Measuring Search Retrieval Accuracy of Uncorrected OCR: Findings from the Harvard-Radcliffe Online Historical Reference Shelf Digitization Project

Harvard University Library
LDI Project Team[1]
August, 2001

Please direct any comments to this report to: Stephen Chapman <stephen_chapman@harvard.edu>

Introduction

This report presents the findings of an investigation to evaluate the conditions for search retrieval successes and failures when using uncorrected OCR[2] for indexing. The purpose of the study was to assess whether low-cost, high-production techniques for text conversion were adequate to produce digital reproductions of consistent quality and usability. We sought to identify attributes of the original material or the OCR-produced text that could predict when additional, costly processes (OCR correction or keying) would be needed to meet retrieval requirements for text digitization projects.[3]

Our primary research question was, "Under what conditions would uncorrected OCR fail to meet baseline requirements for searching?" The corollary question, "Under what conditions would uncorrected OCR succeed?," was also of interest

We developed a test methodology and test system to measure quality according to "search accuracy"—rather than the traditional measure of character accuracy—for two reasons. First, our principal criterion for OCR usability is to facilitate search and retrieval, not display. We display page images,[4] not text, as the default format for digitized printed materials, so the OCR-

[1] Members of Harvard University's Library Digital Initiative technical team collaborated with project managers Robin McElheny of the Harvard University Archives and Jane Knowles of the Radcliffe Archives to develop the test methodology and to gather data. Jim Coleman and Clare McInerney of the Office for Information Systems developed the test systems, and Stephen Chapman of the Weissman Preservation Center wrote the report.

[2] For the purposes of this project, "uncorrected OCR" refers to plain text generated by optical character recognition of 1-bit TIFF images and using no text clean up processes (such as spell checking) following OCR.

[3] We assume that minimum thresholds for search retrieval will vary from collection to collection according to user needs as well as budgetary considerations. Therefore, we did not explore the conditions under which a specific threshold—e.g., 95%—could be achieved. The methodology we used in this investigation (see "Test") serves to answer the project manager's question, "What percentage of searches will succeed with uncorrected OCR?" In cost/benefit analyses, the answer to this question (what one can get) may then be compared to stated or assumed preferences of what users want (e.g., 100% search and retrieval accuracy) in order to develop a realistic budget for a text conversion project.

[4] Created by scanning, "page images" are relatively inexpensive to produce and can adequately serve as authoritative reproductions of historic materials. As photographic formats (like photocopies or photographic negatives and prints), these images convey the original layout, fonts, and dimensions of the original printed pages. They are not searchable. The quality of page images varies according to the specifications used in scanning.

generated text does not need to serve as a reliable surrogate for the originals. We view the key functional requirement for OCR as facilitating searches that retrieve corresponding page images that contain the search word or phrase. (This is not the same as saying that a search engine needs to provide an accurate count of how many times a word or phrase appears on a page or within a larger dataset.) Our search methodology is based on simple retrieval: if any instance of the word on a page is right, the page is retrieved. Some retrieval systems use word counts to provide a "relevancy rating" to help users identify documents of greater probable interest. Our search methodology does not take such ranking algorithms into account.

One might presume that if character accuracy is 100%, then search and retrieval will be 100%, but the quality control overhead to check character accuracy points to our second reason for associating quality with search and retrieval. Counting character errors is a manual, time-consuming process. We wondered, "Could an easier, less error-prone process be developed to ascertain the quality of OCR?"

This report parallels the study, "Measuring the Accuracy of the OCR in the Making of America," conducted in 1998 by the University of Michigan.[5] The physical characteristics of the original source materials and the attributes of the digital reproductions were comparable in both projects, and the motivations for the studies were the same. Both investigations measured failure for uncorrected OCR in service of the goal to minimize costs for text digitization. Uncorrected OCR costs so much less than corrected OCR or keyed text[6] that its acceptance might be *the* underpinning of large-scale text conversion. At the time of the Harvard study, for example, keying costs were approximately 10-13 times more expensive per page than uncorrected OCR.

The two studies differ in the methodology used to determine whether uncorrected OCR was "reasonably accurate." The University of Michigan defined accuracy by the number of character and word errors, and explored whether percentages of failure could be reliably inferred from OCR software confidence scores. Harvard defined accuracy by the number of failed searches, and explored whether these percentages could be inferred from OCR confidence scores, publication dates of original materials, or other page attributes. Harvard did not seek to correlate retrieval accuracy with character accuracy; Michigan did not seek to correlate character accuracy with retrieval accuracy.

Michigan found that OCR character accuracy can be measured automatically if the OCR program generates confidence scores, and provided that one correlates these scores with actual counts of character errors. As documented in this report, Harvard found that OCR search accuracy can be measured by using a test system and reliable (human) subjects; because

---

[5] Douglas A. Bicknese, "Measuring the Accuracy of the OCR in the Making of America," 1998. Available online: http://moa.umdl.umich.edu/moaocr.html.

[6] When working from printed pages (or their analog or digital image surrogates), there are only two ways to create machine-readable text. "Keying" refers to (a human) transcribing text from the source into a computer program. OCR refers to (a machine) acquiring text from digital images automatically. Errors *will* be made in the first pass in both operations. Therefore, "double keying" refers to processes in which two operators transcribe the same pages and their results are compared. "Corrected OCR" refers to the process in which an operator reviews the "raw OCR" and manually corrects spelling and/or formatting errors.

confidence scores did not correlate with search accuracy or failure, Harvard's methodology and quality control procedures could be used with any OCR software.

Sample

In 1999, the Harvard University Archives and Radcliffe Archives initiated a joint project, *The Harvard-Radcliffe Online Historical Reference Shelf*,[7] to digitize frequently-consulted publications documenting Harvard's history. These included the annual reports of Harvard University and Radcliffe College, published between 1825 and 1988, totaling 103,871 pages.

The project plan was to create digital images plus searchable text from each report. Page images would be delivered via Harvard's Page Delivery Service[8] (being developed simultaneously as an infrastructure project), and machine-readable text for each page would be indexed and made available for searching, but not display.

Population selected for the study

The population selected for this investigation numbered 225 reports (49% of the collection), totaling 42,272 page images (41% of the collection). This subset was far larger than needed to form a statistically valid sample, but project managers wanted to give users the experience of viewing and searching at least one of the reports published each year. (Annual reports were published as Presidents' Reports and Treasurers' Reports, for both Harvard and Radcliffe.)

The representative sample included comparable numbers of reports from the 19th and 20th centuries, although reports published from 1825-1844 were not included in the test because they had not yet been digitized. We presumed that failure rates would be higher for 19th-century than for 20th-century reports because of the lower contrast of older printed pages.

Table 1. Reports Included in Sample

publication dates	# reports	# pages	ave. # pages per report
19th century (1845-1899)	110	8,946	81
20th century (1900-1988)	115	33,324	290
Totals	**225**	**42,272**	--

Physical characteristics of original materials

All reports are published in English. Almost all of the content is text, and the text is always machine-printed. The reports contain no pictorial information, but often include covers and line-art illustrations (charts). These relatively complex pages are sometimes printed in color. (In 672 instances, original pages were scanned twice to produce two versions of page images: 1-bit TIFFs for OCR, and grayscale or color TIFFs for archiving and display.)

[7] For more information about the Harvard/Radcliffe On-line Historical Reference Shelf, as well as a gateway to the digitized annual reports, see, http://hul.harvard.edu/huarc/refshelf/HROHRSHome.htm.

[8] For a description of Harvard's "Page Delivery Service," see, http://hul.harvard.edu/ldi/html/delivery.html#page.

The majority of pages are laid out in single columns, although indexes are printed in two columns. Page sizes vary, but are typically close to 6" *x* 9". Contrast varies according differences in original paper and the effects of aging, with paper in older reports generally being of lower contrast.

Production of page images and structural metadata
(Harvard College Library Digital Imaging Group)

With few exceptions, Harvard and Radcliffe owned multiple copies of each report. The copies selected for digitization were guillotined. Single pages, printed on two sides, were scanned in one pass, via an autofeeder, to produce 600 dpi 1-bit TIFF images with Group 4 compression.

When only one copy of a report was available, the original was left intact in its binding and photographed in grayscale at a digital camera. The grayscale images were subsequently converted to produce 1-bit TIFF images, with the same specifications described above.

Following scanning, the page images were optimized for OCR with TMS Sequoia's ScanFix software: deskewing, despeckling and other settings were used. The Digital Imaging Group then created structural metadata. The resulting digital object for each report consisted of a sequence of TIFF images and a single tab-delimited data file that associated sequence number, page number, page orientation, and feature name (when applicable), with each file name. These were copied to CD-ROM to deliver to the University of Michigan for OCR and markup.

Production of OCR and SGML
(University of Michigan Digital Library Production Services)

The production workflow used in this project was identical to the Making of America's, as summarized in the Michigan study.[9] Prime Recognition OCR software was used; from each TIFF image, Prime generates a plain text file and a "Prime Score" number, from 100-900. This score, with 900 being perfect, represents the program's confidence rating of character accuracy. Following OCR, scripts were used to remove end of line hyphens (to rejoin split words) and to apply low-level TEI markup with the Harvard-provided structural metadata. Prime Scores for each page were recorded in the SGML file and subsequently analyzed (see "Summary of Findings"). The TEI-encoded SGML files produced by UM DLPS were used in the Test, but later converted to MOA2 XML files, the master format we use for multi-page objects.

Quality

Every step in the production process described above contributed in some measure to the final quality of the 225 SGML files (the containers for the uncorrected OCR) we used in the test. We did not isolate variables to determine each component's (system, input settings, operator judgment) influence upon search retrieval accuracy. Instead, we presumed that these specifications would at least be as successful for these materials as they had been for the books and journals digitized by similar means in the JSTOR and Making of America projects. The

[9] Bicknese, op. cit.

digitization specifications (and systems) represented state-of-the-art production workflows in 1999.

Test

The test methodology and test system were developed in the fall of 1999 by Jim Coleman and Clare McInerney in the Harvard University Library Office for Information Systems. The system was configured as a web site that stored the 225 SGML files—one per report—and 42,272 page images that comprised the sample. *No pages were removed from the sample prior to the test*. Front matter (covers and cover versos), blank pages, all text pages, tables, and end matter (indexes, and cover rectos) were included.

The tests themselves were conducted in three sessions during the winter of 1999-2000. Participants included the project managers—Robin McElheny and Jane Knowles of the Harvard University Archives and the Radcliffe Archives—their staff, and selected researchers. All had knowledge of the materials and are presumed to have submitted representative real-world queries to the test system.

Procedures

Procedures were followed in the test sessions to produce a log of searches for 5%, randomly selected, of the population selected for the study. This percentage was selected to produce statistically valid results and to complement the findings gathered in the Michigan study.

The test system was programmed to:

- present a list of reports, from which the researcher could randomly select a report to search
- randomly pick one of the page images from the selected report
- present the page image (GIF) and a dialog box for the researcher to enter one search term for that page
  - the researcher entered a "meaningful" search word or phrase at his or her discretion
  - stop words (based on a pre-approved list) were eliminated from string searches
  - the search terms themselves were not saved for analysis, as the purpose of the test was not to study how people would use the materials, only to document retrieval successes and failures
- parse the uncorrected OCR (in the corresponding SGML file) and record the result in a database as either found or not found
  - search successes or failures were not presented to the user
- allow the researcher to instruct the system to present the next randomly selected page from the report, or to return to the list of all reports and choose another one to search

When the three sessions concluded, at least two pages were selected randomly from each report. Due to rounding, however, slightly less than 5% of the sample had been searched: 1,997 of the 42,272 page images, or 4.724%.

Because nine of the randomly selected page images were blanks, these searches (which had been recorded as failures) were eliminated from the total. Thus, the test resulted in 1,986 valid searches.

Summary of Findings

The following table presents the findings of OCR search accuracy, with allowances for sampling errors.[10]

Table 2. Search Retrieval Accuracy of Uncorrected OCR

	population	sample	% of population	% success	+/- sample error
all years	42,270	1,986	4.7%	**96.6%**	**0.8%**
20th century	33,324	1,596	4.8%	96.9%	0.8%
19th century	8,946	390	4.4%	95.1%	2.1%

Although these findings present a lower success rate for 19th century reports, this difference could possibly be explained by the difference in sampling error. In retrospect, we should have ensured that equal percentages of the 19th and 20th century reports were searched in the test.

Having determined that **96.6%** $\pm$ 0.8% of searches (all years) would succeed with uncorrected OCR, we evaluated search accuracy and search failure against the variables of Prime Score (OCR confidence), publication date, and page characteristics.

Correlations

1. Prime Score with Search Accuracy (rates of search retrieval success and failure)

We evaluated the relationship between Prime Scores and search accuracy to see if our findings would support Michigan's conclusion that a Prime Score "cut-off point" could be identified to segregate acceptable from possibly unacceptable results. Michigan concluded that pages with Prime Scores above 880 would be "reasonably accurate"[11]—that is, reliably have a high enough percentage of character accuracy ($\geq$ 99.0%) not to require clean up.

We found no correlation between Prime Score and search accuracy.

Search successes and failures were distributed across the range of Prime Scores, with highest counts for each in the 890-898 range: 40.8% of all success and 39.7% of all failures occurred in the highest range of Prime Scores. (See Appendixes 1.1 and 2.1.) Search failure and success

[10] The formula used to calculate the plus or minus sample error = 1.96*(SQRT(%found in sample*(1-%found in sample)/(sample size-1))). This formula was developed by Jim Self for the Weissman Preservation Center in November, 2000. Error rate varies according to size of collection (10,000 item threshold point), size of sample, and percentage found in the sample.

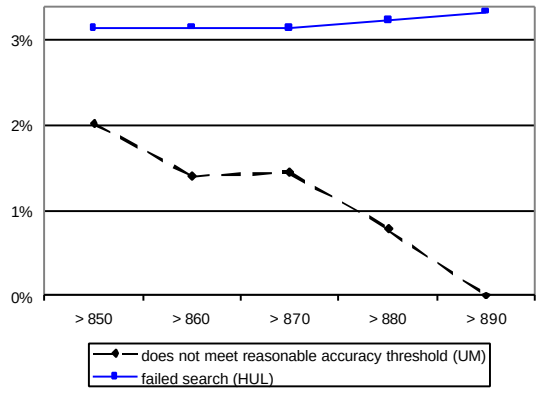
[11] In the Michigan study, "reasonably accurate" was defined as: "99.0% or better character accuracy for pages which did not contain an illustration, blank page, table, or advertisement."

rates were nearly identical for the populations above and below the median Prime Score (882). For pages with Prime Scores of 882-900, thirty-four searches failed and 96.64% were successful. For pages with Prime Scores below 882, thirty-four searches failed and 96.50% were successful. These numbers suggest that the likelihood of search success or failure does not change meaningfully according to Prime Score.

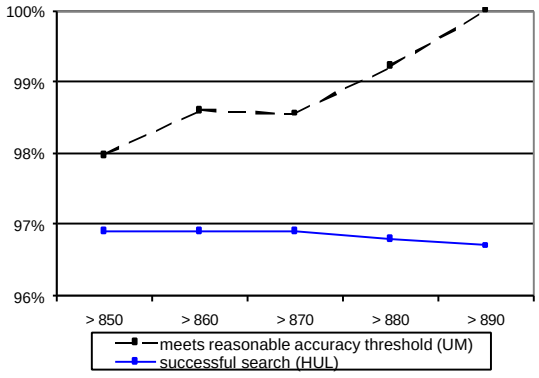
To compare our results with Michigan's, we plotted the percentages of failure and accuracy for Prime Scores above 850, 860, 870, 880, and 890.[12] (Failure in the Michigan study meant that the page did not meet the threshold of 99% or more of the characters being accurate; failure in the Harvard study meant that the search word or phrase did not retrieve the page.)

Table 3. Prime Scores to Accuracy: UM and Harvard Findings Compared

A. Prime Score to % Failure



B. Prime Score to % Success



The trendlines in Tables 3A-B reveal that search failure rates slightly increased with Prime Scores, whereas Michigan found that character errors decreased, particularly with Prime Scores above 870. (See Appendixes 1.2 and 2.2 for additional data.)

These trends (i.e., the shape of the lines in the graphs) are markedly different, and suggest that there is not a correlation between rates of character accuracy and rates of search accuracy. Additional testing would be needed to verify this assumption, but one should not necessarily assume that 99% accuracy (according to Prime Score) means that 99% of searches will retrieve corresponding page images.

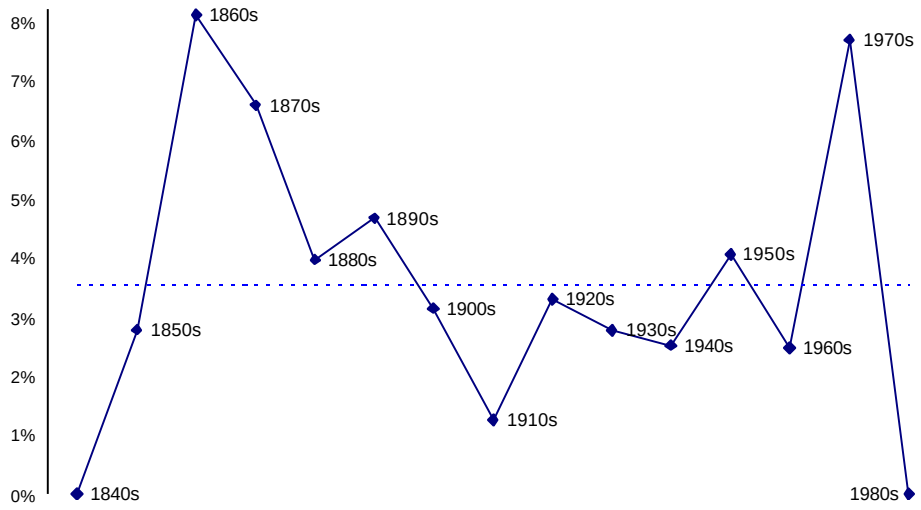
[12] UM Report Section III, "Analyzing the Results."

2. Publication Dates with Search Accuracy

As noted in Table 2 (p. 6), the raw percentages of search successes (without allowing for sampling error) indicated a 1.8% difference between 19th-century and 20th-century reports.

We presumed that percentages of failure would be higher for older reports, but when plotted chronologically by decade, the search failure rates do not show a smooth declining trend (from old to recent materials):

Table 4a. Percentage of Failed Searches Per Decade, Plotted Chronologically



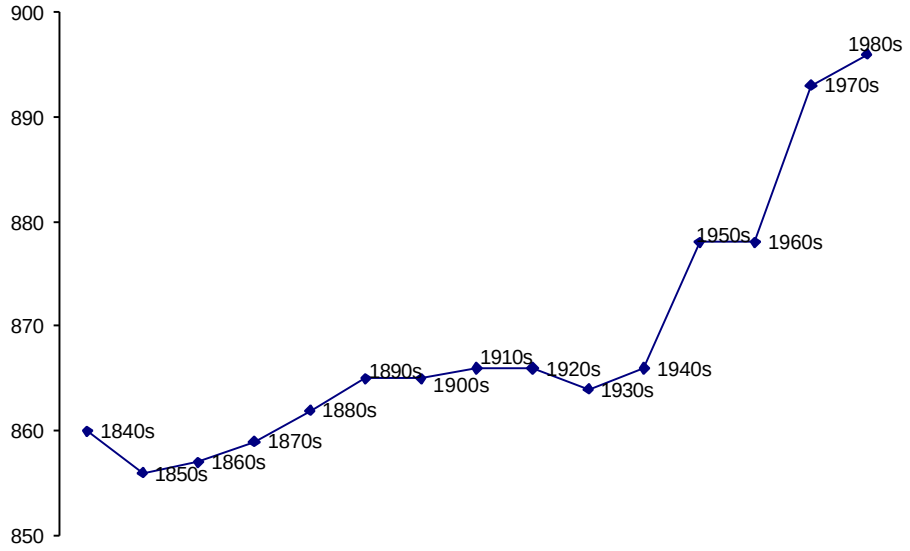
(dotted line = 3.55%, overall average failure rate per decade)

When comparing failure rates by decade to the overall baseline average of 3.55% (dotted trendline above), one finds that four of six 19th century decades had failure rates above average, whereas only two of the nine 20th century decades were in this higher range. (See Appendixes 1.3a-b and 2.3a-b for additional views of the correlations between publication date and search successes and failures.)

Correlation of Prime Score to Publication Date

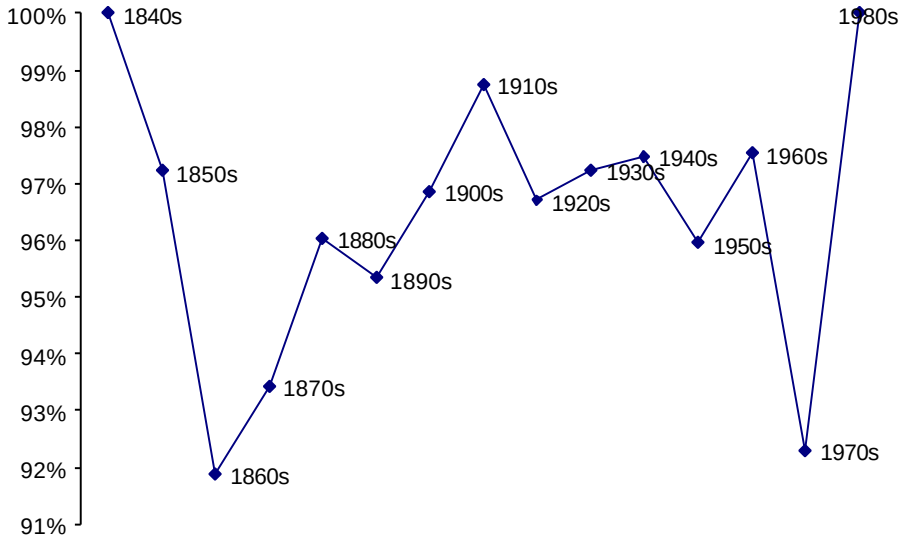
Plotted chronologically, average Prime Scores per decade do increase over time, suggesting that rates of character accuracy improve with modern materials:

Table 4b. Average Prime Scores Per Decade, Plotted Chronologically



This increase in Prime Scores, however, was not matched by search success rates plotted chronologically:

Table 4c. Percentage of Successful Searches Per Decade, Plotted Chronologically



If there are meaningful differences in retrieval by century—and our test suggests that this difference may be as many as three percentage points for rates of successful retrieval (see Table 2, allowing for sampling errors)—Tables 4b and 4c suggest then an attribute other than Prime Score would need to be identified to establish a reliable cut-off point.

4. Page Characteristics with Rates of Search Failure

We did not make a rigorous study of search failures with page features. We can, however, infer some findings from the pages randomly selected for our Test:

- no evidence of failure on front matter: 13 of the 13 first and second page images succeeded
- no evidence of failure on end matter: 9 of the 10 last pages and next-to-last pages succeeded (given this high rate of success, we did not determine whether the failure was a search of a back cover recto, two-column index page, or standard index)

We did not determine whether "non-standard text" pages factored meaningfully in the 68 failures in the test. The annual reports included covers, foldouts, errata notes, and two-column indexes.

Implications for Future Text Conversion Projects

The percentage of search retrieval accuracy can be determined relatively easily with the test methodology and system designed for this investigation. Although the methodology is not fully automated—one needs to use trustworthy subjects to enter search terms—it is quicker than printing OCR-produced text and comparing it to the source (TIFF image or original printed page). More importantly, the methodology reveals the likelihood of success or failure for indexing and retrieval, whereas percentages of character accuracy do not appear to correlate with percentages of search accuracy.

Prime Scores were not found to be a convenient way of narrowing the populations of data that would need to be corrected to meet baseline levels of search accuracy. Although we would very likely continue to collect these confidence ratings from the sophisticated OCR packages that generate them—free metadata, one might reason, should be gathered and stored for any potential value it offers—we would not infer a likelihood of retrieval success or failure based on such scores. We might conclude from this study that one should evaluate uncorrected OCR generated by Prime with the same techniques that would be used for other programs that do not produce confidence ratings: use spell check and other techniques to flag suspected words for correction.

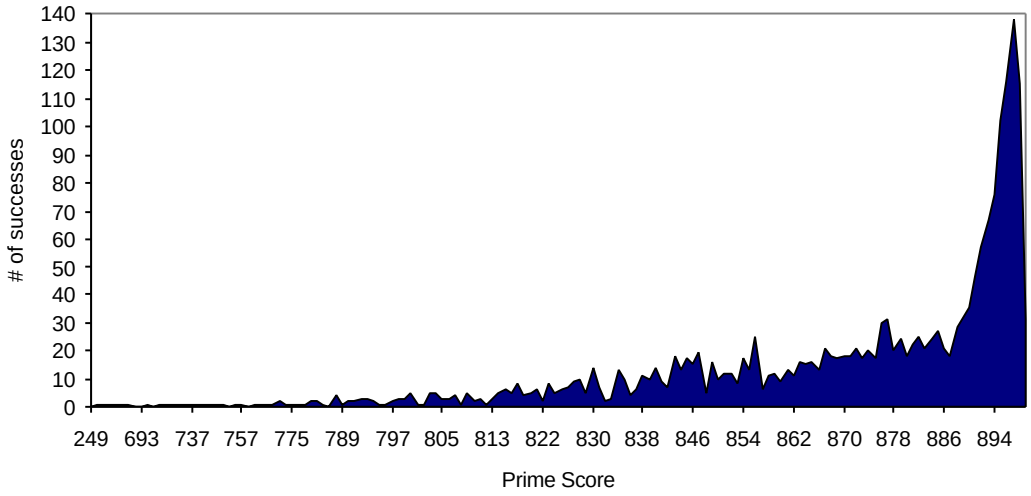
The results of this investigation suggest that it is worthwhile to record the publication dates of the digitized text to determine whether older materials will, as might be assumed, necessitate either OCR cleanup or even keying in order to meet the baseline average rate of retrieval accuracy for more recently published material. Further studies would need to be conducted to determine whether automated methods could be used—at the time of scanning, image processing, during OCR, or in combination—to improve the retrieval accuracy of uncorrected OCR for older materials.

Appendixes: Data Correlations

Appendix 1 (4 charts), *Search Accuracy*, correlates search successes with Prime Score, and successes with publication dates.

Appendix 2 (4 charts), *Search Failures*, correlates search failures with Prime Score, and failures with publication dates.

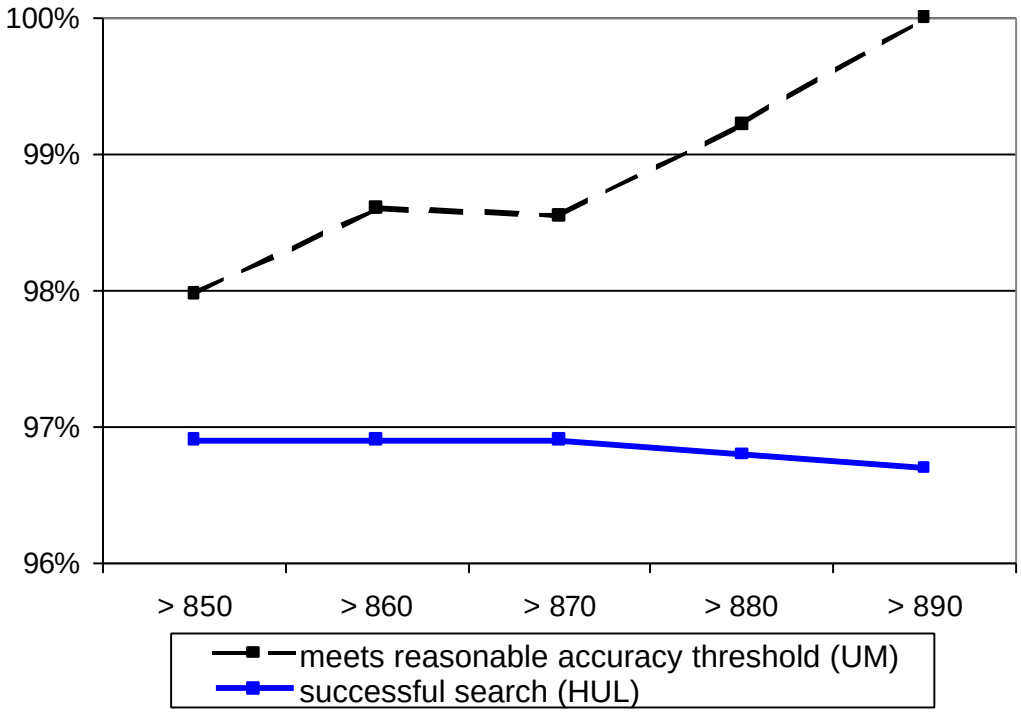
1 SEARCH ACCURACY



1.1 Frequency Distribution of Successful Searches, by Prime Score

Prime Score	total # searches	# successes
< 600	3	2
600-699	8	5
700-749	9	9
750-799	51	46
800-809	35	33
810-819	39	37
820-829	65	63
830-839	83	80
840-849	134	133
850-859	130	126
860-869	154	149
870-879	222	216
880-889	243	236
890-898	810	783

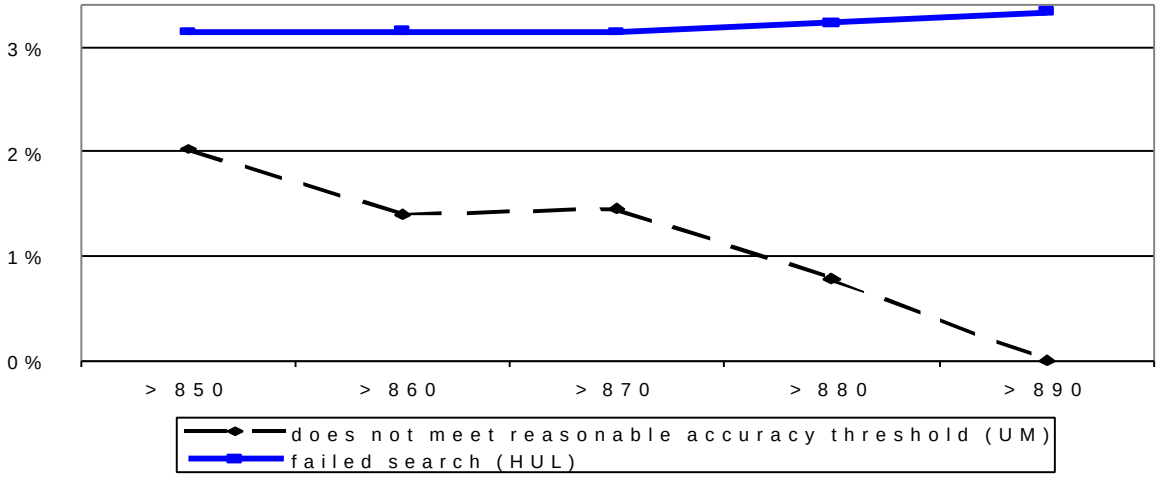
1.2 Prime Score to % Success: Harvard and University of Michigan Findings



Prime Score	character accuracy > 99.0% ("reasonably accurate") (UM)	search accuracy = 100% (page retrieved) (Harvard)
≥ 249	data not available	96.60%
≥ 600	data not available	96.60%
≥ 750	data not available	96.70%
≥ 800	data not available	96.90%
≥ 850	97.98%	96.90%
≥ 860	98.60%	96.90%
≥ 870	98.55%	96.90%
≥ 880	99.22%	96.80%
≥ 890	100.00%	96.70%

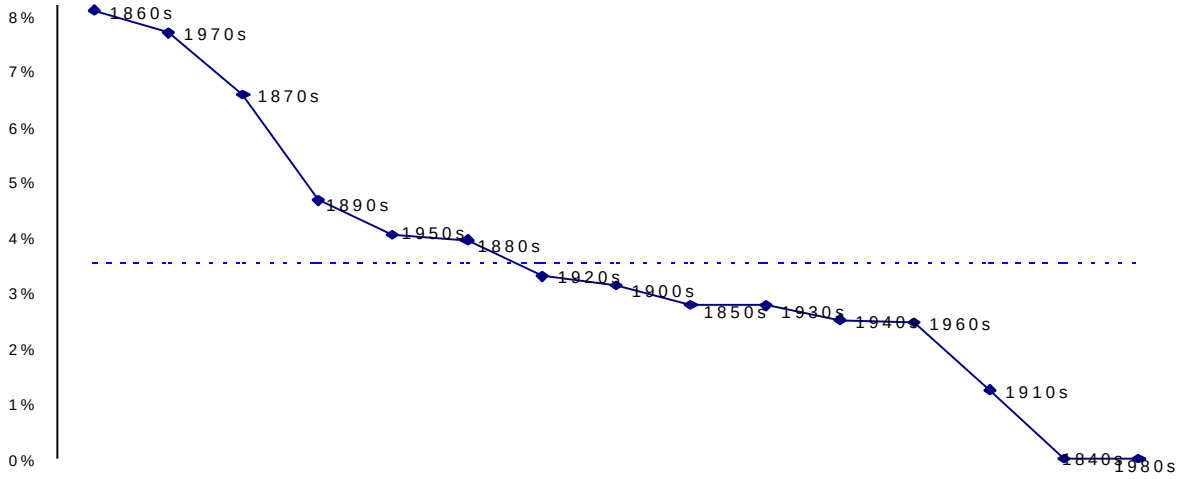
1.3 Date to % Success

(a) Rates of Success By Decade (to baseline average per decade), ascending sort



(dotted line = 96.4%, overall average success rate)

(b) Rates of Success By Decade (to baseline average per decade), chronological sort

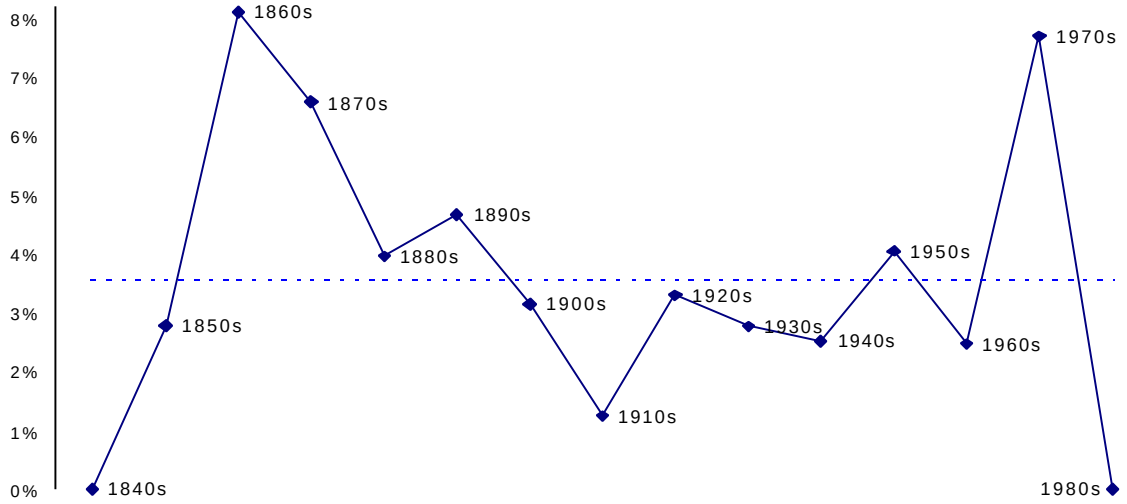


(dotted line = 96.4%, overall average success rate)

decade	% success	total # searches	# successes
1840s	100.0%	10	10
1850s	97.2%	36	35
1860s	91.9%	37	34
1870s	93.4%	76	71
1880s	96.0%	101	97
1890s	95.3%	150	143
1900s	96.9%	159	154
1910s	98.8%	80	79
1920s	96.7%	91	88
1930s	97.2%	289	281
1940s	97.5%	358	349
1950s	96.0%	420	403
1960s	97.5%	162	158
1970s	92.3%	13	12
1980s	100.0%	4	4

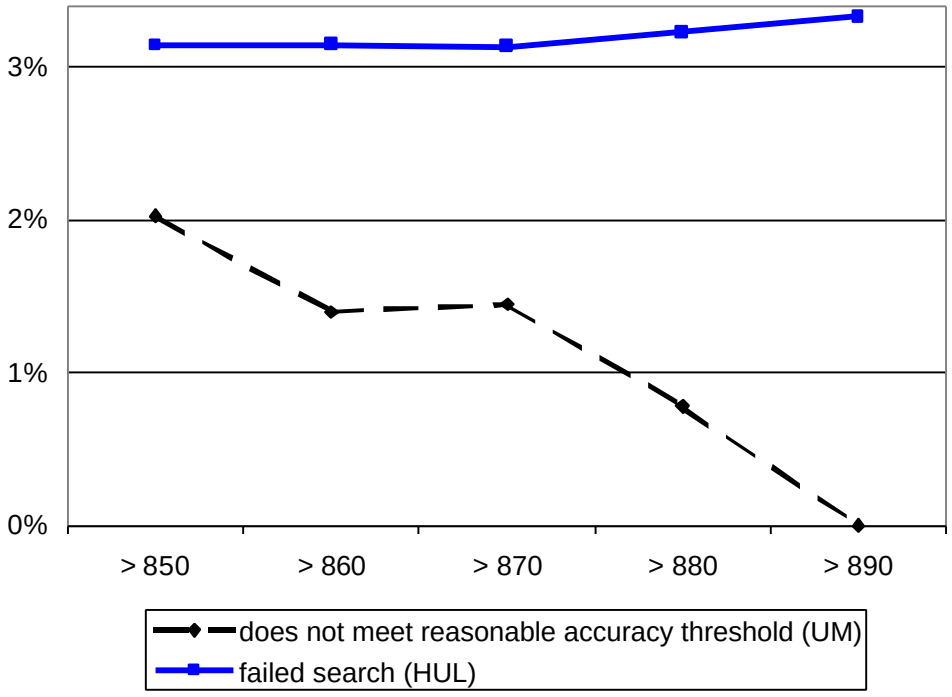
2 SEARCH FAILURES

2.1 Frequency Distribution of Search Failures, by Prime Score



Prime Score	total # searches	# failures
< 600	3	1
600-699	8	3
700-749	9	0
750-799	51	5
800-809	35	2
810-819	39	2
820-829	65	2
830-839	83	3
840-849	134	1
850-859	130	4
860-869	154	5
870-879	222	6
880-889	243	7
890-898	810	27

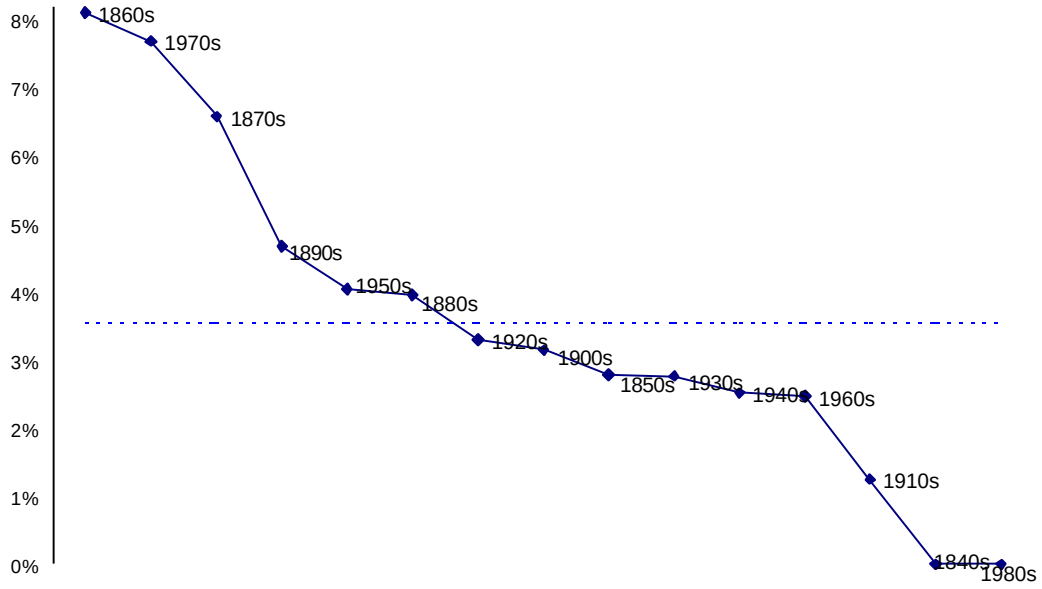
2.2 Prime Score to % Failure: Harvard and University of Michigan Findings



Prime Score	character accuracy < 99.0% (not "reasonably accurate") (UM)	search accuracy = 0% (page not retrieved) (Harvard)
≥ 249	data not available	3.42%
≥ 600	data not available	3.38%
≥ 750	data not available	3.26%
≥ 800	data not available	3.08%
≥ 850	2.03%	3.14%
≥ 860	1.40%	3.15%
≥ 870	1.45%	3.14%
≥ 880	0.78%	3.23%
≥ 890	0.00%	3.33%

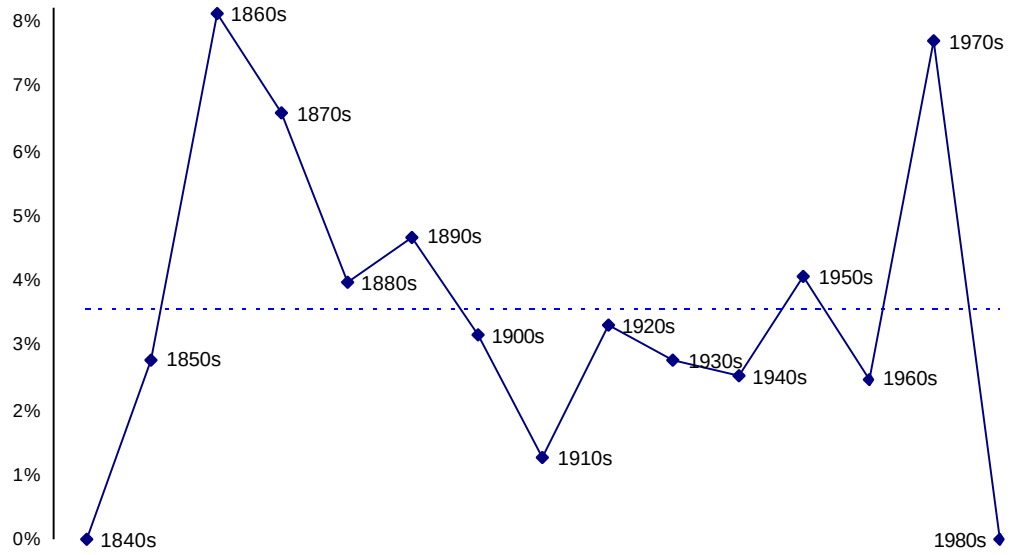
2.3 Date to % Failure

(a) Rates of Failure By Decade (to baseline average per decade), descending sort



(dotted line = 3.55%, overall average failure rate)

(b) Rates of Failure By Decade (to baseline average per decade), chronological sort



(dotted line = 3.55%, overall average failure rate)

decade	% failure	total # searches	# failures
1840s	0.00	10	0
1850s	2.78	36	1
1860s	8.11	37	3
1870s	6.58	76	5
1880s	3.96	101	4
1890s	4.67	150	7
1900s	3.14	159	5
1910s	1.25	80	1
1920s	3.30	91	3
1930s	2.77	289	8
1940s	2.51	358	9
1950s	4.05	420	17
1960s	2.47	162	4
1970s	7.69	13	1
1980s	0.00	4	0

