Modeling the History of Book Design
HTRC Whitepaper: Summary of Activities

David Bamman
School of Information
UC Berkeley
dbamman@berkeley.edu

Björn Hartmann
EECS
UC Berkeley
bjoern@eecs.berkeley.edu

1 Introduction

Much work in the large-scale cultural analysis of books — for questions of genre (Underwood, 2016), character (Underwood et al., 2018), emotion (Heuser et al., 2016), loudness (Katsma, 2014), geographic attention (Wilkens, 2013) and much more — has tended to reason only about the strings of words those books contain. Books of course have a strong visual component as well, with a rich design tradition informing not only their overall structure but also the typographical layout of physical pages.

In this work, we ask the following question: does the visual structure of the printed page change over time, and if so, along which dimensions? Figure 1 illustrates this with five editions of Darwin's *Origin of Species*; these editions are published in 1859, 1904, 1936, 1993, and 2008. Even among these five versions we can see substantial variation in the density of the text and the size of the page, the relative size of the margins. Are there systematic ways in which these typographic elements have changed over the period of publication?

2 Experimental design

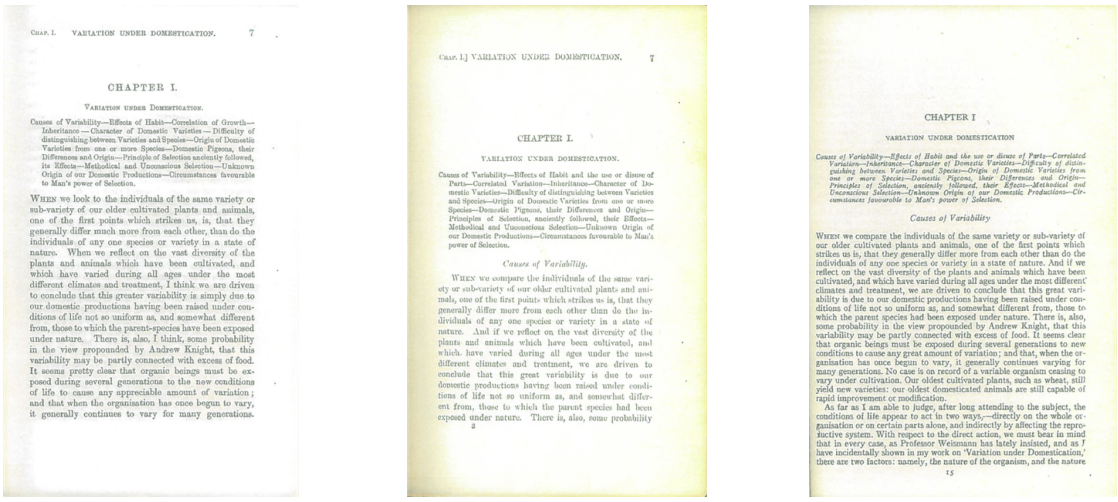
One possibility for exploring design changes would simply involve measuring typographic elements in a large random sample of books and assessing the differences observed between books over time. However, any measurement of typographic quantities would be confounded in such a sample by the many other changes in the history of book publication—such as the rise of mass market paperbacks (with smaller overall dimensions and font sizes and thin margins).

To control for all of the other variables that may have changed over the course of book history, we design an experiment to investigate the changes taking place within a fixed set of books. To borrow terminology from the FRBR hierarchy (IFLA, 1998) the five versions of Darwin's *Origin of Species* in table 1 are each an *expression* (a different typeset edition) of a single *work* (*The Origin of Species*). For this research, we select 4,919 works and two expressions of each—one published in 1928 or before, and one published in 1929 and after. 1928 marks the rough midpoint of our collection (1850–2011) and coincides with the year of publication of *The New Typography* by Tschichold (1928), which advocated for a radical break from the static, symmetric, centered design that came before it.

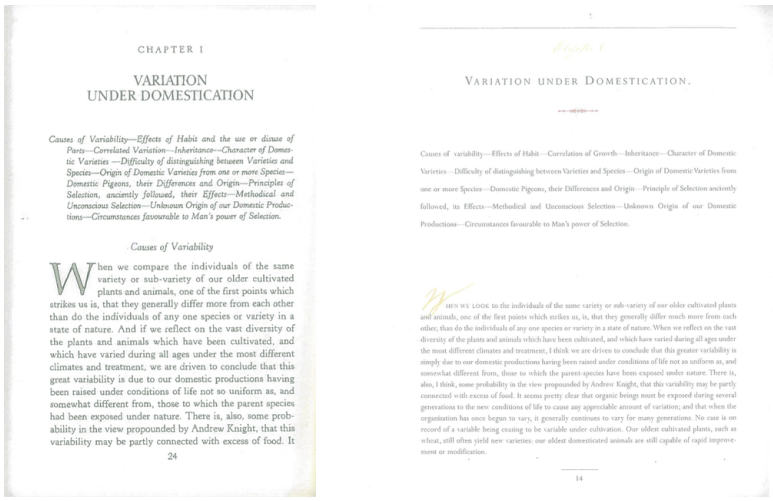
3 Data

The first activity carried out in this collection was identifying a set of works in the HathiTrust that have two distinct expressions (one published in 1928 or before and one in 1929 or after). Importantly, these expressions also need to be distinct; many versions of the same work in the HathiTrust are different FRBR *manifestations* of the same expression (e.g., different scans of the same edition in different libraries).

To identify these candidate texts, we consider the top 4,000 authors by frequency of publication in the HathiTrust (i.e., those who are likely to have multiple copies of their work). For each author, we use the HathiTrust ExtractedFeatures data (which provides unigram counts by page for all words in a text) to identify all pairs of pages from two of



(a) Facsimile of original printing (1859) (b) Darwin (1904) (c) Darwin (1936)



(d) Darwin (1993) (e) Darwin (2008)

Figure 1: Five versions of Darwin's *Origin of Species*, illustrating different design choice in page layout.

their books that match over some threshold. Each book must be published after 1850 and cannot have the term *poem, poet, verse, play* or *dictionary* in the titles (since we are targeting traditional formats).

Since we are specifically looking for pages of the same content in *different* editions, it's likely that only one half of page i will match page j (with the remaining content of page i appearing on page $j-1$ or $j+1$). Rather than defining a match strictly on cosine similarity, we identify the highest scoring page j in book B for each page i in book A; $\text{BEST}(i) = j$. We define a match between i and j if $\forall n \in [-10, \ldots, 10]$, $\text{BEST}(i+n) = j+n$, for at least 70% of n. This yields a large set of pairwise page matches between book A and B.

We attempt to exclude books that are the same editions, and do so again by exploiting the similarity of the page *sequences* between the two books. For candidate pair i and j, we calculate the cosine similarity between $i+n$ and $j+n$ for $n \in [-2, -1, 0, 1, 2]$. The cosine similarity between identical books will be the vector $[1, 1, 1, 1, 1]$, so we calculate the euclidean distance between the observed cosine similarity and that ideal sequence. All pairs that are judged to be too similar (with a distance < 0.10) are excluded.

This process yields a total of 4,919 works (9,838 editions); for analysis, we sample 5 pairs of pages from each work (drawing roughly the same page from each expression).

4 Measurements

We investigate ten different typographic quantities, outlined below. All of the metrics described below rely on estimates of bounding boxes for the page, line and individual words. We create these bounding boxes by running Google's Tesseract OCR software (version 4.0) on each page image. While the HathiTrust provides access to bounding box information in the form of hOCR files, they are generated by different OCR systems for different subsets of the collection; in order to eliminate any confounds that arises as a function of different OCR systems being used for public domain texts published before 1923 and in-copyright texts published afterwards, we judged it necessary to run the same OCR system over all texts to remove that dimension of variation.

We estimate the following quantities for each page in a book:

Page height. We measure page height by the number of pixels between the top and bottom page coordinates, divided by the scanning DPI (to yield a quantity in inches).

Page width. We measure page width by the number of pixels between the left and right page coordinates, divided by the scanning DPI (to yield a quantity in inches).

Median line height. We measure the line height by the number of pixels between the top and bottom line coordinates, divided by the scanning DPI and multiplied by 72 (to yield a quantity in points).

Median line leading. We measure the line leading by the number of pixels between the bottom of line i and the top of line $i + 1$, divided by the scanning DPI and multiplied by 72 (to yield a quantity in points).

Number of lines. We measure the number of lines as simply the count of the number of lines on a page.

Median line length. We measure line length by the number of pixels between the left and right line coordinates, divided by the scanning DPI (to yield a quantity in inches).

Median word spacing We measure the spacing between words as the number of pixels between the right coordinate of word i and the left coordinate of word $i + 1$, divided by the scanning DPI and multiplied by 72 (to yield a quantity in points).

Media number of characters per line We measure the number of lines as simply the count of the number of characters in a line.

Median number of words per line. We measure the number of lines as simply the count of the number of words in a line.

Right justification ratio. We measure whether a page is right justified by counting the ratio of lines whose right bounding coordinate differs from the median right coordinate for the page by 10 pixels; if fewer than 80% of lines differ by this amount, we judge it to be right justified.

For all measures, we create a summary statistic for an entire book as the median value observed in the page samples.

5 Tests

In order to calculate whether a meaningful difference exists between these quantities in the period that divides 1928 from 1929, we carry out a hypothesis test for each metric. Our design choice for this research involves a pair of typographic *expressions* for each *work*; one expression is published in 1928 or before and one in 1929 or after. Accordingly, a natural hypothesis test in this scenario is the paired t-test, which investigates differences between pairs of observations for the same entity; here, we treat each work as a data point and the metric calculated for each expression

as an observation for that point.

Table 1 presents the results of these tests.

Feature	Early	Late	Diff	p-value
right justification ratio	0.916	0.912	-0.004	
page height (inches)	7.817	8.038	0.221	***
page width (inches)	5.113	5.234	0.122	***
median line height (points)	10.082	10.341	0.258	***
median line leading (points)	1.780	1.621	-0.158	***
number of lines	39.480	40.258	0.778	**
median line length (inches)	3.381	3.540	0.159	***
median word spacing (points)	4.433	4.266	-0.166	***
median number of characters per line	43.568	45.556	1.988	***
median number of words per line	9.428	9.802	0.374	***

Table 1: Results. * = $p < 0.01$; ** = $p < 0.001$; *** = $p < 0.0001$

6 Analysis

We leave the analysis of these results to our own future work. Two trends, however, are immediately apparent: a.) the most significant trends appear to involve the change in book size between these periods. Books published in 1929 and after are on average about one quarter of an inch taller and one eighth of an inch wider than books published before. This increase in size leads to consequent differences such as an increase in the number of lines per page, width of each line, and the number of characters and words per line. b.) the font size (as determined by the median line height) also appears to be increasing (by a quarter of a point) over this time period, while compression is also increasing (the spacing between words has decreased, and the line leading has also decreased).

References

Charles Darwin. *The Origin of Species*. D. Appleton and Company, New York, 1904.

Charles Darwin. *Origin of Species*. Random House, New York, 1936.

Charles Darwin. *Origin of Species*. Random House, New York, 1993.

Charles Darwin. *Origin of Species*. Sterling, New York, 2008.

Ryan Heuser, Franco Moretti, and Erik Steiner. The emotions of london. Technical report, Stanford Literary Lab, 2016.

IFLA. *Functional Requirements for Bibliographic Records Final Report*. K.G. Saur Verlag, Munich, 1998.

Holst Katsma. Loudness in the novel. Technical report, Stanford Literary Lab, 2014.

Jan Tschichold. *Die neue Typographie*. Verlag des Bildungsverbandes der Deutschen Buchdrucker, Berlin, 1928.

Ted Underwood. The life cycles of genres. *Cultural Analytics*, 2016.

Ted Underwood, David Bamman, and Sabrina Lee. The transformation of gender in English-language fiction. *Cultural Analytics*, 2018.

Matthew Wilkens. The geographic imagination of Civil War-era American fiction. *American Literary History*, 25(4): 803–840, 2013.

Appendix: Distributions of differences

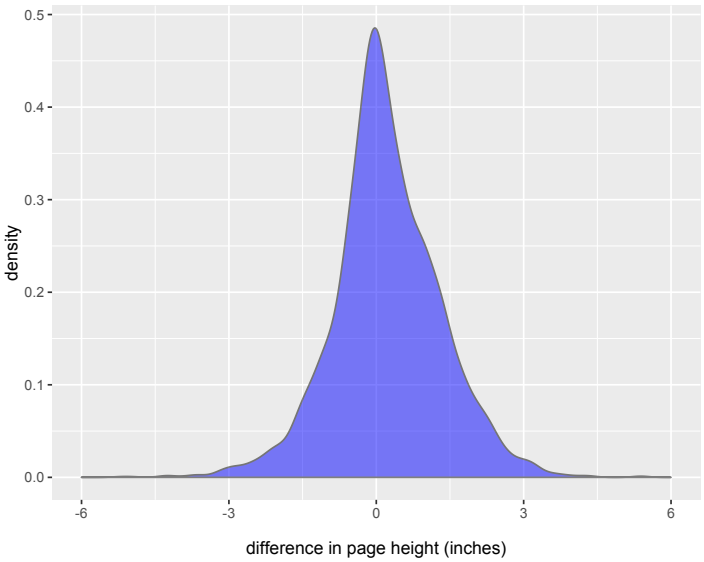


Figure 2: Distribution of differences (late-early) in median page height.

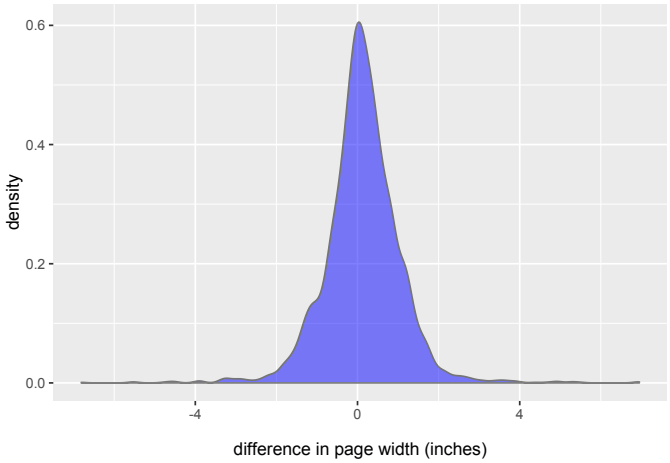


Figure 3: Distribution of differences (late-early) in median page width.

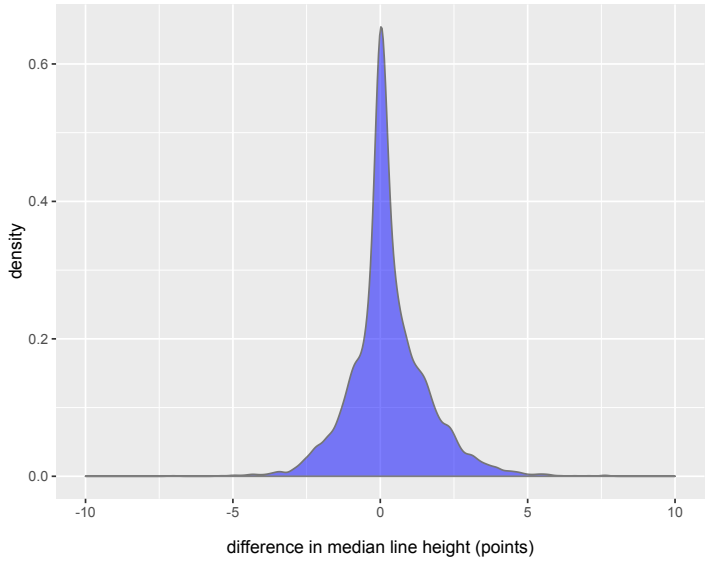


Figure 4: Distribution of differences (late-early) in median line height.

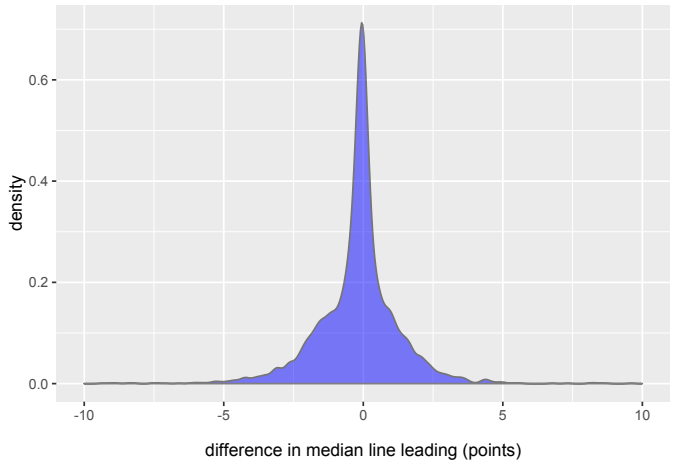


Figure 5: Distribution of differences (late-early) in median line leading.

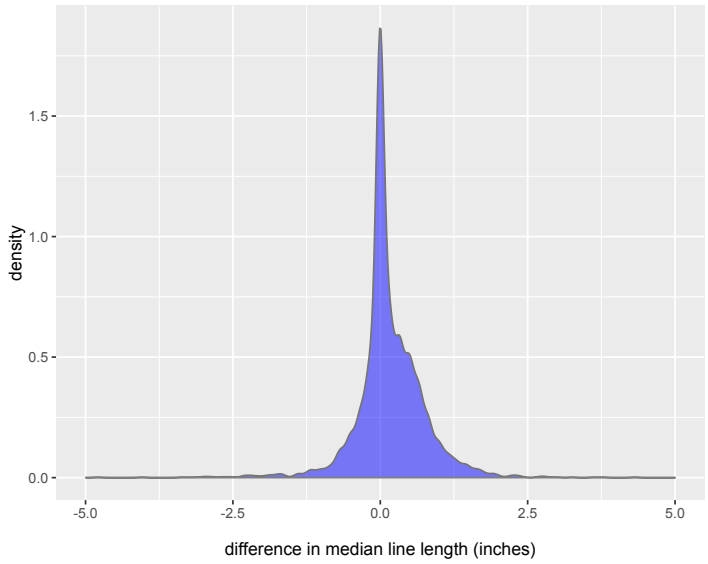


Figure 6: Distribution of differences (late-early) in median line length.

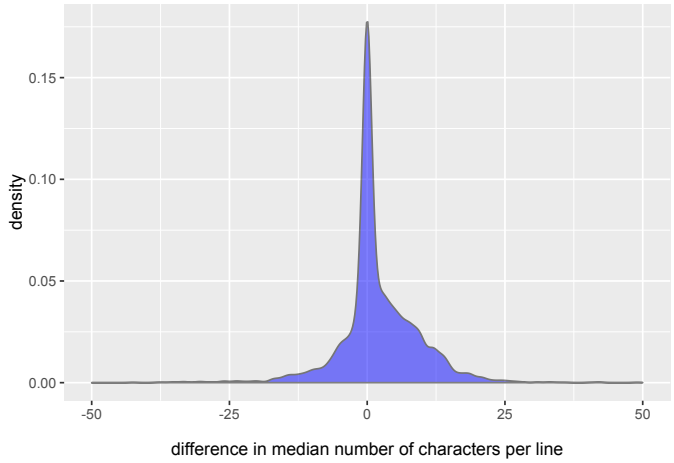


Figure 7: Distribution of differences (late-early) in median number of characters per line.

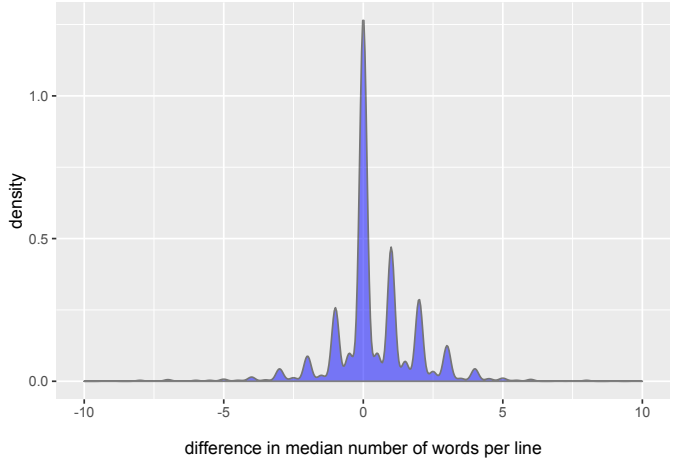


Figure 8: Distribution of differences (late-early) in median number of words per line.

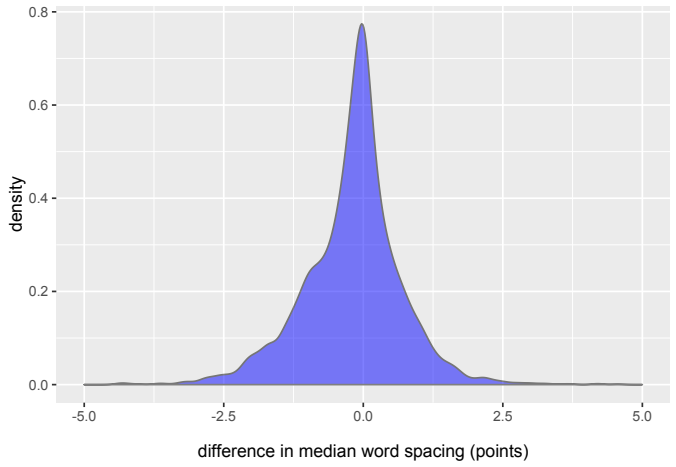


Figure 9: Distribution of differences (late-early) in median word spacing.

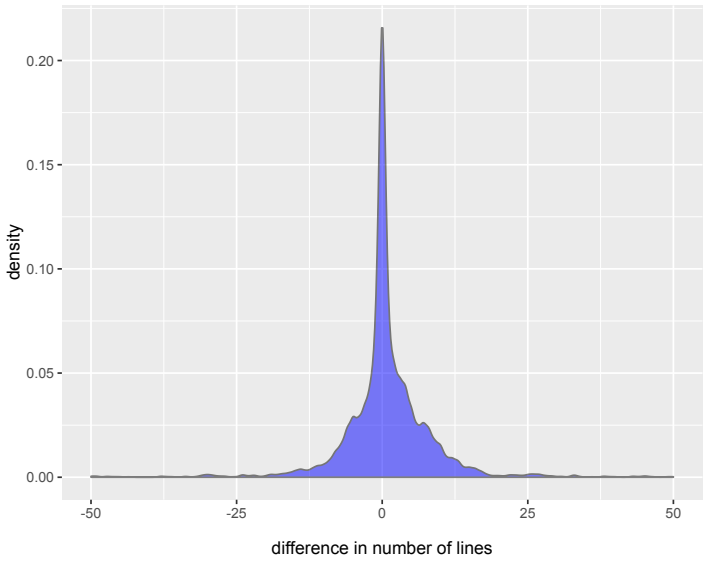


Figure 10: Distribution of differences (late-early) in median number of lines.

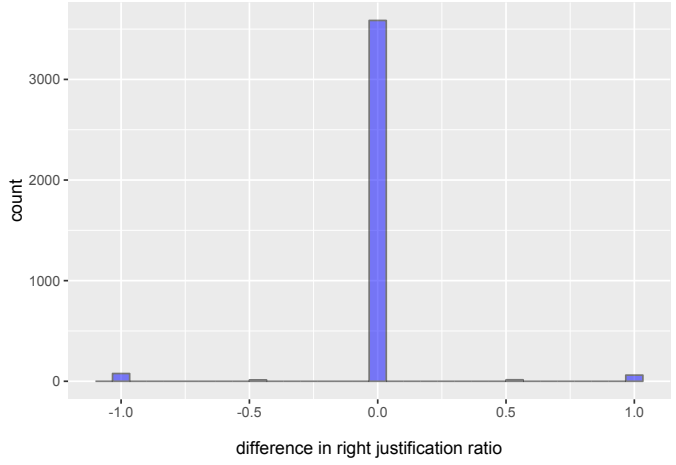


Figure 11: Distribution of differences (late-early) in median right justification ratio.