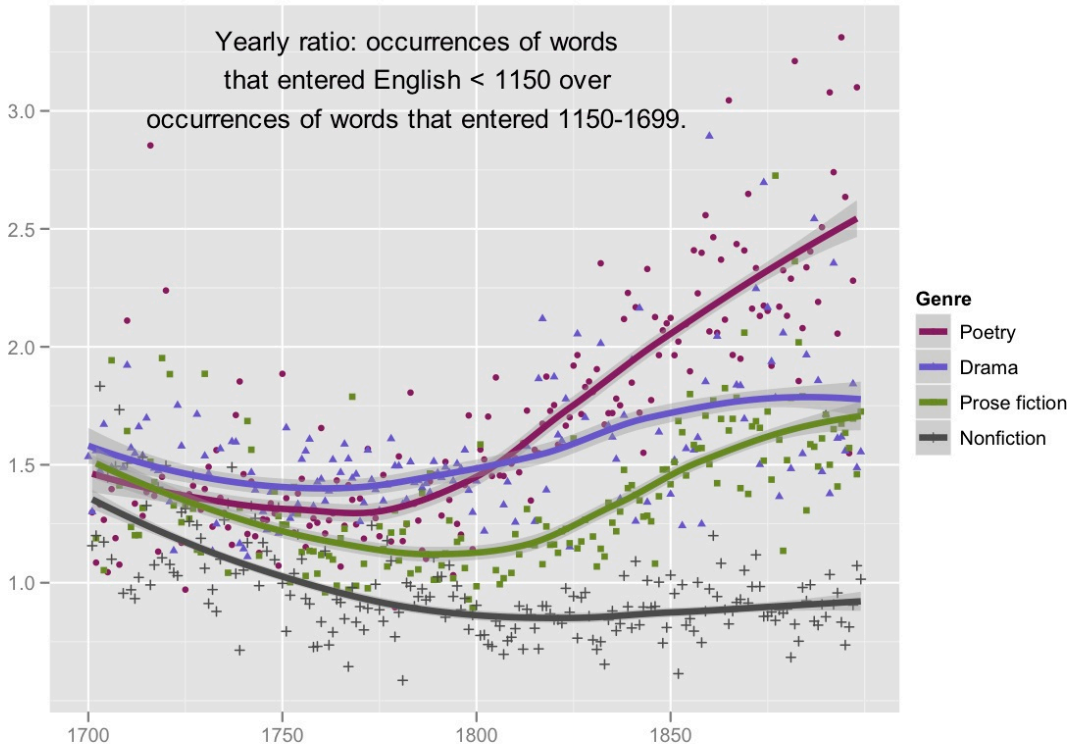
Statement of innovation: In a collection that spans more than a few decades, the definitions of genres are mutable. To address that problem, we will develop a classification strategy that allows the fingerprint of a genre to change continuously over the timeline. Because no single scheme of classification is perfect for every inquiry, our code will allow other researchers to define their own categories. Working with HTRC, we will extend this approach to collections that lie partly outside the public domain (1800-1949).

Statement of humanities significance: Our preliminary work on poetry, fiction, and drama in a collection of 4,275 volumes has already produced new insights about the differentiation of literary genres from nonfiction prose in the period 1700-1899. We propose both to develop those leads, and to explore what can be learned about subtler processes of subgeneric differentiation — for instance, the differentiation of self-consciously literary fiction from so-called "genre" fiction in the nineteenth and early twentieth centuries.

Narrative.

The lively debate about distant reading in the last few years may seem to imply that many scholars are already studying literary history digitally on a macroscopic scale. In fact, we are only beginning to organize the kinds of collections we would need for that research. Many of the best existing collections are limited to a single genre or century. In larger collections, the metadata that would interest literary scholars (genre, form, gender, nationality) are usually missing.

When enriched metadata are combined with scale and breadth, startling results can be discovered. For instance, when Sellers and Underwood classified a collection of 4,275 volumes by genre, they were able to uncover new evidence about the specialization of literary diction. It is already recognized that modern definitions of literature took shape between 1700 and 1900. In the early eighteenth century "literature" meant learning or written discourse generally; by the end of the nineteenth century the word described a category of writing set apart by specifically fictive and aesthetic aims. We're now able to show that this conceptual differentiation of literary and nonliterary genres paralleled a corresponding change in writing practices that set fiction, drama, and poetry apart from nonfiction prose on the level of diction.



In particular, the older part of the lexicon (words that entered English before 1150) became much more common in fiction, poetry, and drama than in nonfiction: a differentiation that is absent in the early eighteenth century (Underwood and Sellers 2012). These findings have the potential to complicate critics' received histories of "poetic diction"; for a fuller account of that argument, please see Appendix A. We are glancing at it only briefly here, to show that digital research on genre can produce leads of broad significance for literary scholarship.

While we have published some of our initial findings, we would need a larger collection to develop these leads to their full potential. A collection of 4,275 volumes represents less than 1% of what is potentially available in the nineteenth century. If we, and others, want to think about genre in a detailed way (biographies, gothic novels, sensation novels), we need to work on a larger scale.

Generic classification on a large scale is far from simple. While the Library of Congress does define genre/form headings, they are often missing in MARC records, and cannot usually be inferred from call numbers. Moreover, scholars might not even want to use a single fixed set of genre classifications. As Carolyn Miller and Amy Devitt have pointed out, genres are social practices rather than fixed forms. The definition of a genre, and even the meaning of generic categorization itself, can vary over time (Devitt 2004). Categories like "fiction" and "nonfiction" may be moderately stable. But genres like the gothic novel mutate and subdivide. Just as importantly, genres are sites of critical contestation, and critics may need to redraw boundaries for particular questions.

These instabilities actually strengthen the argument for an algorithmic approach to genre. We might, arguably, be able to crowdsource generic classification in a collection of a million volumes — once. But we cannot ask dozens of human readers to redo that work every time a scholar wants to redraw the boundary between science fiction and fantasy. It would make more sense to design software that can repeatedly divide large collections into categories appropriate for a given research question. Since membership in a genre is not a black-and-white matter, we will need to allow works to belong to multiple categories with different degrees of confidence; that approach will allow us to trace processes of generic fission without needing to define a single moment of division. We will also need to allow the definition of each genre to change continuously over the time axis. Software with these kinds of flexibility would both allow us to explore the history of genre in its own right, and provide a foundation for other digital projects that become more meaningful when enriched with metadata.

In applying digital methods to the history of genre, our goal is not to eliminate ambiguity, but to describe gradual processes of change more subtly (Unsworth 2008). In our prototype collection, for instance, we provisionally divided prose works into "fiction" and "nonfiction." But at the same time, the illustration above reveals that the space of separation between fiction and nonfiction varied over the course of the eighteenth and

nineteenth centuries. We can now explore that process of differentiation from many different angles. In other words, the point of our project is not simply to categorize works, but to reveal how the significance of categorization has varied over time.

While we're sensitive to these ambiguities, we also aspire to produce metadata that are practically useful for other scholars, at least in their broad outlines (verse, prose, drama, fiction, nonfiction). See the Data Management Plan for further details. We do not expect to replace the work of librarians. We are aiming for 99% accuracy — more than sufficient for distant reading, but not for permanent archival purposes.

History of the project: The early stages of this project have been supported by the Andrew W. Mellon Foundation as a use case under "Expanding SEASR Services." In particular, that grant funded research assistants who helped the PI develop a collection of 4,275 volumes, and write software for regularizing HathiTrust texts and metadata.

We have also negotiated commitments from HathiTrust Research Center, XSEDE, and the University of Illinois that will support the project going forward. HathiTrust has already provided a collection of almost a million English-language volumes, and we expect to receive more. XSEDE has provided an initial allocation of 30,000 hours on the Blacklight supercomputer. The University of Illinois has funded resources for distributed management of large datasets on the Illinois Campus Cluster. Moreover, we have written software that solves most of the data-cleaning problems we will confront (for instance, we can correct optically-scanned texts, automatically separate verse from prose, and make good guesses about article divisions within a serial volume). Finally, in the last two months we have demonstrated that our algorithms can reliably classify volumes by genre (see Appendix C).

We still need to address the issue of historical change, and implement our solutions in a collection of a million volumes. Ted Underwood will be doing most of the coding as well as literary analysis on the project. He is the author of two books on nineteenth-century literature, but also reads extensively in the field of machine learning, and has the coding skills necessary to implement machine-learning solutions at scale. (For instance, he has written a parallel version of Latent Dirichlet Allocation in Java, now being implemented on the Blacklight supercomputer). He is also fluent in Python and R. The literary advisory board and collaborators at NCSA will serve mainly in a consulting capacity, answering technical or literary questions that lie outside Underwood's area of expertise. Much of the manual classification of individual volumes in the training set will be done by undergraduate researchers, supervised by a graduate assistant, and tested for inter-rater reliability.

Finally, because it is important to demonstrate that work of this kind can be carried out in periods still covered by copyright, we have arranged to become an early use case at the HathiTrust Research Center. We will send our data-cleaning algorithms into HTRC, and HTRC will export word counts for volumes (and volume parts) that we can use for classification. Since this part of the project is vulnerable to legal vicissitudes, it is worth

emphasizing that 1924-1949 is only a small part of the period we aim to cover; if the project of non-expressive research were somehow blocked by the courts, it would still be possible for us to produce significant results on a collection of more than a million volumes.

Environmental scan: Researchers already have access to large collections of digital text and data-mining tools. But relatively little literary research is taking place on the scale we envision, because interesting kinds of metadata are missing. While subject classifications are available (for instance, in Harvard's Bookworm collection), these don't illuminate questions of literary genre. Stanford's collection of nineteenth-century novels is categorized by genre. But that collection doesn't include poetry, drama, or nonfiction; it is built using existing bibliographies; and it is roughly the same size as our prototype (five thousand volumes). We will use inductive methods on all the English-language works HathiTrust holds in the period 1800-1949, producing a collection about two orders of magnitude larger.

We need to classify texts ourselves for two reasons. Pragmatically, there is no other option. Genre/form headings are often missing in MARC records, and cannot always be inferred from call numbers. We also believe that doing classification ourselves will give us a more profound understanding of genre: for instance, assigning a measurement of confidence to each generic tag will help us study processes of gradual differentiation.

Several different teams have already shown that it is possible to classify literary works by genre, using evidence as simple as word counts. Alison, Heuser, Jockers, Moretti, and Witmore were able to sort works of fiction into subgenres like "the sensation novel" and "gothic novel" using a limited range of textual features (Alison, Heuser, et. al., 2011). But their work did not address the historical mutability of genre, or consider error rates at scale. To address scalability, we have divided 1,340 nineteenth-century works into "prose fiction," "poetry," "drama," and "nonfiction prose," identified sources of error, and developed a plan to improve performance. (See Appendix C.) Just as importantly, we have identified confidence metrics (drawn from Schneider 2005) that can reliably identify the subset of the corpus that will still require disambiguation by a human eye.

A project like this needs to engage several different disciplinary traditions. The problem of classification is a central area of research in machine learning, and a great deal is known about best practices (Rokach 2010; Han, Kamber, and Pei 2012). But classification strategies developed for short genres, like web pages, may not be fully applicable to literary genre (Yu 2008). Moreover, the subfield of "genre theory" — positioned between literary studies and rhetoric — has raised fundamental challenges to ordinary forms of classification. Scholars like Carolyn Miller and Amy Devitt point out that genres are social practices evolving in conversation with each other, rather than fixed forms (Devitt 2004).

Our goal is to develop a machine-learning strategy that can acknowledge and illuminate these theoretical challenges. For instance, genres overlap: it is possible for a work to be at once poetry and drama, or science fiction and detective fiction. A volume can

also be a mosaic of genres. Nineteenth-century collections of poetry often have long prose introductions. Most importantly, the fingerprint of genre can change over time, as Michael Witmore has pointed out (2012). So we need a strategy that permits classification criteria to vary continuously over the timeline. Just to illustrate one possibility, we could train multiple classifiers on different, overlapping slices of time, and use the precise date of a given document to weight their "votes."

No matter how sophisticated we make our classification strategy, literary scholars will initially be skeptical of results based on an automatically-classified collection. We therefore propose to develop a manually-classified training corpus that is large enough to serve as a research collection in its own right. (We are aiming for 10,000 volumes, or roughly 1% of the whole.) We may need to use this smaller, manually-classified subset in order to confirm the conclusions we draw from the larger, automatically-classified collection. In particular, we will need to use it to respond to charges of circularity that bedevil all research on genre (see Appendix B).

On the other hand, we do not believe that a purely manual, crowdsourced strategy would make good use of the opportunities presented by a collection of a million volumes. Preconceptions about genre might make certain patterns invisible to scholars who are working through the collection at ground level. More importantly, crowdsourcing would be an unrepeatable process, and schemes of classification will in fact need to be redrawn. We are therefore proposing a coordinated combination of manual and automatic strategies.

Work plan: By the time work begins on this grant (May 2013), we will have completed most of the data preparation on a collection of a million volumes 1800-1923. This will require: automatically weeding out works where the optical transcription is too poor to correct, correcting errors in the works that remain, dividing serial volumes automatically into articles (using clues provided by changes in running headers), normalizing metadata, regularizing spelling, and separating poetry from prose in volumes that are a mosaic of both. Most of the software we need has already been written and tested. The next steps:

May 2013. We begin with unsupervised clustering and topic modeling, as exploratory processes, in order to discover genres that don't fit our preconceptions. Guided by occasional input from an advisory board of scholarly experts, we will identify a tentative framework of genres and subgenres (in nonfiction prose as well as fiction, drama, and poetry).

June 2013. Underwood and a graduate research assistant will manually classify a randomly-selected subset of 2,000 volumes, to supplement our existing 4,275-volume collection. This will provide an initial corpus for training classification algorithms, to be expanded as work continues. We will set aside another 2,000 works as an evaluation test set, not to be used in development.

July and August 2013. Test and scale up an ensemble of classification algorithms. We have already found that naive Bayesian and k-nearest-neighbor classifiers work adequately (see Appendix C), but we will also test support vector and logistic regression algorithms. In general, an ensemble of approaches tends to work better than a single one (Rokach 2010). We will also evaluate the utility of adaptive boosting.

Concurrently, August 2013. Send the data-cleaning and volume-segmentation routines we have developed to HathiTrust Research Center, and extract word counts for English-language volumes 1924-1949.

September-December 2013. We will begin to classify the whole collection, using an active learning strategy that permits our algorithm to collaborate with a team of human readers. Our system will begin by identifying the volumes (or volume parts) in the collection that it finds *hardest* to classify (for instance, because different algorithms reach different conclusions). Undergraduate researchers, supervised by a graduate assistant, will manually classify these ambiguous documents.

January-March 2014. Classify evaluation test set; evaluate results. Make our collection and metadata public for other scholars.

February-October 2014. Use our enriched metadata to analyze a collection of a million volumes. Use corpus comparison to trace processes of generic differentiation. Publicly disseminate promising leads, and begin to draft a book project based on the research. Help HTRC offer our classification strategy as a service for other researchers.

How these results will support future research. While there are many quantitative ways to evaluate classification (recall, precision, F-score, and so on), the final measure of success for a project like this is simply that it supports new accounts of literary history.

We intend to ensure that success in three ways: by developing our own theses about generic differentiation, by sharing data and metdata, and by calling other scholars' attention to promising leads. Underwood is already developing a book project about the linguistic differentiation of literature from nonfiction in the eighteenth and nineteenth centuries (Appendix A). With an enlarged collection and enriched metadata, he will be able to make subtler and more persuasive arguments about the kinds of cultural capital embedded in different forms of literary diction.

But we also expect to find a range of interesting leads in areas where the PI is not expert. (For instance, we will no doubt stumble on interestingly ambiguous categories — nineteenth-century novels, say, that look like twentieth-century "genre fiction," but that were published before those genres are supposed to have existed.) These leads will be shared, first with the advisory board and graduate research assistants, and second with the broader community of literary scholars through blogging (at *The Stone and the Shell* and elsewhere). In this way we hope, not just to passively share our data, but to actively foster digital research projects broadly distributed across English-language literary scholarship, and beyond the circuit of discussion ordinarily associated with "digital humanities."

Data management plan.

Classifying a large collection by genre is partly a service for other scholars, and partly a way to start interesting arguments. We don't expect other critics to passively adopt the boundaries we draw for subtle subgeneric categories like "gothic romance" or "the sensation novel." On the contrary, our goal is to highlight interesting ambiguities, and disseminate code that other scholars can use to do their own classification. But we do think it will be possible to establish broadly reliable generic metadata for categories like "verse," "drama," "prose fiction," and "prose nonfiction," and those results should be easy for other researchers to borrow.

Our goal is to make the products of this research as public as possible given United States copyright law. We are using public-domain data wherever possible. Even the portion of the work after 1923 will be done through HathiTrust Research Center, an institution that seeks to facilitate public non-expressive access to works otherwise in copyright.

The collection produced by this project will reside at HTRC. All the metadata we produce will be public; it will be saved both as individual .json files associated with each text, and as a .csv file for the whole collection. Where we identify article divisions or internal boundaries between prose and poetry, these will be marked on the text using TEI-lite. TEI files before 1924 will be public-domain, although special arrangement with HathiTrust may be necessary to access a subset of those files that were originally digitized by Google. After 1924, the metadata will still be public, but the texts themselves will have to reside at HTRC.

The software tools we propose to develop will be open-source, and available on github. Some portions may be built in Python, but the core will be written in Java for reasons of performance and scale; it will be designed to accept HathiTrust data structures. More importantly, all the tools and resources we develop will be embedded in HathiTrust Research Center as parts of a research infrastructure immediately available to other scholars. This will include:

- data-cleaning modules that remove running headers and segment volumes,
- lists of rules for OCR correction in particular periods, and algorithms for contextual correction of phrases ("mortal fin forgiven" => "mortal sin forgiven").
- general-purpose software for automatic genre classification,
- a model specifically trained to classify works in the period 1800-1949,
- manually-classified genre metadata (and author gender) for a collection of roughly 10,000 volumes 1800-1949,
- automatically-classified genre metadata for a collection of more than a million volumes 1800-1949,
- a list of volumes in the collection that proved impossible to classify, which might be one of the most interesting products of this research — providing useful leads even for literary scholars who are not interested in distant reading.

Biographies

Ted Underwood (PI) is Associate Professor of English at the University of Illinois, Urbana-Champaign. He is the author of two books on nineteenth-century literature, *The Work of the Sun: Science, Literature and Political Economy 1760-1860* (Palgrave, 2005) and *Why Literary Periods Mattered: Historical Contrast and the Prestige of English Literature* (finished, under review at Stanford). His articles on nineteenth- and twentieth-century literature have appeared in *PMLA, Representations, MLQ, Studies in Romanticism,* and *The Journal of Digital Humanities.*

Underwood's interest in informatics dates back to the 1980s, when he worked as a software developer for Artificial Intelligence Atlanta, producing a prototype of an information-retrieval expert system in Prolog. More recently, he has collaborated with the Illinois Informatics Institute to correct optical transcription errors in the Google ngrams database so that it can be used to study the eighteenth century. He stays current in the fields of machine learning and knowledge discovery, and has written software to do Bayesian OCR correction, clustering, topic modeling, corpus comparison, and genre classification in Java, Python, and R. He currently holds a fellowship from the Institute for Advanced Computing Applications and Technologies, and serves as Associate Director of NINES.

Graduate research assistant, TBA
Undergraduate research assistants, TBA

The advisory board of literary scholars will guide the project when it confronts particular questions about genre that fall outside the literary expertise of the PI.

The advisory board will include

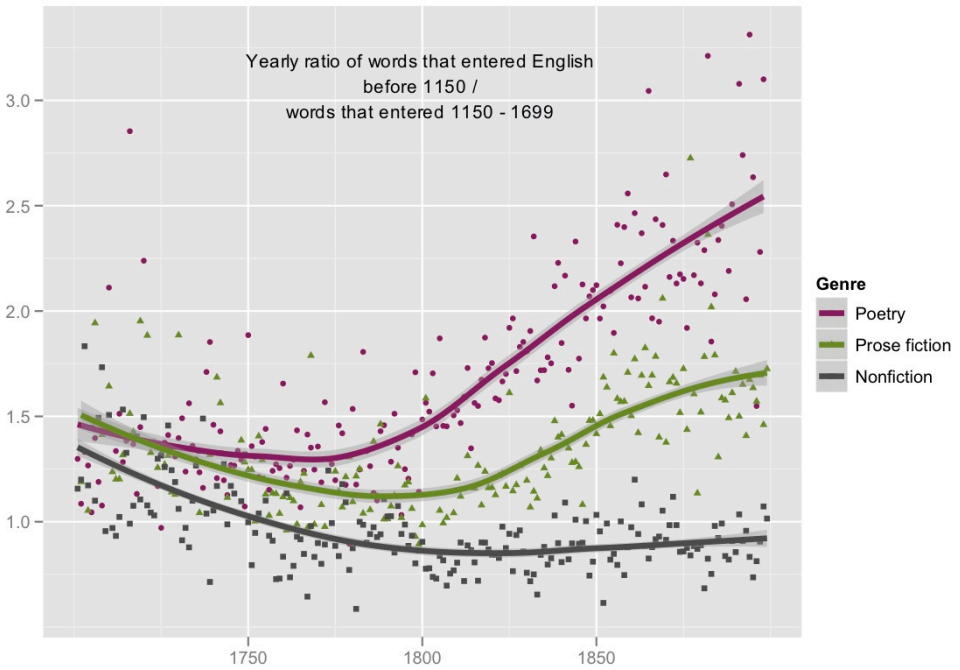
Eleanor Courtemanche, Associate Prof., UIUC (19th century British fiction)
Stephanie Foote, Associate Prof., UIUC (19th/20th century American fiction)
Christopher Freeburg, Assistant Prof., UIUC (19th/20th century American literature)
Andrew Gaedtke, Assistant Prof., UIUC (20th century British literature)
Harriett Green, English and Digital Humanities Librarian, UIUC
Justine Murison, Associate Prof., UIUC (19th century American literature)
Timothy Newcomb, Prof., UIUC (20th century poetry)

Appendix A. On the problem of "literary diction."

This project has two goals. It serves scholars of English literature by enriching a large digital collection with generic metadata. But on that foundation, it also aims to build a more pointed argument that uses diction to trace the differentiation of genres.

The word "diction" is usefully vague. At times, it approximates the sociolinguistic concept of "register" — a variety of language used in a particular social context (formal, casual, intimate, and so on). But there are also narrower phenomena like "poetic diction." Even subject categories can be associated with distinct patterns of word choice. We tend to call the language associated with a subject "specialized vocabulary" rather than "diction," and imagine it as reflecting content rather than style. But in reality that boundary is blurry. Consider literary criticism. Critics have an overtly specialized vocabulary ("overdetermination"), but we may also deploy certain ordinary words with heightened frequency ("deploy" and "heightened," for instance). Similarly, a subgenre like detective fiction will display both explicit jargon ("gumshoe," "dame") and patterns of word choice that are shaped more subtly by the characteristic attitudes and situations of the genre.

Even when we talk about diction in a fairly narrow sense — like "poetic diction" — literary critics lack a firm grasp on the phenomenon the word describes. For instance, students are often taught that William Wordsworth's intervention in *Lyrical Ballads* mattered because it challenged artificial poetic diction, in order to reduce "the space of separation betwixt Prose and Metrical composition" (Wordsworth 1800). In one sense this is true: poetry did lose some specialized vocabulary in the nineteenth century. But in a broader sense the truth may be closer to the reverse. Where register is concerned, the "space of separation" between poetry and prose increased, because poets began to use the older part of the lexicon much more heavily than writers of prose (and especially writers of nonfiction).



The graph above is based on a collection of 4,275 (mostly book-length) documents; to make it readable, we have plotted yearly values rather than individual works. In each year, we count the number of words (tokens) that entered English before 1150, and divide it by the number of words that entered the language between 1150 and 1699. (We consider only the most common ten thousand words in the collection, and exclude function words: determiners, prepositions, conjunctions, and pronouns. Prose introductions and notes were removed from volumes of poetry.)

Why do this? and what can it tell us about the register of writing? In English, etymology often has social implications, because the English language was for 200 years (1066-1250) almost exclusively spoken, while French was used for writing. The learned part of the Old English lexicon didn't survive this period. Instead, when English began to be written again, literate vocabulary was borrowed from French and Latin. As a result, the boundary between older and more recent parts of the lexicon also tends to be a social distinction between relatively informal and learned/literate language. This was clearly true in the thirteenth century, and linguists Laly Bar-Ilan and Ruth A. Berman have recently demonstrated that it remains true today (Bar-Ilan and Berman 2007).

So the graph above shows that, while all genres of writing tended to adopt a more learned diction in the eighteenth century, poetry and fiction decisively reversed course in the nineteenth. (Drama did as well, although I have left it out of the graph above for clarity. See the main narrative for the proposal.) As a result there was by the end of the nineteenth century a new, sharply marked distinction between literary and nonliterary diction: novels were using the older part of the lexicon at a rate almost double that of nonfiction prose. Poems were using the older part of the lexicon at a rate almost three times higher.

Moreover, these changes affected individual words very similarly in different literary genres. So while poetry may have lost some of its specialized vocabulary and orthography ("poetic diction"), it appears that the nineteenth century also saw the emergence of a new, broadly literary diction shared by poetry, fiction, and drama. We have argued elsewhere that these new writing practices were associated with a new definition of "literature" as a category of writing set apart from nonfiction by fictive and aesthetic aims, and especially by an emphasis on individual subjectivity (Underwood and Sellers 2012).

In the work supported by this grant, our goal is partly to extend this thesis and test it more rigorously against a larger collection. But we are also hypothesizing that it will be possible to find similar patterns of linguistic differentiation between other subgenres.

By tracing these changes at a linguistic level, we hope to better understand processes of generic differentiation. The emergence of a genre like "science fiction" can be traced partly by graphing rising production, which is something we already know how to do (see Moretti 2005). But the emergence of a genre is also a matter of establishing clearly differentiated conventions and expectations, and we have found that those changes always leave traces at the level of diction. We can already see that it is possible, for instance, to

identify lexical features that mark the gradual differentiation of the novel from nonfiction, and in doing so we have learned that different literary genres parallel each other to a surprising degree in the nienteenth century.

We don't expect that the particular etymological metric used above will always be appropriate. But we trust that it will be possible to define some appropriate metric of differentiation in each instance. In some cases, it may be possible to use a metric as simple as cosine similarity to trace the increasing distance between two genres. In other cases, the classification algorithms we use to define a genre may double as measures of differentiation — for instance, it seems likely that our metrics of classification confidence will increase as two genres disentangle from each other. More firmly defined best practices for tracing generic differentiation are part of what we propose to develop under this grant.

Appendix B. Problems of circularity in the study of genre.

The study of genre has always confronted problems of circularity. As Andrew Tudor pointed out in 1974, we cannot describe the "principal characteristics" of a genre until we isolate a group of exemplary works. But how can we identify examples of a genre before we know the genre's principal characteristics?

To break that vicious cycle, Tudor suggested that scholars simply have to begin with examples of a genre that are recognized by "common cultural consensus" (Tudor 1974). That is of course why we propose to begin with a manually classified training corpus. Moreover, digital methods make it easy to build outward from a set of examples without ever formally defining a fixed set of characteristics. We can instead use tacit affinities to classify works, and even allow the fingerprint of a genre to vary continuously over the timeline. But some level of circularity remains unavoidable: it is still true that we will use diction to classify works, and then study generic differentiation by tracing changes in diction.

We believe that abstract kind of circularity poses a mainly rhetorical obstacle. If our criteria of classification vary flexibly over time, shaped by examples of a genre in a given period, we won't have to impose any fixed standard of diction. So it should still be possible to use automatically-classified volumes to trace the history of generic differentiation.

But rhetorical obstacles are not insignificant; we want to develop methods, after all, that actually convince literary scholars. That is why we have insisted on also producing a smaller, but substantial, manually-classified collection of 10,000 volumes. We may need to use that small collection to address questions about circularity, while we use the larger, automatically-classified collection to address questions about inclusiveness.

Of course, we will also have to rely on close readings of individual texts in order to interpret the literary significance of the changes we discover. Quantitative methods can reveal patterns of change that wouldn't otherwise fit into a reader's field of vision. But no matter how rigorous our methods become, they can't replace literary interpretation itself.

Appendix C. The technical feasibility of genre classification.

Alison, Heuser, Jockers, Moretti, and Witmore have demonstrated that clustering methods can highlight literary genres, including subgenres like "tragedies" and "history plays" (Alison, Heuser, et. al., 2011). But before attempting this in collection of a million volumes, we need to think carefully about error tolerance. We explicitly anticipate that there will be many ambiguous cases requiring manual classification; that is not itself a problem. But we need to be confident that we can catch those ambiguous cases, and reduce their numbers to a point where manual classification is practical.

To assess the difficulty of this problem, we have carried out some simple proof-of-concept experiments on our existing collection of 4,275 volumes — focusing more specifically on a subset of 1,340 nineteenth-century volumes, including poetry, drama, prose fiction, and nonfiction. (We ensured that biographies were well represented in the nonfiction corpus, since we anticipate that biography will be the genre of nonfiction hardest to distinguish from fiction.)

Results

In these initial experiments, we tested both naive Bayesian and k-nearest-neighbor algorithms. Both performed well, and it seems clear to us that the technical problem is going to be soluble. It's not solved yet, but that is after all the work we propose to do.

For Bayesian classification, we built six different classifiers, so that all four genres could be contrasted separately against three possible alternatives. We then allowed the classifiers to vote. In each classifier, 1000 features were selected using a Wilcoxon test (used heavily for feature selection in bioinformatics). The confusion matrix below indicates how many works in each genre (column) were classified as one of the four alternatives.

	really drama	really fiction	really nonfiction	really poetry
as "drama"	85.4%	0%	0%	6.1%
as "fiction"	0%	95.1%	6.0%	0.8%
as "nonfiction"	0%	2.8%	94.0%	3.0%
as "poetry"	14.6%	2.1%	0%	90.1%

This classifier performed well with fiction and nonfiction, but relatively poorly with drama and poetry. The primary reason for this is that "drama" and "poetry" are not really

exclusive categories in the nineteenth century. You'll notice that all the misclassified dramatic works were misclassified as "poetry." In fact, they were all poetic dramas: Lord Byron's *Marino Faliero,* Felicia Hemans' *Vespers of Palermo*, and so on. Clearly, we're going to need to allow the categories of "poetry" and "drama" to overlap. Once we do that I think it will be possible to identify drama at least as reliably as prose.

The k-nearest-neighbors algorithm performed slightly better than naive Bayesian classification across the board

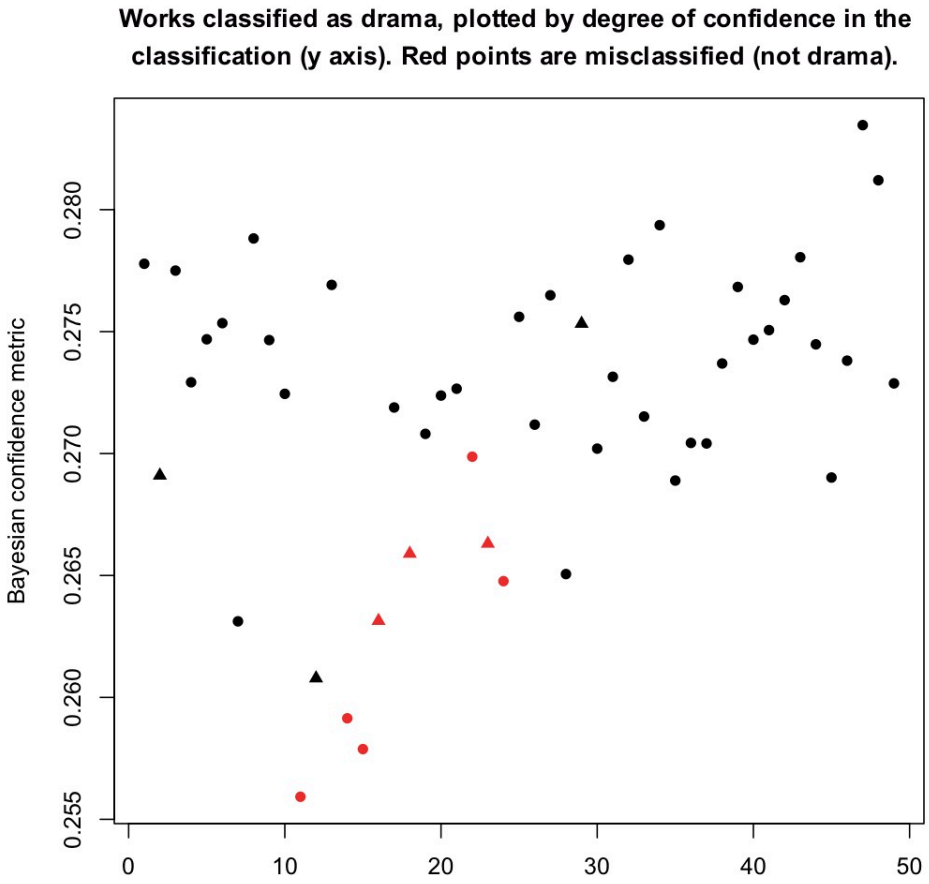
	drama	fiction	nonfiction	poetry
as "drama"	91.7%	0%	0%	3.0%
as "fiction"	4.2%	95.8%	4.5%	1.5%
as "nonfiction"	0%	2.8%	94.8%	3.1%
as "poetry"	4.1%	1.4%	0.7%	92.4%

In the final project, we expect to combine several classifiers in an ensemble. We will test not just the approaches demonstrated here, but SVM and logistic regression. According to Ng and Jordan (2002) those algorithms tend to outperform naive Bayes as the training set grows larger. We eventually aim to increase accuracy to 98-99%, using several strategies:

1) combining algorithms in an ensemble,
2) using clues from metadata; for instance, call numbers will often support a strong suspicion that a given volume is (some kind of) nonfiction,
3) gradually developing a larger training corpus, with active learning,
4) recognizing internal divisions in a volume, so that (for instance) the long prose introduction to a volume of poetry won't prevent the classifier from recognizing the rest of the volume as verse.

Even more important than the question of accuracy is the goal of recognizing ambiguity. Our active learning strategy requires identifying a subset of ambiguous works first, so that we can "teach" the classifier to discriminate between hard cases. But this strategy will only work if we can recognize ambiguity! Fortunately, we have found that the metrics of confidence described by Schneider (2005) work well for this purpose. The illustration below graphs all works that the Bayesian classifier categorized as "drama,"

including works mistakenly so classified (in red). The x axis is not meaningful (it's just an index distinguishing volumes from each other), but the y axis is a Bayesian metric for our confidence in the classification. The important detail here is that our algorithm tended to have significantly lower confidence in the red (misclassified) works.



Similar patterns appeared in other genres. This suggests that it will be possible to improve classification using a strategy of active learning that starts with the "bottom" (most ambiguous) data points and works up from there. Incidentally, the triangles above are works marked as "less confident" by the kNN classifier. So far that metric is less reliable than the Bayesian confidence metric, but it may be possible to combine multiple metrics as recommended by Delany, Cunningham, and Doyle (2005).

Appendix D. Other kinds of metadata.

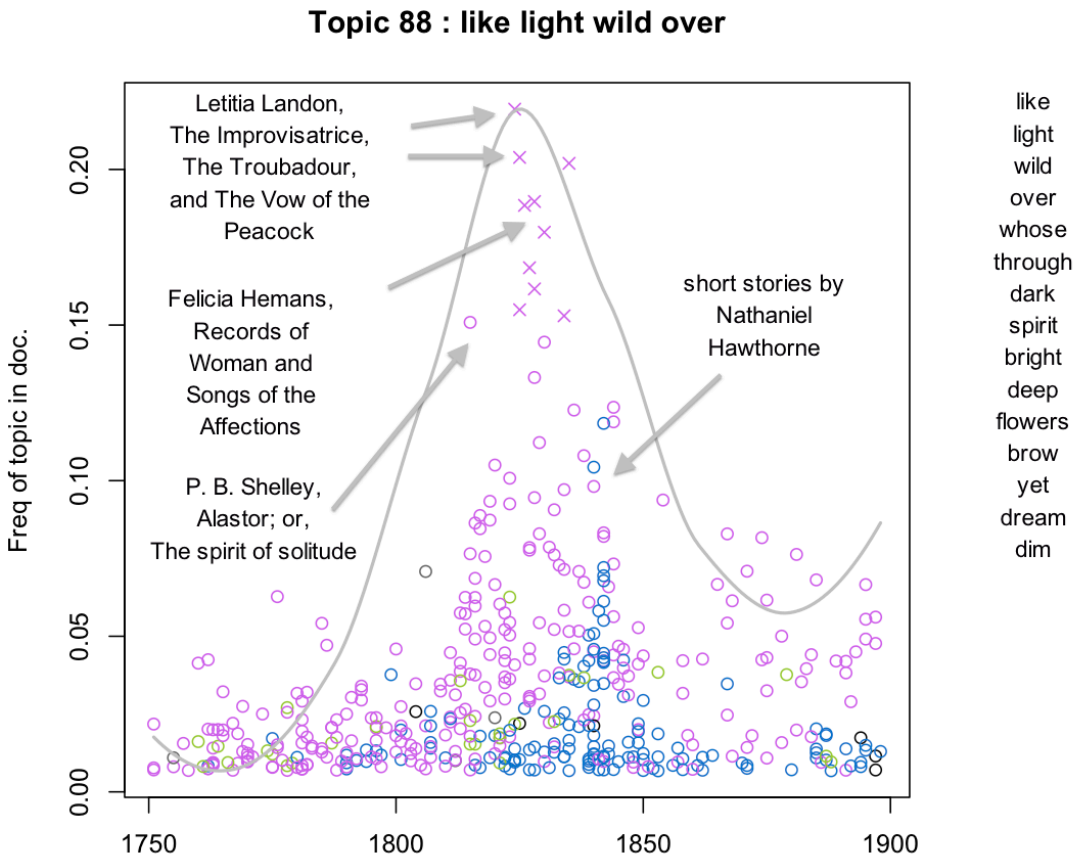
There are many kinds of metadata that literary scholars might desire beyond genre — including, for instance, the nationality, ethnicity, and gender of authors. Those are important desiderata, but a large-scale solution probably requires a different strategy than the one we are developing here, based in all likelihood on linked data. We will, however, record gender when we manually classify volumes, because it would be a shame to neglect that opportunity.

"Form" is another category of metadata, potentially distinct from genre. The difference between a genre and a form is not always easy to define, and the Library of Congress does not always in practice distinguish the two concepts (LoC 2011). But for the purpose of this project, we'll characterize short stories, novels, and novellas as fictional forms, so that the distinction between short and long fiction can cut across boundaries of genre.

We aspire to categorize forms as well as genres, but believe it might be unwise to promise this in an eighteen-month project. For instance, although length is obviously a useful way to distinguish novels from novellas and short stories, these formal distinctions may also involve factors other than length. Also, "length" is not quite as easy to assess as it sounds. In serial volumes, it may not be easy to distinguish short stories from pieces of a serialized novel. In book-length works, we may need to map patterns of proper-name recurrence in order to automatically distinguish a collection of short stories from a novel. These are not straightforward tasks. So we are not proposing to solve the problem of form yet, although it is certainly a problem we'll want to address in the near future.

Appendix E. Other ways to use generic metadata.

This proposal focuses mainly on the process of generic classification itself, and on simple kinds of corpus comparison. But generic metadata can also be useful in more complex kinds of analysis. We have found, for instance, that topic modeling is more legible when you're working with a generically-tagged collection.



Blue/fic, purple/poe, green/drama, black/bio, brown/nonfic, triangle/letters or orations.

This is a topic I generated on a collection of about 1,800 volumes; it's mostly late-Romantic poetry, especially, as it happens, poetry by women. But because this collection is classified by genre, the nice detail that becomes visible is that column of blue circles. Those are fiction — specifically, short stories by Nathaniel Hawthorne. The inclusion of fiction in this predominantly poetic topic suggests interesting critical questions about Hawthorne's style.

This is just a quick example of the way generic classification can add significance to text mining, even in contexts where genre itself is not the primary object of inquiry. This visualization is still rough: please note that the gray curve, for instance, is not to scale.

References.

Allison, Sarah, and Ryan Heuser, Matthew Jockers, Franco Moretti and Michael Witmore. *Quantitative Formalism: An Experiment.* Stanford Literary Lab Pamphlet Series, January 2011. http://litlab.stanford.edu/?page_id=255

Bar-Ilan, Laly and Ruth A. Berman. "Developing Register Differentiation: the Latinate-Germanic Divide in English." *Linguistics* 45 (2007): 1-35.

Delaney, Sarah Jane, Pádraig Cunningham, and Dónal Doyle. "Estimates of Classification Confidence for a Case-Based Spam Filter." 2005. International Conference on Case-Based Reasoning.

Derrida, Jacques. "The Law of Genre." In *On Narrative*. Ed. W. J. T. Mitchell. Chicago: University of Chicago Press, 1981.

Devitt, Amy J. *Writing Genres.* Carbondale: Southern Illinois University Press, 2004.

Han, Jiawei, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques.* Waltham, MA: Morgan Kaufman, 2012.

Jackson, Peter, and Isabelle Moulinier. *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization*. John Benjamins: 2002.

Library of Congress. *Frequently Asked Questions about Library of Congress Genre/Form Terms for Library and Archival Materials.* 2011. http://www.loc.gov/catdir/cpso/genreformgeneral.html

Moretti, Franco. *Graphs, Maps, Trees.* London: Verso, 2005.

Ng, Andrew Y, and Michael I. Jordan. "On Discriminative Versus Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes." 2002. Stanford AI Lab.

Rokach, Lior. "Ensemble-Based Classifiers." *Artificial Intelligence Review.* 33 (2010): 1-39.

Santini, Marina, and Mark Rosso. "Testing a Genre-Enabled Application: A Preliminary Assessment." *BCS-IRSG.* 2008.

Schneider, Karl-Michael. "Techniques for Improving the Performance of Naive Bayes for Text Classification." *Proceedings of CICLing.* 2005.

Sculley, D., and Bradley M. Pasanek. "Meaning and Mining: The Impact of Implicit Assumptions in Data Mining for the Humanities." *Literary and Linguistic Computing* 23 (2008): 409-24.

Tudor, Andrew. *Theories of Film.* New York: Viking, 1974.

Underwood, Ted, and Jordan Sellers. "The Emergence of Literary Diction." *Journal of Digital Humanities* 1.2 (June 2012). http://journalofdigitalhumanities.org/1-2/the-emergence-of-literary-diction-by-ted-underwood-and-jordan-sellers/

Unsworth, John, and Tanya Clement, Sara Steger, and Kirsten Uszkalo. "How Not to Read a Million Books." http://people.lis.illinois.edu/~unsworth/hownot2read.html

Witmore, Michael. "The Time Problem: Rigid Classifiers, Classifier Postmarks." *Wine-Dark Sea.* April 16, 2012. http://winedarksea.org/?p=1507 - comments

Wordsworth, William. *Lyrical Ballads, with Other Poems.* 2 vols. London: Longman, 1800.
Yu, Bei. "An Evaluation of Text Classification Methods for Literary Study." *Literary and Linguistic Computing.* 23 (2008): 327-343.

NATIONAL ENDOWMENT FOR THE HUMANITIES

DH start-up proposal

Applicant Institution: University of Illinois, Urbana
Project Director: Ted Underwood
Project Grant Period: 5/1/2013-11/1/2014

	Computational Details/Notes	(notes)	Year 1	(notes)	Year 2	(notes)	Year 3	Project Total
			5/1/2013-5/1/2014		5/1/2014-11/1/2014			
1. Salaries & Wages								
Project Director Ted Underwood	1 course/semester for two semesters at $10,000/course	50%	$20,000		$0			**$20,000**
Undergraduate research	hourly		$6,000		$0			**$6,000**
2. Fringe Benefits								
Project Director Ted Underwood	44.67% of salary		$8,934		$0			**$8,934**
Undergraduste benefits	0.14% of hourly wages		$8					**$8**
3. Consultant Fees								
4. Travel								
Project Director Ted Underwood	one visit to the NEH		$1,100		$0			**$1,100**
5. Supplies & Materials								
	0							**$0**
6. Services								
								$0
7. Other Costs								
								$0

	Computational Details/Notes	(notes)	Year 1	(notes)	Year 2	(notes)	Year 3	Project Total
8. Total Direct Costs	**Per Year**		**$36,042**		**$0**		**$0**	**$36,042**
9. Total Indirect Costs	**Per Year**		$21,121		$0		$0	**$21,121**
Indirect Cost Calculation: a. Rate: 58.6% of direct cost per year. b. Federal Agency: Office of Naval Research c. Date: 03/01/2012								
10. Total Project Costs (Direct and Indirect costs for entire project)								**$57,163**
11 Project Funding								
a. Requested from NEH	Outright:							**$57,163**
	Matching Funds:							
	Total Requested from NEH:							**$57,163**
b. Cost Sharing	Applicant's Contributions:							**$0**
	Third Party Contributions:							**$0**
	Project Income:							**$0**
	Other Federal Agencies:							**$0**
	Total Cost Share:							**$0**
12. Total Project Funding								**$57,163**

UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

Office of the Vice Chancellor for Research

Fourth Floor Swanlund Building
601 East John Street
Champaign, IL 61820-5711

September 21, 2012

Professor William E. Underwood
Department of English
University of Illinois at Urbana-Champaign

Dear Ted:

I am very pleased to see that you are submitting the start-up grant proposal, "Understanding Genre in a Collection of a Million Volumes," to the National Endowment for the Humanities. Illinois is one of the country's pre-eminent centers for digital humanities scholarship, and will provide a highly supportive environment for this project.

An important component of this support will be the computing and storage resources needed, first for cleaning and organizing your collection of volumes, and then for classifying the volumes by genre. I understand that your initial work will be conducted using a portion of the University of Illinois campus cluster https://campuscluster.illinois.edu managed by M. Scott Poole, the Director of Illinois Computing for the Humanities, Arts and Social Sciences (I-CHASS). More intensive processing will make use of an NSF XSEDE allocation of 30,000 on the Blacklight supercomputer, in consultation with the National Center for Supercomputing Applications.

In order to extend your analysis beyond 1923, to a period when many volumes are still covered by copyright, I am glad to see that you will be working with the HathiTrust Research Center here at Illinois. The HathiTrust Research Center (HTRC) permits non-consumptive research, which will allow you to utilize the content of the English-language volumes from 1924 to 1949 contained in the HathiTrust Library. As many of these volumes are not yet in the public domain, this will enable you to pursue your research while preventing intellectual property misuse within the confines of current U.S. copyright law.

The University of Illinois is uniquely well qualified to be the site of this project, thanks to the domain expertise of its faculty and staff, its digital library resources, and its excellence in computing, database and information technology. Your project promises to make a significant contribution to the digital humanities, and we look forward to supporting your efforts.

Sincerely yours,

Peter Schiffer
Vice Chancellor for Research

telephone 217-333-0034 • *fax* 217-244-3716

Initiative for Digital Humanities, Media, and Culture

Serving the Colleges of Architecture, Education,
Engineering, and Liberal Arts

TEXAS A&M UNIVERSITY

Laura Mandell
TAMU 4227
mandell@tamu.edu
Ph. 979.845.8345
Fax 979.826.2292

September 24, 2012

National Endowment for the Humanities
Office of Digital Humanities
Digital Humanities Startup Grants

Dear NEH Office of Digital Humanities:

I write to recommend that you award a large DH Startup Grant to Ted Underwood for his project, "Understanding Genre in a Collection of a Million Volumes." This project is eminently fund-worthy for multiple reasons that I will now enumerate:

1) In addition to being one of the star scholars in the field of Romantic-era studies in English, Ted Underwood is experienced at topic and data-mining: both his article in the prominent Journal of Digital Humanities and his presentation at the DH2012 Conference in Hamburg, Germany, reveal his capacity to mine big data for unexpected results. Listening to his talk in Germany, I realized that, by tweaking the genre measures he used to make them work, Professor Underwood had uncovered an essential truth about Romantic poetry that I have seen nowhere else, viz., that it was BECAUSE and not IN SPITE OF using plain language (the "language of men speaking to men") that Romantic poetry became more specialized, its linguistic characteristics separating it MORE dramatically from prose than it had been before.
2) As director of ARC, the Advanced Research Consortium, that supports NINES, 18thConnect, REKn, MESA, and ModNets, I can say that we need automated genre detection in the worst way. Genre is a feature of our metadata, but we are under pressure to dispense with it as a metadata category for those contributing to our online communities and search engines: managers of unruly datasets rarely can detect genres of the metadata and texts at their disposal. That is why, if you go to NINES, so many items are labled "bibliographic citation": we cannot determine genre for the resources beyond that fact, that we have a metadata item (even if there is text attached). We would use and develop any genre-determining mechanism, and we resist abandoning genre as a category because of the value of searching for things in that manner, not only to literary scholars, but to lawyers, for instance, who might like to know when eighteenth-century documents are employing legal diction in legal cases or in fiction.
3) Less selfishly on my part, efforts to detect genre besides Dr. Underwood's have had mixed results, as evinced by the paper from the LitLab at Stanford co-authored by Matthew Jockers and Michael Witmore. We need many many experiments in order to build this crucial knowledge.

227 John R. Blocker Building
4227 TAMU
College Station, TX 778434227

Tel. 979.845.3451 Fax 979.862.2292
www-english.tamu.edu

In short, you cannot go wrong in funding this project: Ted Underwood and his collaborators will certainly create something worth using, knowing about, and building upon.

Sincerely,

Laura Mandell
Professor, English
Director, Initiative for Digital Humanities, Media, and Culture (IDHMC)

Maryland Institute for
Technology in the Humanities
0301 Hornbake Library
College Park, MD 20742

Travis Brown
Assistant Director
of Research and Development

September 19, 2012

To: Office of Digital Humanities, NEH
Re: Level II Digital Humanities Start-Up Grant Proposal
Understanding Genre in a Collection of a Million Volumes

I am writing to recommend Ted Underwood's "Understanding Genre in a Collection of a Million Volumes" proposal. In my work at MITH I've often seen the need for the kind of resource it describes, and the project itself promises to have a substantial additional value as a model for future large-scale metadata-enrichment projects.

Many of MITH's recent text analysis projects have focused on the development and application of *unsupervised* machine learning methods, such as Latent Dirichlet Allocation topic modeling. Unsupervised methods have a particular value for humanities research at the current moment, since many of the large text collections we work with are lacking in accurate and detailed annotations or metadata. Topic modeling allows us to identify thematic patterns in a collection of thousands of eighteenth-century texts from the HathiTrust Digital Library, for example, even if we do not have reliable information about the genre — or even the language, in many cases — of those texts.

While these unsupervised methods can provide an extremely useful way to explore large text collections with imperfect metadata at a high level, their limitations quickly become apparent when they are applied to many specific kinds of research question — to trace lines of influence from the Gothic novel to science fiction, for example. Having access to reliable metadata about genre on the scale this proposal describes would allow us to create more sophisticated models, to train our existing models in more targeted ways, and to ask new questions about the output of our systems.

Travis Brown

Email: trbrown@umd.edu • *url* http://mith.umd.edu• *phone* (301) 405-8927

UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

Graduate School of Library and Information Science
Library and Information Science Building
501 East Daniel Street
Champaign, IL 61820-6211

September 19, 2012

To Whom It May Concern,

I am pleased to support Ted Underwood's research project "Understanding Genre in a Collection of a Million Volumes." Professor Underwood's project represents an exciting and high-impact collaboration for the HathiTrust Research Center, testing a range of capacities that we have already developed or are currently developing.

HathiTrust Research Center allows researchers to query public-domain portions of the HathiTrust collection and export word counts, as required by this proposal. Underwood's project will also require initial data cleaning before word counts are exported, and we are currently developing capacities that support a variety of approaches to data cleaning, permitting researchers to contribute their own Python or Java modules to the HTRC workflow.

By the time Underwood's project would need our fuller technical support (late 2013), we plan to have these capacities available for non-consumptive use on post-1923 data. If unforeseen legal obstacles make it impossible to export word counts from volumes on the HathiTrust's post-1923 data, Professor Underwood's project would still usefully dovetail with HTRC, since the HTRC strives to be the primary repository for metadata and algorithms developed using the HathiTrust texts.

In the case of Dr. Underwood's project, this would involve hosting collections of data and metadata linked to HathiTrust volume IDs, along with algorithms that can be used to normalize, segment, or classify HathiTrust volumes. We plan to incorporate this project's tools and outputs with an ever-growing collection of such resources created by leading scholars and researchers over the very long term. In this way, Professor Underwood's cutting-edge research will influence the widest audience of scholars over the longest period of time.

Sincerely,

Professor J. Stephen Downie
Co-Director, HathiTrust Research Center

telephone 217-333-3280 • *fax* 217-244-3302 • *e-mail* gslis@illinois.edu • *url* http://www.lis.illinois.edu

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

DEPARTMENT OF ENGLISH
Curtis Perry, Head

608 South Wright Street
Urbana, Illinois 61801
phone: *217-333-2390*
fax: *217-333-4321*
email: *cperry@illinois.edu*

September 12, 2012

To whom it may concern,

I am writing, in my capacity as Head of the Department of English at UIUC, to express my support for my colleague Ted Underwood's Digital Humanities Start-Up Grant ("Understanding Genre in a Collection of a Million Volumes"). I have read the proposal and its budget and find both to be feasible.

One budget item includes money to buy Professor Underwood out of teaching two of the classes he would otherwise be scheduled for during the 2013-14 academic year. This teaching release is of course essential to the proposed project, freeing up time to conduct the work described. I am writing here specifically to offer my assurances that if the proposal as budgeted is funded we will indeed allow Underwood to buy out these classes, thereby releasing him from classroom obligations in order to conduct the research program decribed in his proposal.

Please do not hesitate to contact me if you have any further questions.

Sincerely,

Curtis Perry
Professor and Dept. Head