

The Labors of the Novel: A Digital Database of Vocations in German Narrative Fiction, 1750-1950

A. Enhancing the Humanities Through the Use of Emerging Technologies:

i. Project Description:

The goal of this project is to create a digital database of roughly 6,000 German novels published between 1750 and 1950 and to tag them with keywords that identify the protagonist's vocation and (for many entries) the geographical location of the principal setting.

These keywords will be extracted from existing bibliographic metadata (*i.e.*, descriptive data about books) in both print and electronic forms. The sources of such metadata include library catalog subject headings and keywords, detailed finding aids written for rare books collections, specialized reference works and existing databases.

The emphasis of the project is on the collection of quantitative data. Users will interact with the resulting database *via* an interface that will allow them to track the development of individual vocations in German fiction both chronologically and geographically. Researchers interested in the history of the industrial revolution in Germany, for instance, will be able to track, on a year-by-year basis, how many novels were published about factory workers, where these novels were set, and where they were published. (See the Appendix for mock-up screenshots). Researchers will further be able to combine individual vocations into larger categories (for instance "painter," "musician," and "poet" into the larger category of "artist") and to plot vocational trends against one another (*e.g.*, "shopkeeper" versus "factory worker").

The political, cultural and linguistic history of the German-speaking regions is complex. Our project solves this problem by aiming for maximum inclusivity. The database will thus contain works written in dialect and regional language varieties such as *Plattdeutsch* (Low German) or *Schwyzerdütsch* (Swiss German). Similarly, we will include German works published by foreign presses as well as those whose principal setting takes place outside of the German-speaking countries.

The metadata used in this project will derive from a number of sources and institutions. In each case, the researchers will obtain prior copyright clearance and will communicate with participating institutions about acceptable procedures for data acquisition. In addition, search results in the completed database will include listings of the various bibliographic datasets that were used to create these results, complete with hyperlinks to the original sources.

ii. Project Rationale:

The German novel is traditionally said to have followed a very different trajectory from that of its Western European neighbors. Largely because of the political fragmentation of the German principalities prior to national unification in 1871, the German novel ostensibly focused on the inner life of its protagonists instead of cultivating a broad social realism of the kind that emerged in Britain and in France during the mid-nineteenth century. This inwardness, in turn, supposedly discouraged German readers from imagining themselves as social actors in the same way that readers of Dickens or Balzac did and thereby contributed to a culture of political passivity that culminated with the rise of Nazism.

This traditional account first came under attack during the 1980s, when authors such as Jeffrey Sammons and Peter Uwe Hohendahl pointed out that the notion of an inwardly-oriented German novel is the result

of critical occlusions made during the Wilhelmine period, often for overtly political reasons. The *real* German novel, so it was claimed, was actually much more varied than the canon would seem to suggest. These critical interventions were never followed up in systematic fashion, because the critical climate during the 1980s and 1990s was hostile to the broad sociological treatment that would have been necessary to establish a more accurate picture of the literary field.

An intelligent evaluation of vocational metadata, carried out with digital tools and with modern data visualization techniques, could thus yield valuable insights into the true nature of German novel production and how it interacted with concurrent social developments. The questions that might reasonably be asked of such survey data would include:

- Were industrial occupations over- or underrepresented when compared to artisanal or agrarian vocations?
- How long did it take for the industrial revolution to emerge as a dominant topic in German novels?
- Did novels about artists and other “spiritual” professions really constitute a large subset of German novels?

In addition to vocational data, the database will also encompass entries on the principal settings of a smaller subset of novels. These entries will give researchers new tools to pursue queries such as:

- What effect did political fragmentation have on the German novel?
- Were the majority of nineteenth-century German novels set in urban, provincial or rural settings? How do the settings of novels correlate with demographic trends?
- Did publishing houses favor novels with local settings?

Throughout, the goal of the project will be to give researchers concrete statistical tools that will allow them to formulate hypotheses about how the German novel relates to social trends in a more precise fashion than is possible with detailed textual analysis alone.

To facilitate this goal, innovative data visualization will be an important component of this project. There is a growing number of visualization tools for bibliographic metadata, both commercial and open source, but many of these were designed to reveal the relationships between papers (citation networks), or to enable a user to explore reference collections based on content and more easily find the most relevant citations for papers. By contrast, this project centers on the identification of chronological trends in the vocations of literary protagonists and the relationships between these trends and other bibliographic metadata. While standard graphing tools, such as Microsoft Excel, could be used to plot and analyze some of the data in this project, developing a customized interactive visualization tool that can be incorporated directly into a web interface is extremely important to facilitate flexible exploration and discovery. This tool will allow researchers to query the database quickly and easily, visualize trends, and investigate new questions and ideas inspired by what they see. One example of interactive visualization of somewhat similar data is Job Voyager (http://flare.prefuse.org/apps/job_voyager), a stacked time series that shows the distribution of different occupations in the United States labor force between 1850 and 2000. Users can study trends in all of the occupations at once, sort by gender to compare trends in males and females, or focus on a single occupation.

B. Environmental Scan:

The Labors of the Novel contributes to what Stanford literary scholar Franco Moretti has influentially called “distant reading” approaches to literature, *i.e.*, approaches that privilege quantitative surveys and descriptive accounts over the in-depth analysis of individual works associated with traditional literary studies (“close reading”). The phrase “distant reading” was coined by Moretti in the article “Conjectures on World Literature” (*New Left Review* 1, Jan-Feb 2000), but found its most influential expression in his later book *Graphs, Maps, Trees: Abstract Models for Literary History* (London: Verso, 2005).

While Moretti’s theories have generated considerable discussion among literary scholars (for a representative summary with many links to external sources, see the archived discussion at http://www.thevalve.org/go/valve/archive_asc/C48), its practical impact, measured by the number of projects that it has triggered, has been much smaller, especially within the field of German Studies.

The existing “distant reading” projects can usefully be divided into two categories. The first category encompasses all those projects that attempt to “read” literary texts themselves, *i.e.*, to use raw computing power to analyze vast amounts of text files and subject them either to keyword searches or to various kinds of statistical analysis. These approaches are in the majority and have frequently received both methodological inspiration and financial backing from the search-engine company Google. An especially useful tool in this context is the *Google Ngram Viewer* (<http://ngrams.googlelabs.com/>), which allows limited keyword searches on a vast corpus of books. Current projects that from within German Studies that utilize quantitative approaches include Paul Youngman’s “*Black Devil and Iron Angel Revisited: N-Gramming the Railway in 19th Century German Fiction*” at UNC Charlotte, and Allen Beye Riddell’s “Quantitative Analysis of Nineteenth-Century German Studies, 1928-2006” at Duke University.

Projects from within the first category that rely on statistical analysis of literary texts frequently make use of the DocuScope software developed at Carnegie Mellon University (http://www.betterwriting.net/projects/fed01/dsc_fed01.html). A practical example of such an endeavor would be Matthew L. Jocker’s attempt at “Machine-Classifying Novels and Plays by Genre” at Stanford (<http://www.stanford.edu/~mjockers/cgi-bin/drupal/node/27>).

The Labors of the Novel belongs to a second category of “distant reading” projects, which focus their analysis on metadata rather than on the texts themselves. This second category is much smaller, in part because metadata first needs to be painstakingly generated by subject experts. Precisely for this reason, however, our project will hopefully have a signal function within the field of digital humanities: in an age in which scholarly searches are increasingly conducted with brute computing power, through tools such as *Google Books*, it demonstrates the enduring significance of traditional bibliographic expertise and charts out a way in which “old” and “new” humanities might cooperate with one another.

A project that directly inspired our own is Moretti’s inquiry into the length and syntactical features of 7,000 English novel titles, published as “Style Inc.” in *Critical Inquiry* 36 (Autumn 2009). While this attempt at quantitative metadata analysis has many similarities to ours, there is an important difference: book titles function as both data and metadata, and are both part of the literary work itself and part of its description in the library catalog. They thus need to be treated with considerable suspicion, for, unlike other metadata, they owe as much to poetic inspiration as they do to analytic neutrality. Significantly, Katie Trumpener, an expert in German literature, has charged Moretti with underplaying linguistic and national peculiarities of literary naming conventions, and thus producing an extremely biased account of his corpus (see her “Paratext and Genre System: A Reply to Franco Moretti” in the same issue).

[History and Work Plan Omitted]

C. Staff (see section 6 for biographical details and section 5 for time allocation)

Tobias Boes, Assistant Professor of German, University of Notre Dame; Dr. Robert L. Kusmer, Associate Librarian, University of Notre Dame; Dr. Kristina Davis, Prof. Dilkushi Pitts and Prof. Christopher Sweet, Center for Research Computing, University of Notre Dame.

D. Final Product and Dissemination

The primary investigator, Prof. Tobias Boes, will give a preliminary presentation on the *Labors of the Novel* project at the German Studies Association conference in Louisville, KY, in September 2011. Initial results will be presented at a symposium at Washington University in April 2012 and will then be published in a forthcoming volume on digital humanities approaches to German literature. A final presentation likely can be scheduled for an international symposium on German studies and the digital humanities at the University of Notre Dame that would be sponsored through bilateral assistance from the DFG and NEH in the spring of 2013. An alternate solution would be to hold this presentation at the German Studies Association conference in the fall of 2013.

The presentations are intended primarily as a means to attract scholarly attention to the project. The outcome of this endeavor will be measured by monitoring web traffic to the project server (www.novelvocations.crc.nd.edu), which is currently accessible only to project participants but will open to the general public by February 2012. Research of the kind facilitated by this database has the potential not only to alter our larger understanding of the field of German literature, but can also have an impact on the way scholars approach individual well-known works. For instance, Thomas Mann's famous novel *Buddenbrooks* has long been recognized as owing a debt to Scandinavian and Russian novels of social realism. Much less known, however, is that *Buddenbrooks* can also be placed into a longer tradition of novels featuring merchant protagonists and set in the port towns of the former Hanseatic League.

The software infrastructure created to support the database will also prove useful to the scholarly community. In particular, the GIS components, which can layer a wide variety of data sets onto a historically flexible map of the German-speaking countries, should be adaptable to a variety of other projects. The CRC will publicize information about this software via the project website and track its adoption by other researchers.

A secondary metric by which to judge the success of the project will be the impact it has on undergraduate education, both at Notre Dame and beyond. At Notre Dame, this impact will be measured by the effect it has on undergraduate research, particularly at the senior thesis level. To encourage adoption of our project in classrooms beyond Notre Dame, the project participants plan to describe the database and illustrate some of its pedagogical uses in journals such as *Profession*, *Digital Humanities Quarterly*, or *Modern Language Journal*. The final web portal for the project will also include supplementary information about potential uses in undergraduate teaching.

One clear pedagogic benefit of this project is that it sheds light on an area of German culture that is neglected in most American classrooms: the novel prior to 1945. While narrative fiction has long been recognized as an important component of national identity formation, the length of most eighteenth- and nineteenth-century novels, together with the unavailability of English translations or even editions in Roman typeface, have effectively precluded their formal study at the undergraduate level. While the proposed database will never substitute for close textual analysis, it will give both literary scholars and historians a valuable tool to shed light on what would otherwise remain an unapproachable topic.