CHARACTERIZING DESCRIPTIVITY IN WRITING THROUGH TEXT ANALYSIS OF BOOKS FROM THE HATHITRUST DIGITAL LIBRARY

Allocation: Exploratory/50 Knh
PI: J. Stephen Downie[1]
Co-PIs: Sayan Bhattacharyya[2], José Eduardo González[1]
Collaborators: Boris Capitanu[1], Craig Willis[1], Peter Organisciak[3]

[1]University of Illinois at Urbana–Champaign
[2]Singapore University of Technology and Design
[3]University of Denver

EXECUTIVE SUMMARY

This project approaches quantifying the notion of descriptivity in text. The immediate objective was to explore descriptivity in forms of writing that have been characterized as exemplifying different writing styles and genres. Digital text analysis offers an opportunity to operationalize the anecdotal notion of descriptivity by developing quantified metrics for descriptivity. This work leveraged the resource represented by the HathiTrust Digital Library repository. The requested allocation was used to create an updated data set of preprocessed, extracted features from the HathiTrust corpus with exploratory methods to support the research. These extracted "features" were quantifiable facts about the pages of the books, most usefully counts of words (unigrams) or strings of words (bigrams).

RESEARCH CHALLENGE

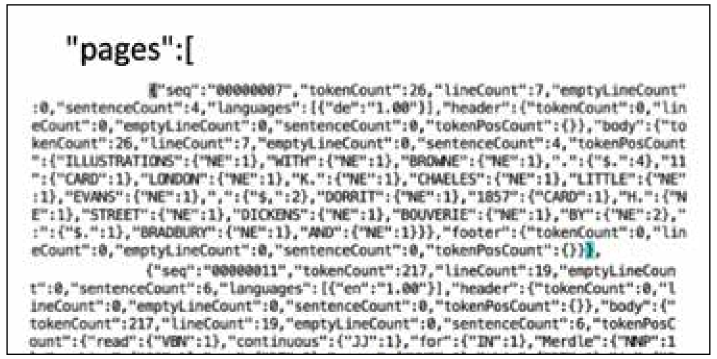
There is a belief that writing has, since World War II, taken an overall turn away from "tell" toward "show"; this suggests that writers are becoming increasingly more interested in description [1]. However, estimating descriptivity is difficult. This is useful not only for literary and historical studies (the research team's immediate focus of interest) but also for researchers from beyond these areas. Characterizing books or pages in books by such a metric in the form of user-generated metadata is helpful for the use of a text corpus resource because users of the resource may want to find more descriptive or less descriptive books, depending on their needs, especially since treating text as data opens up new experiences of reading [2].

The notion of description has been the topic of a longstanding discussion in literary and historical studies. In his influential 1936 essay "Narrate or Describe?" György Lukács distinguished, in the case of fiction, between a dynamic concept of "narration," in which verbal description of material objects is intertwined with the progression of the protagonists' character development through action, and mere "description," in which description is static and isolated from the flow of action [3]. Such a distinction becomes easy to operationalize in terms of the research team's metric. Theorists from literary studies and historiography, such as Hayden White, have, likewise, associated description without narrativity with lack of meaning [4]—a notion that could similarly be operationalized to engage with ongoing debates in the humanities [5,6] and in cultural analytics [7].

In particular, bigrams are useful in several ways. In certain circumstances (such as when queries can be explicitly identified as good candidates for bigram use), greater improvements in information retrieval tasks are obtained from using bigrams as queries rather than other queries [8]. The inclusion of bigrams provides consistent gains in sentiment analysis tasks [9].

METHODS & CODES

The research team developed code to compute a simple proxy metric for "descriptiveness" with part-of-speech-tagged unigrams (tagged with Penn Treebank part of speech categories). The script counts the total count of two "per-page co-occurring" parts of speech for any volume. When using HathiTrust Research Center unigrams, "per-page co-occurring" simply means min (x,y), where x is the number of occurrences on a page of the part-of-speech-tag X, and y is the number of occurrences on a page of the part-of-speech-tag Y. This estimation using unigrams, however, typically produces an overcount, as the count generated is not restricted only to contiguous tokens among the equal number of paired tokens but also includes tokens from within the page that are not contiguously paired; this can be rectified by the use of bigrams.



```
"pages":[
        {"seq":"00000007","tokenCount":26,"lineCount":7,"emptyLineCount"
:0,"sentenceCount":4,"languages":[{"de":"1.00"}],"header":{"tokenCount":0,"lin
eCount":0,"emptyLineCount":0,"sentenceCount":0,"tokenPosCount":{}},"body":{"to
kenCount":26,"lineCount":7,"emptyLineCount":0,"sentenceCount":4,"tokenPosCount
":{"ILLUSTRATIONS":{"NE":1},"WITH":{"NE":1},"BROWNE":{"NE":1},".":{"$.":4},"11
":{"CARD":1},"LONDON":{"NE":1},"K.":{"NE":1},"CHAELES":{"NE":1},"LITTLE":{"NE"
:1},"EVANS":{"NE":1},",":{"$,":2},"DORRIT":{"NE":1},"1857":{"CARD":1},"H.":{"N
E":1},"STREET":{"NE":1},"DICKENS":{"NE":1},"BOUVERIE":{"NE":1},"BY":{"NE":2},"
:":{"$.":1},"BRADBURY":{"NE":1},"AND":{"NE":1}}},"footer":{"tokenCount":0,"lin
eCount":0,"emptyLineCount":0,"sentenceCount":0,"tokenPosCount":{}}},
        {"seq":"00000011","tokenCount":217,"lineCount":19,"emptyLineCoun
t":0,"sentenceCount":6,"languages":[{"en":"1.00"}],"header":{"tokenCount":0,"l
ineCount":0,"emptyLineCount":0,"sentenceCount":0,"tokenPosCount":{}},"body":{"
tokenCount":217,"lineCount":19,"emptyLineCount":0,"sentenceCount":6,"tokenPosC
ount":{"read":{"VBN":1},"continuous":{"JJ":1},"for":{"IN":1},"Merdle":{"NNP":1
```

Figure 1: "Data" as bags of words (developer's view). The content of digitized text made available after crunching in the form of unigrams or bigrams is accessible to a developer as data. Each page is a separate bag of words, available for processing.

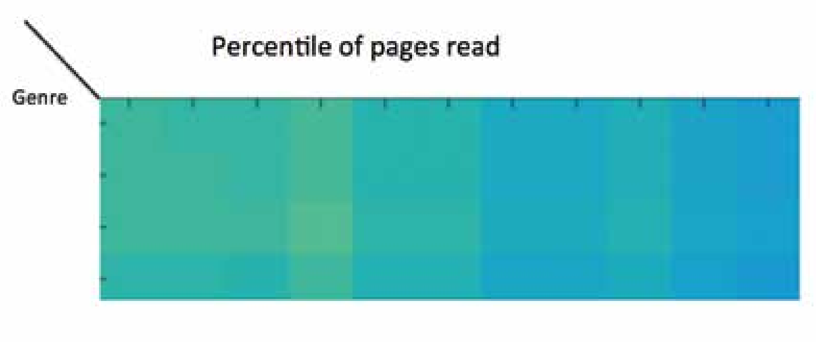


Figure 2: How to conceptually visualize "descriptiveness." Descriptiveness scores of the kind the research team explores, in different genres (such as fiction, drama, etc.) and across "narrative time" (the relative position within the work, measured in percentile, of a page whose descriptiveness is being considered) can be visualized to get an intuitive comparative sense of how descriptiveness tends to wax and wane in collections of works drawn from different genres.

RESULTS & IMPACT

The useful impact of this work is the demonstration that as more computational power is becoming available, more complex and finer-grained analysis (such as that using bigrams) can be undertaken to approach the problem at progressively deeper levels. The research team has used this continuing work in teaching an undergraduate independent/directed study offered through the English Department of the University of Pennsylvania in 2018 and in a digital humanities course in Singapore in September 2019.

WHY BLUE WATERS

The sheer size of the data set and the embarrassingly parallel computational nature of generating the needed features made Blue Waters an ideal environment for conducting this exploratory analysis. Using advanced features such as bigrams allowed for a better estimation of co-occurrence. This has allowed the team to use adjectives and nouns as well as adverbs and verbs in the form of actual pairs in the per-page-co-occurring parts of speech used in computing the metric to measure descriptiveness. For extracting these advanced features, the team created 100 partitions of approximately equal size from the 5.4-million input volumes of text and ran 100 jobs, with each job using 100 nodes and each node processing the text from twenty books (volumes) simultaneously. The average number of volumes per partition was 54,200, and the total number of nodes used was 10,000 (100 x 100). The total number of cores used was 200,000 (100 x 2,000), and the average time taken to process one partition was one hour, 45 minutes. The output produced was 6.25 terabytes (TB), consisting of 55 gigabytes of entities, 2.11 TB of bigrams, and 4.08 TB of trigrams.

PRESENTATIONS & DATA SETS

S. Bhattacharyya, "Textual digital humanities for critique, with Lukács," presented at Lukács and the World: Rethinking Global Circuits of Cultural Production, Santa Barbara, CA, U.S.A., Apr. 20–21, 2018.