Estimating Word Error Rate in PDF Files of Old Newspapers

by Paul Bullock

For more than 10 years I have been using the Old Fulton NY Post Card Website to search for newspaper articles about the Bullocks and Rectors of old. Although I haven't counted, my conservative guess is that I've had more than 1000 hits from articles in a number of newspapers. The papers include the Penn Yan Democrat, Geneva Daily Times, Finger Lakes Times, Naples Record, Clifton Springs Press, Shortsville Enterprise, Dobbs Ferry Register, Cuba Patriot, Livonia Gazette and Corning Journal. The site at http://www.fultonhistory.com/Fulton.html has more than 30,000,000 old newspaper pages in PDF files that are searchable. The site is free and is the go-to place to search old New York State newspapers. Tom Tryniski, the one man team at the Old Fulton site, takes existing microfilm of the old newspapers and uses high speed and high quality equipment to convert the film images to digital files. He then uses OCR (optical character recognition) software to produce PDF files that are searchable. On his website he is quick to say that not all available microfilm is amenable to the OCR process.[1]

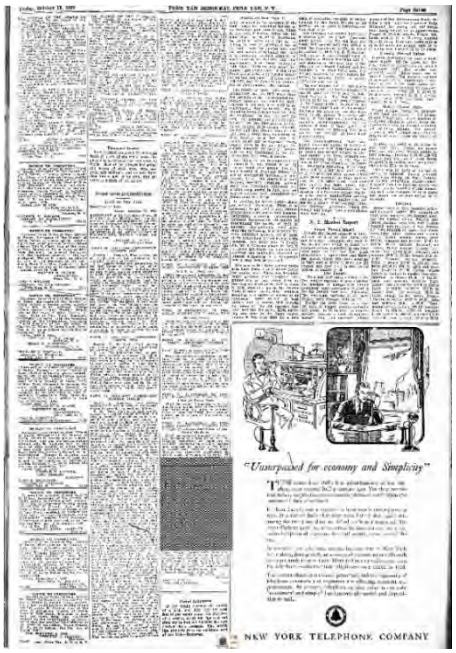
We, at the Yates County History Center, have established an Imaging Center to image our paper holdings. Our primary efforts will be aimed at the almost 1,000,000 pages we have of old newspapers. Since Mr. Tryniski at the Old Fulton site has used microfilm images to make a large number of Penn Yan Democrat pages searchable, we will first attack the approximately 100,000 pages we have of the Yates County Chronicle and Chronicle-Express. The microfilm for the Chronicle has been judged not amenable to the OCR process. Our approach is to use a digital SLR camera to image the pages. These images will then be converted to searchable PDF files using OCR software.

The process of imaging old newspaper text and then OCRing the image to make the text searchable is not perfect and errors will be made. As we start out we ask, "What is a reasonable standard for word error rate?" Because of its demonstrated usefulness the Old Fulton site immediately comes to mind. For study, 20 issues of the Penn Yan Democrat that are on the Old Fulton site were selected. The issues that range from the 1850s to the 1940s are listed in Table 1.

---

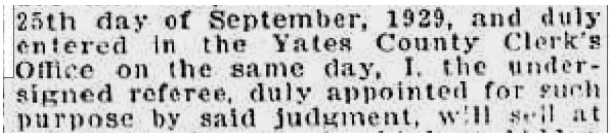
[1] A number of microfilm companies filmed the old newspapers for the Newspaper Project. Some of the filming was of very poor quality newsprint (faded, torn, creased, excessive ink bleed) so that an acceptable image (for OCR) was next to impossible to obtain. The results are a perfect example of WYSIWYG - what you see is what you get. To make this process even more interesting, microfilm comes in three (3) generations. The First Generation consists of Master Negative or originals. The Second Generation is the Print Master Negative which is made from the Master Negative. The Third Generation, or Service Copy positive, is made from the Print Master Negative. Each time you go down a generation, you lose image quality. This website almost always gets 3rd generation Microfilm (also called Service copy positive), or, as I call it, "bottom of the barrel" microfilm. This generation of film has been used many times at the various libraries and has a good deal of wear, tear, rips, scratches, dirt and splices. Despite these imperfections, remarkable technologies have been developed [word recognition, aka OCR software] which extract acceptable images to text from this very poor source material (remember some of the Newspaper pages go back to the early 1800's). Not withstanding the marvels of this technology, it cannot find or correct what is not there. As a result, although most pages are legible, some, regrettably, are not.

The PDF files of all of the pages in each of the 20 issues selected were downloaded from the Old Fulton site. In all, 128 pages were downloaded and studied out of the approximately 20,000 pages of the Penn Yan Democrat available on the Old Fulton site.



A random sample of 32 or more words was selected from each issue. The early issues had 4 pages so 8 words were selected from each page. In the issues with 8 pages, 4 or more words were selected from each page. Each word on a page was selected at random by the following procedure. The page was set up as a grid with the number columns across the page and 8 rows down the page. The column number and row number were selected at random using a random number generator. Table 2 shows the worksheet for sampling the October 11, 1929 issue of the Penn Yan Democrat. The worksheet, developed using Excel, randomly selects the column and the row where each sample word is to be selected. For example, the third sample word from page 7 of the issue was selected from the words contained in the intersection of column 3 and row 7 as shown in the image of page 7 to the right. The final word selection within the small rectangle was made on the computer screen with the PDF file of page 7 showing. That word turned out to be "same." All 32 samples from the issue were taken in this manner.

Next, each of the 32 sampled words were searched for in the 8 pages of the October 11th issue using Adobe reader. If the word was found in the search it was called a hit. If it was not found it was called a miss or an error. In this case there were 5 misses and 27 hits. One of the errors was that word "same" in column 3, row 7 of page 7.


25th day of September, 1929, and duly entered in the Yates County Clerk's Office on the same day, I. the undersigned referee, duly appointed for such purpose by said judgment, will sell at

25th day of September, 1929, and duly e n t e r e d in t h e Yates County Clerk's Office on t he s a m e day, I. t h e unders i g n e d referee, duly appointed for such p u r p o s e by said judgment, will sell

It's important to discuss here the limitations of OCR and PDF files when dealing with old newsprint. For purposes of this particular application, think of a PDF file as having two layers. The top layer is a view of the document as it appears. The second layer, which under normal circumstances can't be seen by the user, is the text of the document that results from the OCR process. The example to the left, containing the word "same" that was deemed a miss, shows the top layer on top and the second layer on the bottom. Note that the word "same" on the third line was misinterpreted by the OCR process as "s a m e"; spaces were inserted between the letters. As can be seen this is a fairly common OCR misinterpretation.

Looking first at the results for the October 11, 1929 issue in Table shows 8 pages with 4 sample words selected per page for a total of 32 words. The number of missed words, 5, out of the 32 is shown as well as the percent misses, 16%. The 95% confidence limits on the true but unknown percent misses in the thousands of words in that issue is 7% through 32%. The true value could be as high as 32% or as low as 7%. The results of the sampling of all of the 20 issues are given in the table. Figure 1 shows a plot of the word error percentage versus the date published. This clearly shows that the older issues are more prone to higher error rates. Looking just at the group of issues published after 1880 gives a total of 576 sample words selected with 80 words in error. The error rate is for that combined group of newspapers is 14% with confidence interval from 11% to 17%.

Assuming the worst case that the word error rate is actually 17% a search for a name in PDF files of old newspaper pages from that era would find that name, when it's there, 5 times out of 6. The actual error is probably closer to the 14% level so the name would be found 6 times out of 7. That seems to be an error rate one could live with.

How does this error rate compare with published information about old newspaper OCRing studies. Published studies talk about character error rate and mention word error only in passing. There is an interesting and informative article in D-Lib Magazine, March/April 2009 titled "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs" by Rose Holley, Manager - Australian Newspaper Digitisation Program, National Library of Australia. This article about OCRing old newspapers contains a table that speaks to good, average and poor OCR "accuracy." Here is a quote from the article:

"*The question of what is acceptable* (character hit rate or accuracy) *has not been answered, but in speaking to other libraries and OCR contractors, it was generally agreed for historic newspapers that when we talk about good, average and bad OCR we mean:*

Good OCR accuracy = 98-99% accurate (1-2% of OCR incorrect)
Average OCR accuracy = 90-98% accurate (2-10% of OCR incorrect)
Poor OCR accuracy = below 90% accurate (more than 10% of OCR incorrect)"

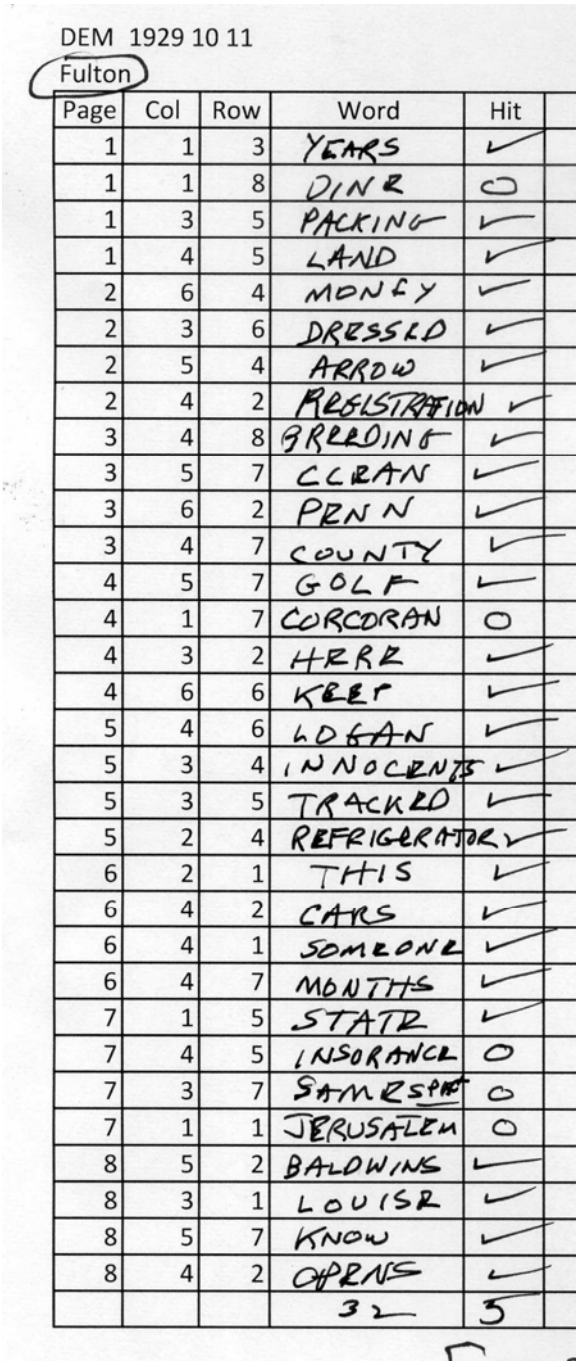
The Old Fulton results for the newer newspapers showed a 14% error rate or an 86% hit rate relative to searching for a word. How does this compare to the above figures of character hit rate? The words we searched for were on average six characters long. To get a word of six characters correct 86% of the time we need to have a character hit rate that is the 6^{th} root of 0.86 or 0.975 or a 97.5% character hit rate. That's true since to get the word correct we need to get the first character correct and the second character correct and the third character correct and so forth up through the sixth character. Since the character hit rate was calculated to be 97.5% it falls close to the article's Good OCR accuracy category.

We will use this sampling plan to monitor our word error rate as we proceed with our project and hope that we will have an error rate as low as the Old Fulton site.

Table 1
Data Concerning the Sampling Study
of 20 Penn Yan Democrat Issues on the Old Fulton Post Card Site

Penn Yan Democrat Issue Date	# of Pages in the Issue	# of Sample Words Selected per Page	# of Sample Words in the Issue	# of Word Misses (Errors) in the Sample	Word Error in %	Lower 95 % Confidence Interval	Upper 95% Confidence Interval
October 12, 1852	4	8	32	12	38	23	55
December 21, 1852	4	8	32	24	75	58	87
August 1, 1860	4	8	32	22	69	51	82
September 29, 1865	4	8	32	15	53	31	64
November 10, 1865	4	8	32	22	69	51	82
August 4, 1876	4	8	32	16	50	33	68
June 17, 1881	4	8	32	10	31	18	48
April 8, 1892	4	8	32	10	31	18	49
December 9, 1898	8	8	64	8	12	6	23
October 18, 1901	8	4	32	2	6	2	20
May 7, 1909	8	4	32	2	6	2	20
January 7, 1910	8	8	64	3	5	2	13
May 3, 1918	8	4	32	3	9	3	24
May 18, 1923	8	8	64	13	20	12	32
October 11, 1929	8	4	32	5	16	7	32
May 11, 1934	8	4	32	3	9	3	27
December 29, 1939	8	4	32	2	6	2	20
April 18, 1941	8	4	32	4	12	5	28
February 1, 1946	8	8	64	6	9	4	19
May 30, 1947	8	4	32	9	28	16	45

Table 2
Sampling Work Sheet
October 11, 1929 Issue of the Penn Yan Democrat
on the Old Fulton NY Post Card Website

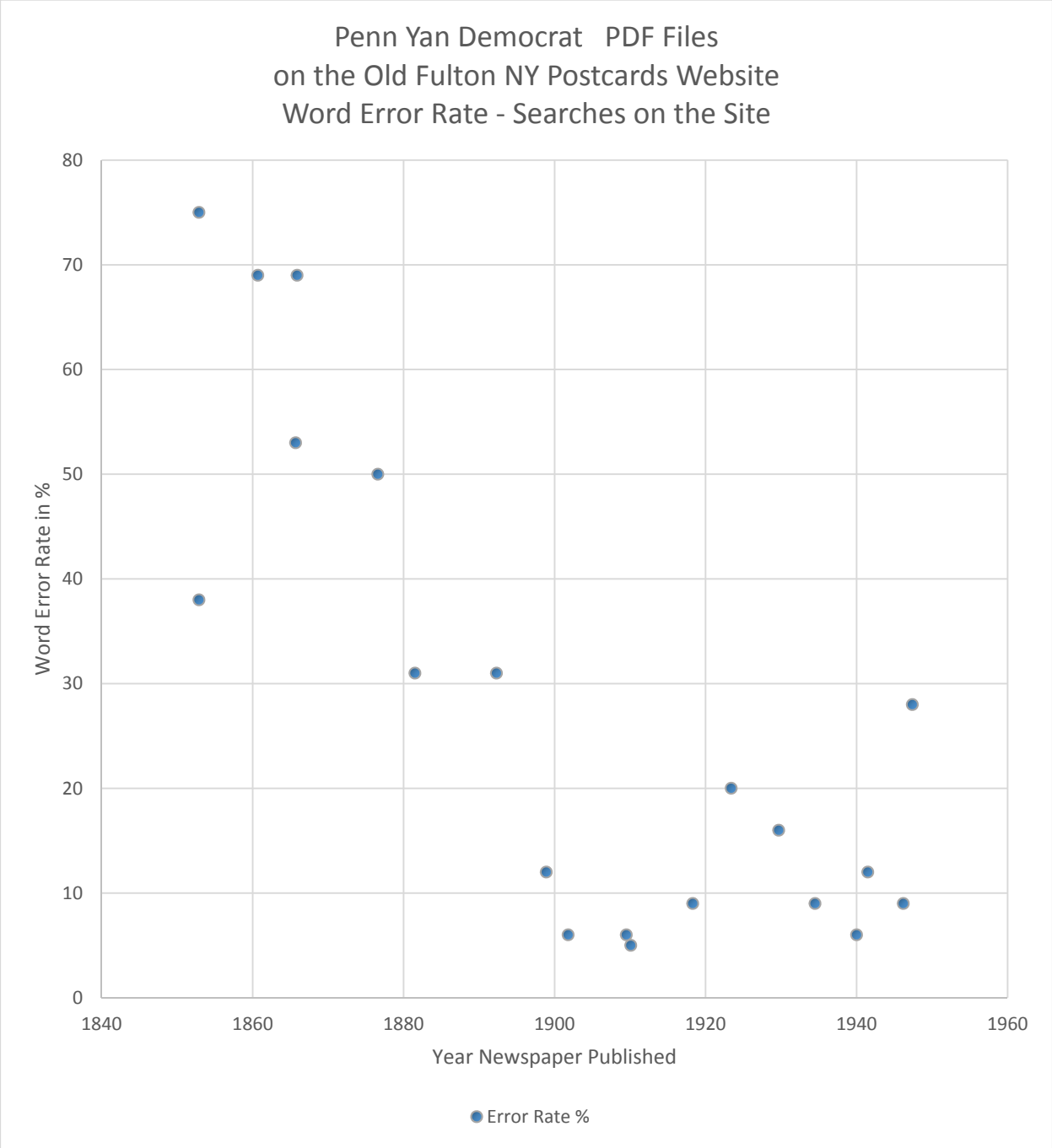


DEM 1929 10 11

Fulton

Page	Col	Row	Word	Hit
1	1	3	YEARS	✓
1	1	8	DINR	○
1	3	5	PACKING	✓
1	4	5	LAND	✓
2	6	4	MONEY	✓
2	3	6	DRESSED	✓
2	5	4	ARROW	✓
2	4	2	REGISTRATION	✓
3	4	8	BRRDING	✓
3	5	7	CCRAN	✓
3	6	2	PRN N	✓
3	4	7	COUNTY	✓
4	5	7	GOLF	✓
4	1	7	CORCORAN	○
4	3	2	HRRR	✓
4	6	6	KRRP	✓
5	4	6	LOGAN	✓
5	3	4	INNOCENTS	✓
5	3	5	TRACKED	✓
5	2	4	REFRIGERATOR	✓
6	2	1	THIS	✓
6	4	2	CARS	✓
6	4	1	SOMEONE	✓
6	4	7	MONTHS	✓
7	1	5	STATE	✓
7	4	5	INSORANCE	○
7	3	7	SAMERSET	○
7	1	1	JERUSALEM	○
8	5	2	BALDWINS	✓
8	3	1	LOUISE	✓
8	5	7	KNOW	✓
8	4	2	OPENS	✓
			32	3

Figure 1


Penn Yan Democrat PDF Files
on the Old Fulton NY Postcards Website
Word Error Rate - Searches on the Site
Word Error Rate in %
80
70
60
50
40
30
20
10
0
1840
1860
1880
1900
1920
1940
1960
Year Newspaper Published
Error Rate %