An Assessment of the Legibility of Google Books

Ryan James, University of Hawaii at Manoa, rsjames@hawaii.edu

Abstract:

Google books has been criticized for faulty metadata, problems with search functionality, copyright infringement, and legibility of scanned texts. This paper explores only the legibility of texts scanned by Google Books. A review of 2500 pages from 50 randomly selected books was undertaken. The results show less than 1% of pages had errors that affected their legibility. This paper concludes that while Google Books is not perfect, the majority of texts sampled were legible.

Introduction:

Many of the recent articles published about Google Books focus on copyright and issues surrounding Google's business arrangements with publishers, authors, and libraries. Fewer articles have directly evaluated the quality of the scanned images provided by this service.

Duguid (2007) evaluated the quality of Google Book's scan of *The Life and Opinions of Tristram Shandy, Gentleman*. While he acknowledges that the focus on a single title imposes problems in generalizing the results to all of the books scanned by Google, there are additional confounding factors. The popularity of a title makes it more likely users of Google Books would notify the company of errors in the text-block which then might be corrected. Less popular books might not receive the same attention.

The problems with Google Books extend beyond legibility to encompass faulty metadata, public domain/copyright concerns, and other related issues (Townsend, 2007). Like Duguid, Townsend limits his review of the quality of books digitized by Google to a single title.

Google Books also has problems with metadata and flaws created by their apparent attempt to automatically extract metadata from scanned texts. The instances of faulty metadata generated by an automated process are troubling, as the errors described in this article would likely also be a problem for an automated optical character recognition system (Nunberg, 2009, August 31). Several instances of faulty metadata have been found, and questions were raised about Google's approach to the processing metadata (Nunberg, 2009, August 29).

Further, there are limitations to Google Books' search functionality and its subject classification (Jacso, 2008, 2009). These errors would necessarily increase the difficulty of the user in finding a desired text, whether searching or browsing.

The purpose of this paper is to assess the integrity of the texts scanned by Google Books with a focus on legibility. Broadly defined legibility is the capability of a text of being

read or deciphered (Merriam-Webster Online, 2009). From the reader's perspective the essential question is "can I read this book?". The information must be presented in such a way that it can be readily consumed, free of errors and defects, introduced during the scanning process, that would frustrate and hinder the reader's attempts to use the information contained in the book.

In this paper major errors are defined as ones which render the text indecipherable (see figure 1). Examples are extremely blurred pages, missing pages, and scans of fingers turning pages. The end result is that significant information was lost during the scanning process with entire pages of text being unreadable.

Minor errors are defined as ones where the text is still decipherable, but the effort required on the part of the user rises to such a level as to be an obstacle to legibility (see figure 2). Examples include missing letters from a sizable number of words, blurred sections of pages, and contrast and resolution problems that "muddy" the scan. It is still possible for the reader to read the information on the page, but there is considerable difficulty.

The key difference between these classifications is that major errors represent information that is lost where minor errors represent information that is present but has reduced legibility.

Errors introduced into the text during the scanning process that do not affect legibility are not considered in this paper (i.e. small variations in color and contrast, skewed pages, and slight blurring.) While interesting from a quality control perspective, they are not as relevant to the user attempting to read or decipher the text.

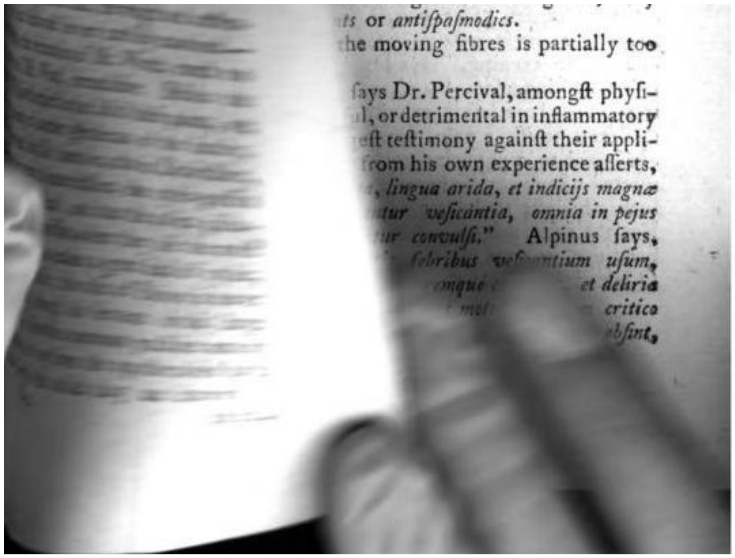

Figure 1: major error

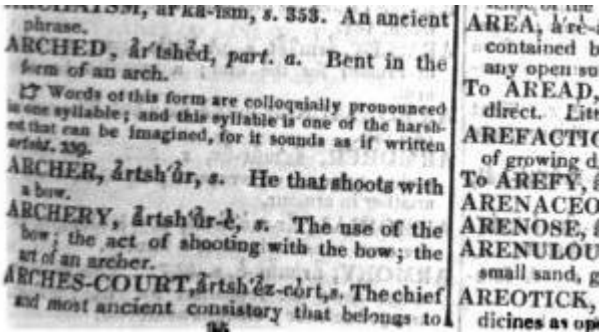
phrase.
ARCHED, ârtshĕd, part. a. Bent in the form of an arch.
☞ Words of this form are colloquially pronounced in one syllable; and this syllable is one of the harshest that can be imagined, for it sounds as if written artsht. 339.
ARCHER, ârtsh'ûr, s. He that shoots with a bow.
ARCHERY, ârtsh'ûr-ê, s. The use of the bow; the act of shooting with the bow; the art of an archer.
ARCHES-COURT, ârtsh'ĕz-côrt, s. The chief and most ancient consistory that belongs to
AREA, à'rê-
contained b
any open su
To AREAD,
direct. Litt
AREFACTI
of growing d
To AREFY,
ARENACEO
ARENOSE,
ARENULOU
small sand, g
AREOTICK,
dicines as op

Figure 2: minor error

Methodology:

A word was randomly selected from the Oxford English Dictionary and inputted into Google Books to generate a results list (Oxford English Dictionary Online, 2009). The list was then limited to full text only. Next a random number was generated between 1 and 100 and was used to select a title to examine from the results list (Random.org, 2009). Words generating results lists of fewer than one hundred titles were discarded. Selected titles that were fewer than 50 pages were also discarded and a new results list was generated from a new word.

Fifty titles were identified using this method and the first fifty pages of the text were examined. Tables of contents, title pages, and other prefatory material were excluded, as Google books does some additional processing on tables of contents and the legibility of pages which contain little or no text is difficult to assess given the methodology of this paper.

The chief advantage of the methodology used in this paper is that multiple titles were chosen randomly for examination. This helps to control for the biases resulting from examining a single known title, as popularity of the title, defects in the original scanned object, and chance could all play a role in altering the results.

This methodology focuses on books that are in the public domain. Copyrighted works are available for a "limited preview" through Google Books, but 50 consecutive pages usually can not be viewed, and are thus excluded from review. The average publishing date of all books reviewed was 1846, with the oldest being 1768 and the newest 1969 (a public domain work published by the US government).

Results:

Data was gathered during June and July of 2009. The first fifty pages of fifty titles were examined for a total of 2,500 pages. Fifteen pages with major errors and nine pages with minor errors were found. Four books had major errors and four books had minor errors. Only one book had both major and minor errors.

Of the books containing major errors the average number of errors was 3.75 per 50 pages reviewed. For minor errors the average was 2.25 per 50 pages reviewed.

Summary of Results	
Number of Books	50
Number of Pages	2500
Major Errors	15
% of Pages with Major Errors	0.6
Minor Errors	9
% of Pages with Minor Errors	0.36
% of Pages with both types of errors	0.96

Table 1: results

Discussion:

The overall error rate of 0.96% shows that while Google Books is not perfect, the majority of the texts sampled in this study were legible. The users of Google Books are likely to encounter the rare, but frustrating examples of errors explored in this paper. But the frequency of those errors should be low enough that the database of texts should prove valuable as a means to enhance access to books without seriously impeding the reading experience.

Minor errors is a slightly more subjective measure than major errors. Individual readers are likely to report variations in the number, especially given their use of the text. A person reading the text primarily for recreation may report fewer minor errors than a scholar using the text for in depth research.

In this light the 0.96% error rate found in texts scanned by Google Books can be interpreted in two ways. It is unlikely Google Books will replace the traditional rare books collection in a library. Scholars needing to perform a detailed analysis of a book are likely to view the error rate as unacceptable.

However, to a reader primarily interested in reading the same text for recreation and enjoyment, the 0.96% error rate might be a minor annoyance. To them it is likely more important that they have access to the text than that the text be 100% perfect.

This is not to say that Google Book's error rate is trivial, but that it exists on a continuum. Some readers will find Google Books more or less useful based on their intended use for the text.

In the future, an examination of the quality and legibility of materials available from Open Library, JSTOR, and other similar mass digitization projects is necessary. Such an analysis is complicated by the fact that other mass digitization projects sometimes use texts that were scanned by Google Books. Where this sharing of digitized texts occurs it becomes increasingly difficult to assess the quality of a given database of texts.

A comparison of the quality of texts in the different databases remains a necessary area to explore. From a practical perspective libraries and users can use these comparisons to better understand the value and utility of a given database. Further these comparisons can help establish best practices and standards for mass digitization projects.

Google Books is not perfect, but it does provide access to many books that would otherwise be difficult to read. However, the scholar's lament of poor quality scans does deserve serious consideration. Yet when we balance access versus perfection, access would seem to be the more important of the two.

References:

Duguid, P. (2007). Inheritance and loss? A brief survey of Google Books. *First Monday*, 12(8). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1972/1847

Jacso, P. (2008). Peter's Picks & Pans: Amazon, Google Book Search, and Google Scholar. *Online*, 32(2). Retrieved from http://www.onlinemag.net/mar08/index.shtml

Jacso, P. (2009, September 24). Newswire Analysis: Google Scholar's Ghost Authors, Lost Authors, and Other Problems. *Library Journal Newswire*. Retrieve from http://www.libraryjournal.com/article/CA6698580.html?q=jacso

Merriam-Webster Online Dictionary. Retrieved from http://www.merriam-webster.com

Nunberg, G. (2009, August 29). Google Books: A Metadata Train Wreck. *Language Log*. Retrieved from http://languagelog.ldc.upenn.edu/nll/?p=1701

Nunberg, G. (2009, August 31). Google's Book Search: A Disaster for Scholars. *The Chronicle Review*. Retrieved from http://chronicle.com/article/Googles-Book-Search-A/48245/

Oxford English Dictionary Online. (2009). Retrieved from http://www.oed.com

Random.org. (2009) Retrieved from http://www.random.org

Townsend, R. (2007). Google Books: Is It Good for History?. *Perspectives*, 45(6). Retrieved from http://www.historians.org/Perspectives/issues/2007/0709/0709vie1.cfm