An Evaluation of OCR Accuracy

Stephen V. Rice, Junichi Kanai and Thomas A. Nartker

1 Introduction

ISRI has conducted its second annual assessment of the accuracy of devices for optical character recognition (OCR) of machine-printed, English-language documents. This year's test featured more devices, more data, and more measures of performance than the test conducted a year ago [Rice 92a].

2 Devices

ISRI has attempted to acquire one copy of every OCR technology available. Only one version was tested from each vendor, but vendors were allowed to submit their latest, most accurate version. In many cases, this was a "pre-release" or "beta" version. The deadline for submissions was January 31, 1993. Table 1 lists the versions that were evaluated.

Vendor	Version Name	Version #
Caere Corp.	Caere OCR	109
Calera Recognition Systems, Inc.	Calera MM600	mm24su
Cognitive Technology Corp.	Cognitive Cuneiform	0.8
CTA, Inc.	CTA TextPert DTK	1.2.9
ExperVision, Inc.	ExperVision RTK	2.0
OCRON, Inc.	OCRON Recore	2.0.5
Recognita Corp. of America	Recognita Plus DTK	2.0β.BC3
Xerox Imaging Systems, Inc.	XIS ScanWorX API	2.0β3

Table 1: Participating Vendors and Versions Submitted

Each vendor was required to submit a version that could be operated in an entirely automatic (non-interactive) way. Hence, each provided a "toolkit" that allows this, which may be reflected in the version name; for example, "DTK" refers to the "Developer's ToolKit."

Both Caere Corp. and Xerox Imaging Systems submitted a version for the Sun SPARCstation; all others submitted a PC DOS version.

3 Data

The data used in the test consisted of 500 pages selected at random from a collection of approximately 2,500 documents containing 100,000 pages. The documents in this collection were chosen by the U.S. Department of Energy (DOE) to represent the kinds of documents from which the DOE plans to build large, full-text retrieval databases using OCR for document conversion. The documents are mostly scientific and technical papers [Nartker 92].

The pages in the random sample are quite diverse. There is a considerable variety of typefaces and type sizes, and page quality ranges from "perfect" originals to illegible photocopies. There are no fax or dot-matrix pages.

Twenty-five of the 500 pages are so degraded that they are essentially unreadable by humans; these pages were excluded from the sample since no OCR device could be expected to decipher them. Another 15 pages containing no text were excluded, bringing the sample size to 460.

Each page was scanned at 300 dpi using a Fujitsu M3096E+ scanner to produce a binary image. The default threshold was used for each page.

Every page image was manually "zoned," i.e., rectangular regions containing text were delineated. All text on a page was zoned, including "main body" text, tables, captions, footnotes, and page headers and footers. The only text that was not zoned were mathematical equations, and text that is part of a figure, such as labels on the axes of a graph, or location names on a map.

Correct text was manually prepared corresponding to each zone. Considerable time and effort was expended to make certain that the correct text is in fact correct. It is believed that this text is at least 99.99% accurate, i.e., less than one error per 10,000 characters.

A total of 1,313 zones were defined on the 460 pages, containing a total of 817,946 characters. See Table 2 for a breakdown by zone type. This sample is about three times the size of the sample used in last year's evaluation [Rice 92a]. It is also considerably more diverse because last year's sample was drawn from a small subset (about 10%) of this document collection.

Zone Type	# Zones	# Characters
"Main body" Text	512	667,161
Table	133	99,839
Caption	125	18,042
Footnote	67	13,981
Header/Footer	448	7,053
Other Text	28	11,870
Total	1,313	817,946

Table 2: Sample Data by Zone Type

4 Methodology

Each device processed the same zoned portions of the same binary images. This processing was carried out in an entirely automated manner, i.e., there was no human interaction with the devices.

No interactive "learning" modes were used, and no device received any special training.

For each device supporting an optional "system" lexicon, this lexicon was enabled. No "user-defined" lexicon was utilized.

"De-columnization" was disabled for each device since it was not needed. With the exception of "table" zones, each zone contains only a single column.

The identification and tabulation of errors was performed entirely under computer control. Generated text was matched with the correct text using a difference algorithm based on the detection of long common substrings [Rice 92b].

Horizontal and vertical spacing is compressed prior to applying this algorithm. Blank lines, and leading and trailing blanks on a line, are eliminated. Consecutive blanks within a line are compressed to a single blank.

Non-ASCII symbols appearing on a page, such as a bullet symbol (•) or a Greek letter (γ), require special handling. Each non-ASCII symbol is represented in the correct text by a tilde ($\sim$). Since the OCR devices are not expected to recognize these symbols, the error-counting software allows zero or one arbitrary character to be generated for each tilde without charging an error.

All software tools used in this evaluation are part of the ISRI OCR Experimental Environment [Rice 93].

5 Character Accuracy

Each character insertion, substitution or deletion required to correct the generated text is counted as an error. This metric is attributed to Levenshtein [Levenshtein 66]; the number of errors has been termed *edit distance* by Wagner and Fischer [Wagner 74].

Character accuracy is defined as

$$\frac{n - (\#errors)}{n}$$

where n is the number of correct characters.

Table 3 shows the number of errors made by each device, and the corresponding character accuracy, in processing the 460-page sample containing 817,946 characters.

	# Errors	% Accuracy
Caere OCR	24,074	97.06
Calera MM600	16,013	98.04
Cognitive Cuneiform	42,354	94.82
CTA TextPert DTK	43,964	94.63
ExperVision RTK	15,186	98.14
OCRON Recore	43,159	94.72
Recognita Plus DTK	36,250	95.57
XIS ScanWorX API	16,750	97.95

Table 3: Character Accuracy for the Entire Sample

5.1 Confusions

Generating the letter "c" when the correct character is an "e" is an example of a *confusion*; one error is charged for this confusion because it can be corrected by one character substitution. If "rn" is generated for the letter "m," two errors are charged, since this confusion requires one substitution and one deletion.

Table 4 shows the 60 most common confusions based on the median number of occurrences for the eight devices. The most common confusion was an introduced space, which causes a word to be split (e.g., "Neva da"). The fourth most common confusion was a missing space, which causes two words to be joined (e.g., "Universityof").

5.2 Page Quality

The median character accuracy achieved by the eight devices in processing a given page is a good measure of the quality of the page, or at least its "OCR difficulty." The 460 pages were sorted by this measure and divided into five "Page Quality Groups" containing approximately the same number of characters in each. Group 1 contains the pages with the highest median accuracy, and Group 5 contains the pages with the lowest median accuracy (see Table 5). Figures 1-5 show examples of page images from the middle of each group.

Table 6 shows the number of errors made by the devices in each Page Quality Group, and Table 7 shows the corresponding character accuracies. Graph 1 displays a graph of this data. It is interesting to note that about 70% of the total errors are made on the worst 20% of the sample, i.e., Group 5.

	Correct	Generated
1		space
2	e	c
3	,	.
4	space	
5	1	l
6	i	l
7	O	0
8	0	O
9	1	I
10	a	s
11		.
12	l	i
13	.	,
14	c	e
15	a	e
16	f	t
17	l	1
18	t	c
19		’
20	.	

	Correct	Generated
21	y	v
22	a	o
23	s	a
24	i	t
25	i	I
26	t	l
27	M	N
28	a	
29	5	S
30	-	
31	.	space
32	0	8
33		,
34	I	l
35	e	a
36	g	q
37	l	I
38	e	o
39	0	o
40	o	c

	Correct	Generated
41	M	H
42	h	b
43	i	
44	s	e
45	a	n
46	m	rn
47	D	O
48	,	
49	S	s
50	b	h
51	;	:
52	f	l
53	m	n
54	o	a
55	D	0
56	e	
57	h	n
58	o	O
59	rn	m
60	O	o

Table 4: Most Common Confusions

Page Quality Group	# Pages	# Characters	Median % Accuracy
1	80	165,110	99.69-100.00
2	77	163,019	99.31- 99.69
3	85	162,367	98.46- 99.30
4	96	163,176	96.58- 98.45
5	122	164,274	0.00- 96.57
Total	460	817,946	

Table 5: Page Quality Groups

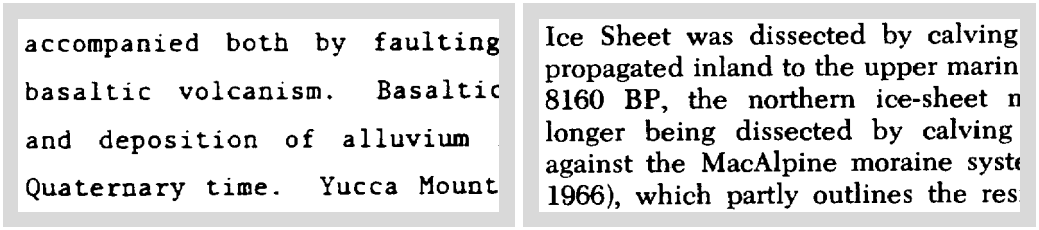


Figure 1: Examples from Page Quality Group 1

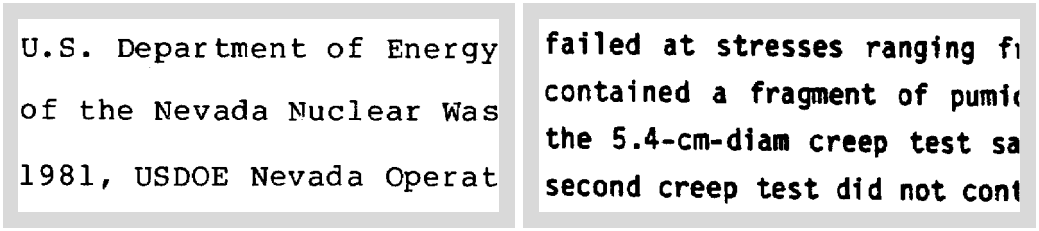


Figure 2: Examples from Page Quality Group 2

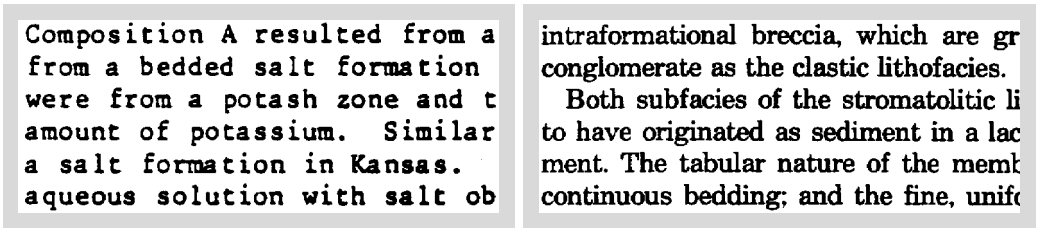


Figure 3: Examples from Page Quality Group 3

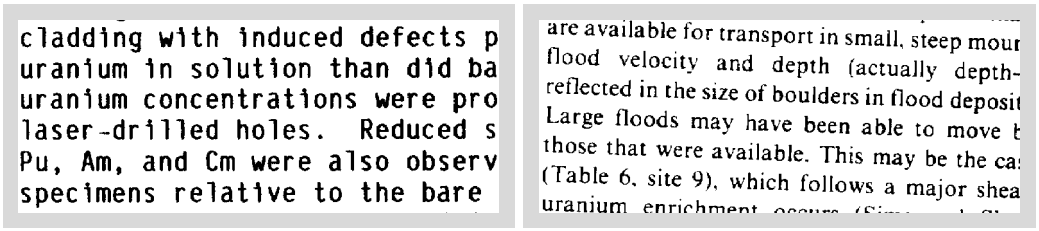


Figure 4: Examples from Page Quality Group 4

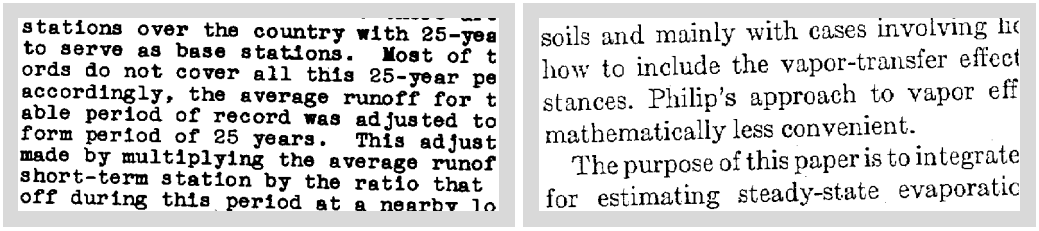


Figure 5: Examples from Page Quality Group 5

	Group 1	Group 2	Group 3	Group 4	Group 5
Caere OCR	202	596	1,281	3,234	18,761
Calera MM600	133	722	1,124	2,519	11,515
Cognitive Cuneiform	1,217	2,518	4,495	9,945	24,179
CTA TextPert DTK	802	2,748	4,305	8,794	27,315
ExperVision RTK	163	477	1,072	2,211	11,263
OCRON Recore	545	1,880	3,203	7,130	30,401
Recognita Plus DTK	539	1,722	2,912	5,482	25,595
XIS ScanWorX API	281	695	1,411	3,240	11,123

Table 6: Number of Errors in Each Page Quality Group

	Group 1	Group 2	Group 3	Group 4	Group 5
Caere OCR	99.88	99.63	99.21	98.02	88.58
Calera MM600	99.92	99.56	99.31	98.46	92.99
Cognitive Cuneiform	99.26	98.46	97.23	93.91	85.28
CTA TextPert DTK	99.51	98.31	97.35	94.61	83.37
ExperVision RTK	99.90	99.71	99.34	98.65	93.14
OCRON Recore	99.67	98.85	98.03	95.63	81.49
Recognita Plus DTK	99.67	98.94	98.21	96.64	84.42
XIS ScanWorX API	99.83	99.57	99.13	98.01	93.23

Table 7: Character Accuracy for Each Page Quality Group

6 Word Accuracy

In a text retrieval application, the correct recognition of words is much more important than the correct recognition of numbers or punctuation. We define a *word* to be any sequence of one or more letters. If m out of n words are recognized correctly, the *word accuracy* is m/n. Since full-text searching is almost always performed on a case-insensitive basis, we consider a word to be correctly recognized even if one or more letters of the generated word are in the wrong case (e.g., "transPortatIon").

The 460-page sample contains 119,497 words. Table 8 shows the number of misrecognized words and the word accuracy for each device.

Graph 2 displays the word accuracies for each Page Quality Group. As page quality declines, word accuracy drops dramatically.

6.1 Stopwords and Non-stopwords

In text retrieval, common words such as "the," "of," "and," "in," etc., are normally not indexed because they provide essentially no retrieval value, and they substantially increase the overhead of maintaining the index. These words are termed *stopwords*; we refer to all other words as *non-stopwords*. *Non-stopword accuracy* is even more relevant than word accuracy to a text retrieval application.

We make use of the default set of 110 stopwords provided by the BASISplus text retrieval product [IDI 90]. Using this definition of what is a stopword, it was determined that 42,116, or 35%, of the words in the sample are stopwords.

Table 9 shows the stopword accuracy and non-stopword accuracy for each device. Not surprisingly, the stopword accuracy is significantly higher; stopwords are short and well-known, and should be part of every

	# Misrecognized Words	Word % Accuracy
Caere OCR	7,187	93.99
Calera MM600	4,100	96.57
Cognitive Cuneiform	10,472	91.24
CTA TextPert DTK	13,579	88.64
ExperVision RTK	4,213	96.47
OCRON Recore	14,710	87.69
Recognita Plus DTK	11,842	90.09
XIS ScanWorX API	5,466	95.43

Table 8: Word Accuracy for the Entire Sample

device's lexicon.

	Stopword % Accuracy	Non-stopword % Accuracy
Caere OCR	96.72	92.50
Calera MM600	98.23	95.66
Cognitive Cuneiform	95.28	89.04
CTA TextPert DTK	94.06	85.69
ExperVision RTK	98.51	95.37
OCRON Recore	92.73	84.95
Recognita Plus DTK	94.55	87.66
XIS ScanWorX API	98.15	93.94

Table 9: Stopword and Non-stopword Accuracy for the Entire Sample

Graph 3 is similar to Graph 2, but displays the non-stopword accuracies for each Page Quality Group, and an even more dramatic drop in accuracy as page quality declines.

Graph 4 shows a plot of the character accuracy vs. the non-stopword accuracy achieved by the devices for each Page Quality Group, and for the entire sample. The relationship between character accuracy and non-stopword accuracy is almost linear; for every 1% drop in character accuracy, there is roughly a 2% decline in non-stopword accuracy.

6.2 Word Length

We are interested in how the length of a word, in number of characters, affects word accuracy. Graphs 5a-5h show, for each device, the stopword and non-stopword accuracy for each word length.

In general, the accuracy is quite low for very short non-stopwords, i.e., lengths 1 to 3. This may be due to the lack of contextual clues provided by such short words. Most of these words are actually abbreviations. The accuracy is especially low for non-stopwords of length two. Perhaps this is due to an over-reliance on "bi-gram" data, as many of these words are uncommon character-pairs. Examples of two-character non-stopwords include "km," "ft," "cm," "Mr," "pH," *et al.*

For longer non-stopwords, some devices show a significant decline in accuracy as the length increases, while others, perhaps taking better advantage of contextual information, perform equally well on words of length 4 through 12.

7 Marked Characters

OCR devices provide the end-user with some help in correcting errors. A device generates a *reject character*, usually a tilde (∼), when it is unable to recognize a character. If it is able to recognize the character, but has low confidence in its decision, it generates the character preceded by a *suspect marker* (∧ is often used). These special characters in the output draw the attention of the end-user to potential errors. We refer to reject characters, and characters marked as suspect, as *marked characters*.

Example:

This sentenc∼ conta∧lns both reject charact∼rs an∧d suspect markars.

We define a *marked error* to be any error that can be identified by examining marked characters.[1] We define a *false mark* to be any correctly-generated character that is marked as suspect. In the example, there are three marked errors (the two reject characters and the marked "l" in "contains"), one unmarked error (the second "a" in "markers"), and one false mark (the marked "d" in "and").

The objective is to provide an efficient mechanism for correcting OCR errors. If too many characters are marked, the process of examining these characters and correcting the identified errors could be as tedious (and costly) as proofreading the entire page. If too few characters are marked, this process fails to correct a large percentage of the errors. Clearly, the goal is to mark as many of the errors as possible, while minimizing the number of false marks.

Five of the devices support more than one level of suspect markers; the end-user can select a discrete level, or choose a value on a continuous scale, to increase or decrease the number of characters that are marked. One of the devices (Calera MM600) supports only a single level, and two of the devices (CTA TextPert DTK and Recognita Plus DTK) provide no support for suspect markers. All of the devices generate reject characters.

In Graph 6, we present a picture of "marked character efficiency." For each device, there is a curve determined by 2 to 5 points, depending on the number of suspect levels supported. Each point shows the character accuracy after correcting the marked errors for a certain percentage of marked characters.

The first point on the curve shows the base character accuracy, and the second point indicates the character accuracy if only reject characters are examined. The third point shows the character accuracy if

[1] More precisely, in terms of the difference algorithm [Rice 92b], if an unmatched substring of the generated text contains at least one marked character, then all errors attributed to that substring are considered to be marked.

both reject characters and "level 1" suspect markers are examined. Similarly, the fourth and fifth points correspond to "level 2" and "level 3" suspect markers.

It is not surprising that the slope of the curve is highest between the first two points. Correcting the errors identified by reject characters is a very efficient operation, since a false mark occurs only when the correct character happens to be a tilde. But this corrects only 10-30% of the errors.

In general, examining the first level of suspect markers appears to be quite efficient, but the second and third levels are much less useful, as the flattening of the curve indicates.

The effect of page quality on the shape of these curves, and on the percentage of marked characters, can be seen in Graphs 7a-7h.

The DOE wishes to build full-text databases that are at least 99.8% accurate. Using the most accurate available device, with a base character accuracy of 98.1%, it will be necessary to correct 90% of its errors to reach this level of accuracy. But only 60-65% of its errors are marked; correcting these will bring the accuracy to 99.3%.

The manual correction of the unmarked errors will certainly be costly. But research is underway at ISRI in the automatic correction of OCR errors. It has been demonstrated that by applying a voting algorithm to the outputs of multiple OCR devices, at least 40% of the errors can be corrected automatically [Handley 91, Bradford 91, Rice 92a]. This work is being extended to make use of suspect markers produced by the devices, and to generate improved suspect markers in the voting output. Another project involves the automatic correction of misspelled words using word clusters [Taghva 93].

8 Conclusion

This evaluation presents a picture of the state-of-the-art in OCR when processing the DOE data. It must be emphasized that a different picture of performance may result from processing other types of documents. ISRI plans to prepare test sets from other document collections for use in upcoming evaluations.

For the DOE data, the recognition of poor quality pages represents a major hurdle, as 70% of the errors are made on the worst 20% of the data. Non-stopword accuracy, which is particularly relevant to text retrieval, is especially low on these pages, ranging from 57% to 86% for the eight devices.

For many applications, the cost of post-editing dominates the cost of document conversion. We measured the degree to which marked characters facilitate the correction of OCR errors.

Finally, we state emphatically that ISRI does not endorse any particular device or devices. The purpose of this evaluation is to provide end-users, vendors and researchers with a greater understanding of the performance and behavior of current OCR devices.

References

[Bradford 91] R. Bradford and T. Nartker, "Error Correlation in Contemporary OCR Systems," *Proc. First International Conference on Document Analysis and Recognition*, Saint-Malo, France, September 1991.

[Handley 91] J. C. Handley and T. B. Hickey, "Merging Optical Character Recognition Outputs for Improved Accuracy," *Proc. RIAO 91 Conference*, Barcelona, Spain, April 1991.

[IDI 90] BASISplus *Database Administration Reference, Release L*, Information Dimensions, Inc., Dublin, Ohio, June 1990.

[Levenshtein 66] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals," *Soviet Phys. Dokl.*, vol. 10, no. 8, pp. 707-710, 1966.

[Nartker 92] T. A. Nartker, R. B. Bradford, and B. A. Cerny, "A Preliminary Report on UNLV/GT1: A Database for Ground-Truth Testing in Document Analysis and Character Recognition," *Proc. First Symposium on Document Analysis and Information Retrieval*, Las Vegas, Nevada, March 1992.

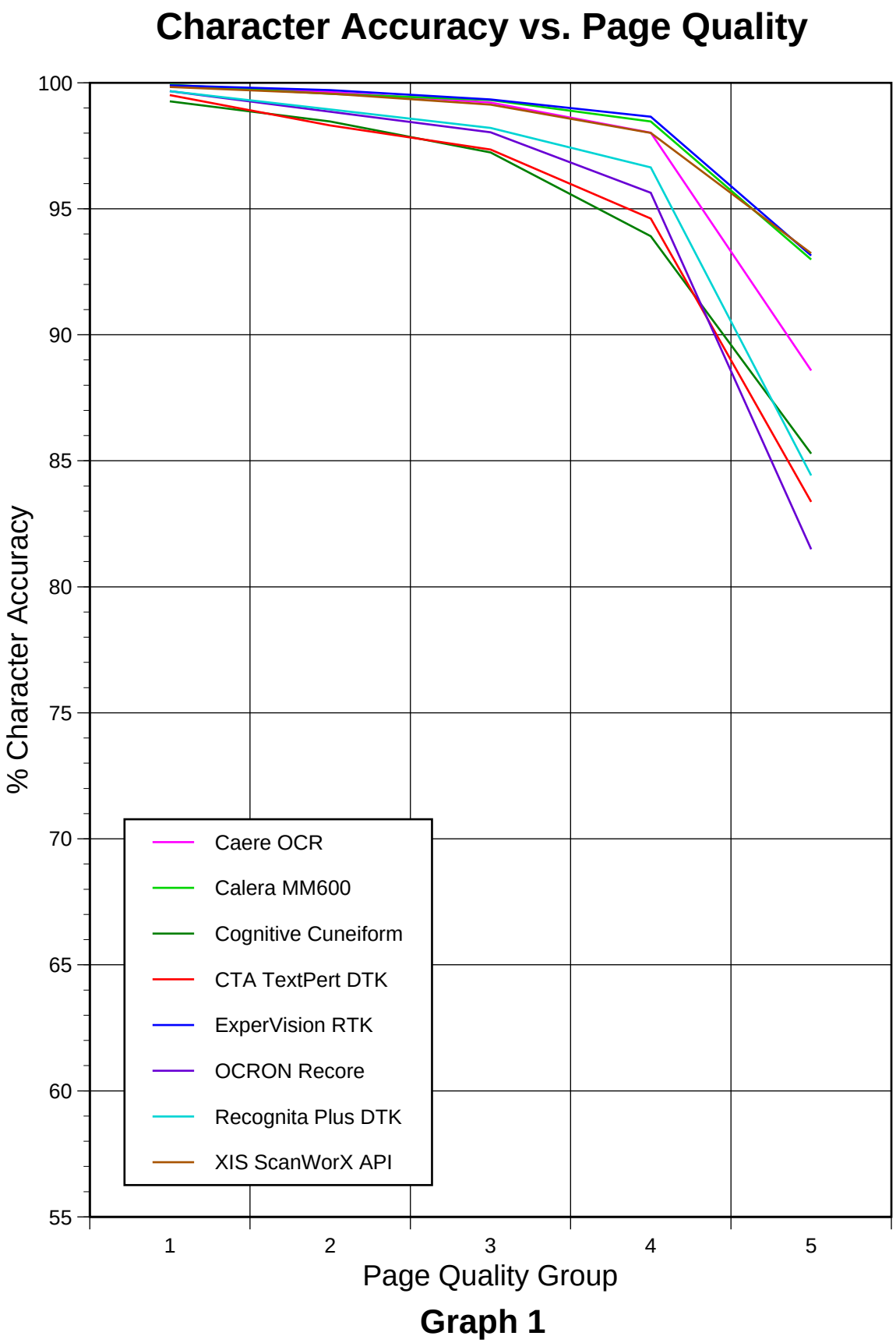
[Rice 92a] S. V. Rice, J. Kanai, and T. A. Nartker, "A Report on the Accuracy of OCR Devices," Technical Report ISRI TR-92-02, University of Nevada, Las Vegas, March 1992.

[Rice 92b] S. V. Rice, J. Kanai, and T. A. Nartker, "A Difference Algorithm for OCR-Generated Text," *Proc. IAPR Workshop on Structural and Syntactic Pattern Recognition*, Bern, Switzerland, August 1992.

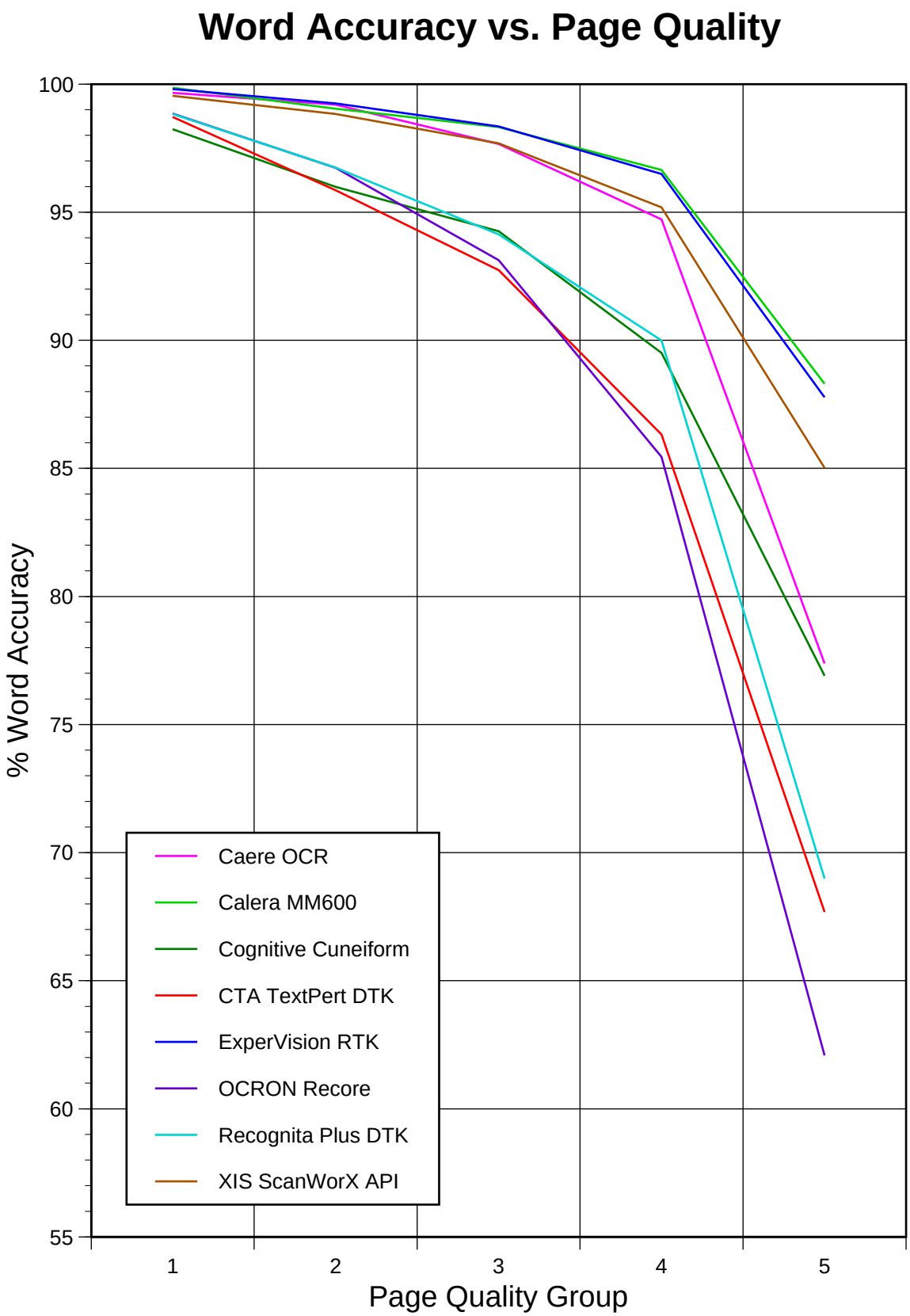
[Rice 93] S. V. Rice, "The OCR Experimental Environment, Version 3," in this report.

[Taghva 93] K. Taghva, J. Borsack, B. Bullard, and A. Condit, "Post-Editing through Approximation and Global Correction," in this report.

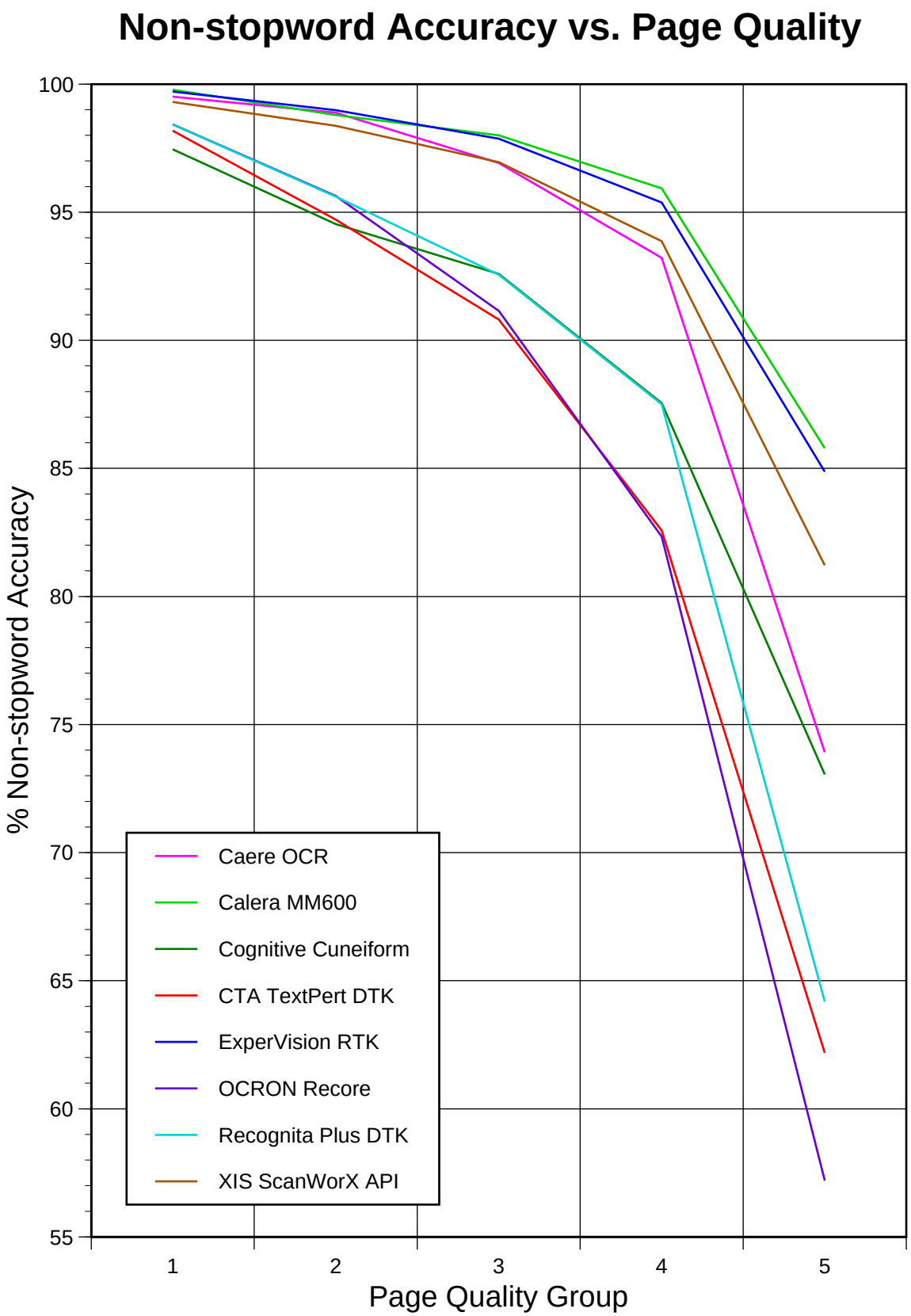
[Wagner 74] R. A. Wagner and M. J. Fischer, "The String-to-String Correction Problem," *Journal of the Association for Computing Machinery*, vol. 21, no. 1, pp. 168-173, 1974.


Character Accuracy vs. Page Quality
100
95
90
85
80
75
70
65
60
55
% Character Accuracy
1
2
3
4
5
Page Quality Group
Caere OCR
Calera MM600
Cognitive Cuneiform
CTA TextPert DTK
ExperVision RTK
OCRON Recore
Recognita Plus DTK
XIS ScanWorX API

Graph 1

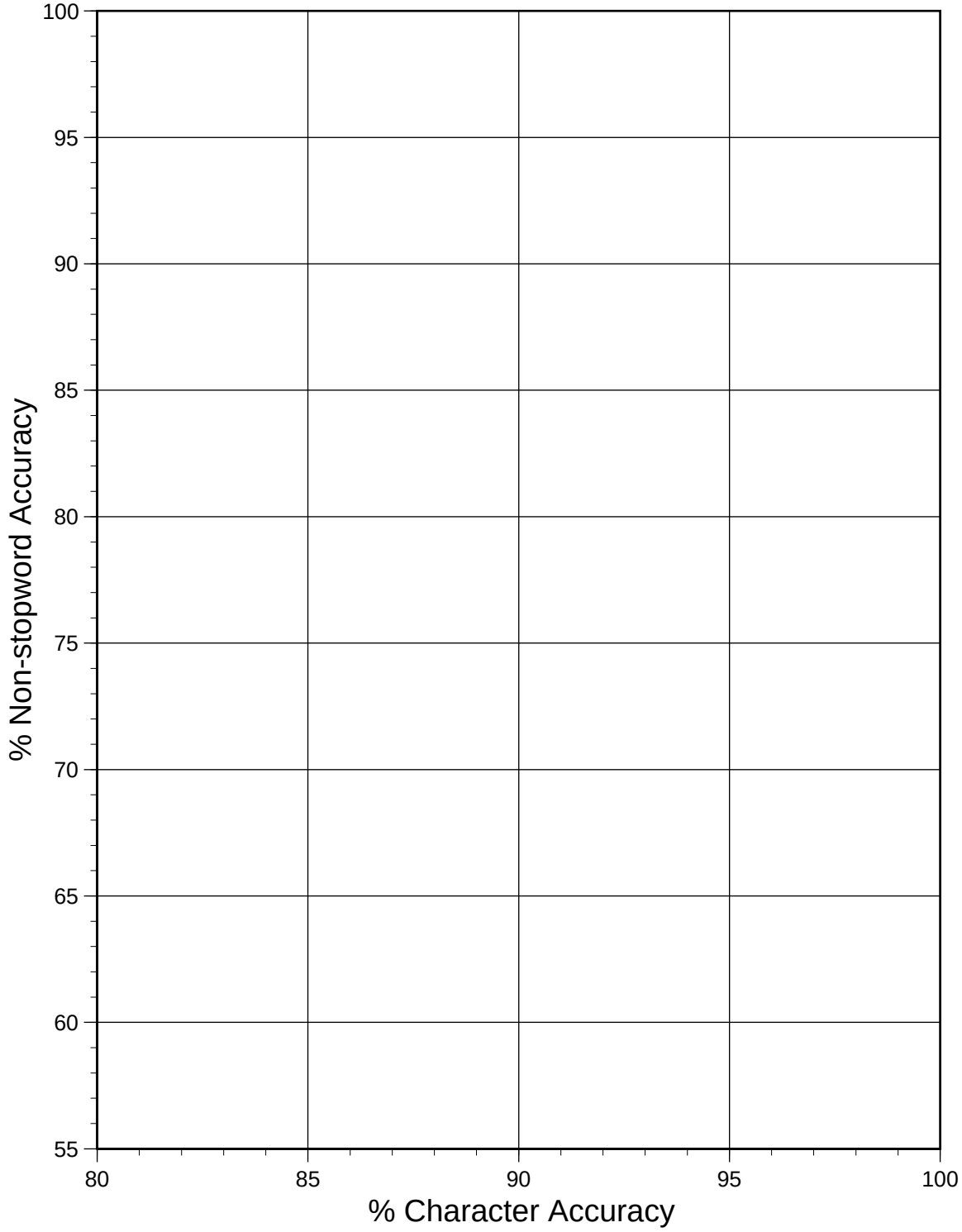

Word Accuracy vs. Page Quality
% Word Accuracy
100
95
90
85
80
75
70
65
60
55
1
2
3
4
5
Page Quality Group
Caere OCR
Calera MM600
Cognitive Cuneiform
CTA TextPert DTK
ExperVision RTK
OCRON Recore
Recognita Plus DTK
XIS ScanWorX API

Graph 2

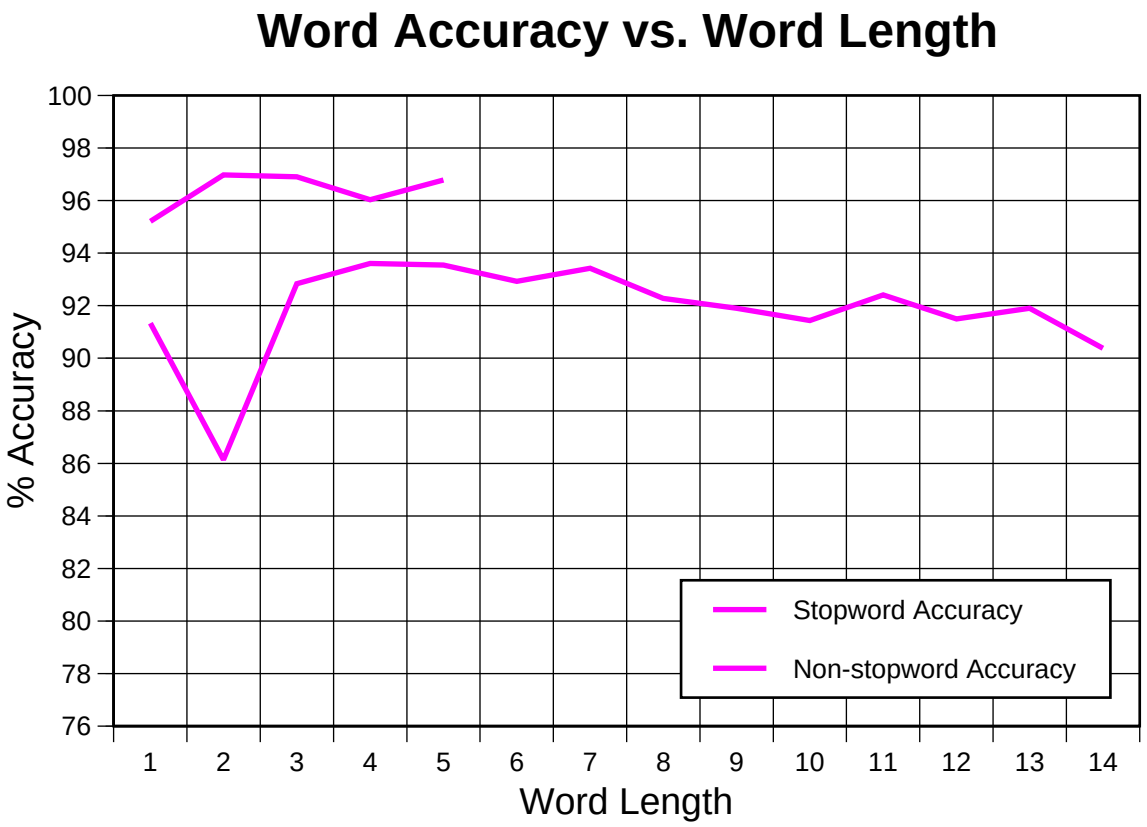

Non-stopword Accuracy vs. Page Quality
% Non-stopword Accuracy
100
95
90
85
80
75
70
65
60
55
1
2
3
4
5
Page Quality Group
Caere OCR
Calera MM600
Cognitive Cuneiform
CTA TextPert DTK
ExperVision RTK
OCRON Recore
Recognita Plus DTK
XIS ScanWorX API

Graph 3

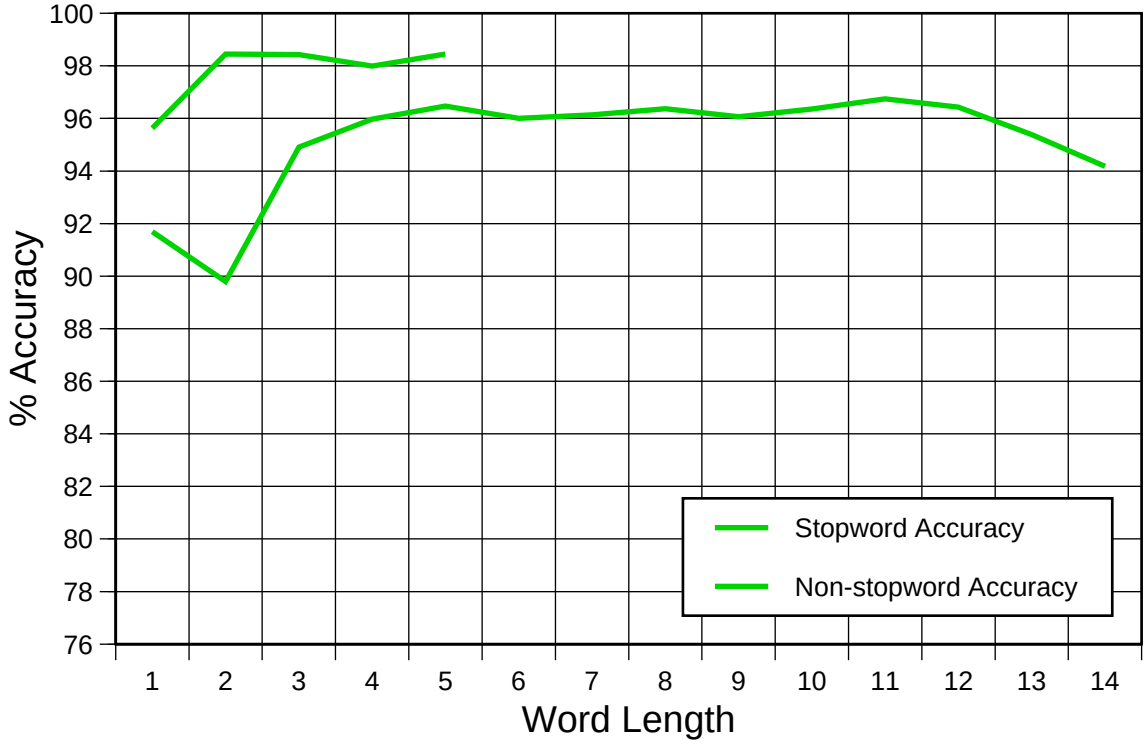
Relationship Between Character Accuracy and Non-stopword Accuracy



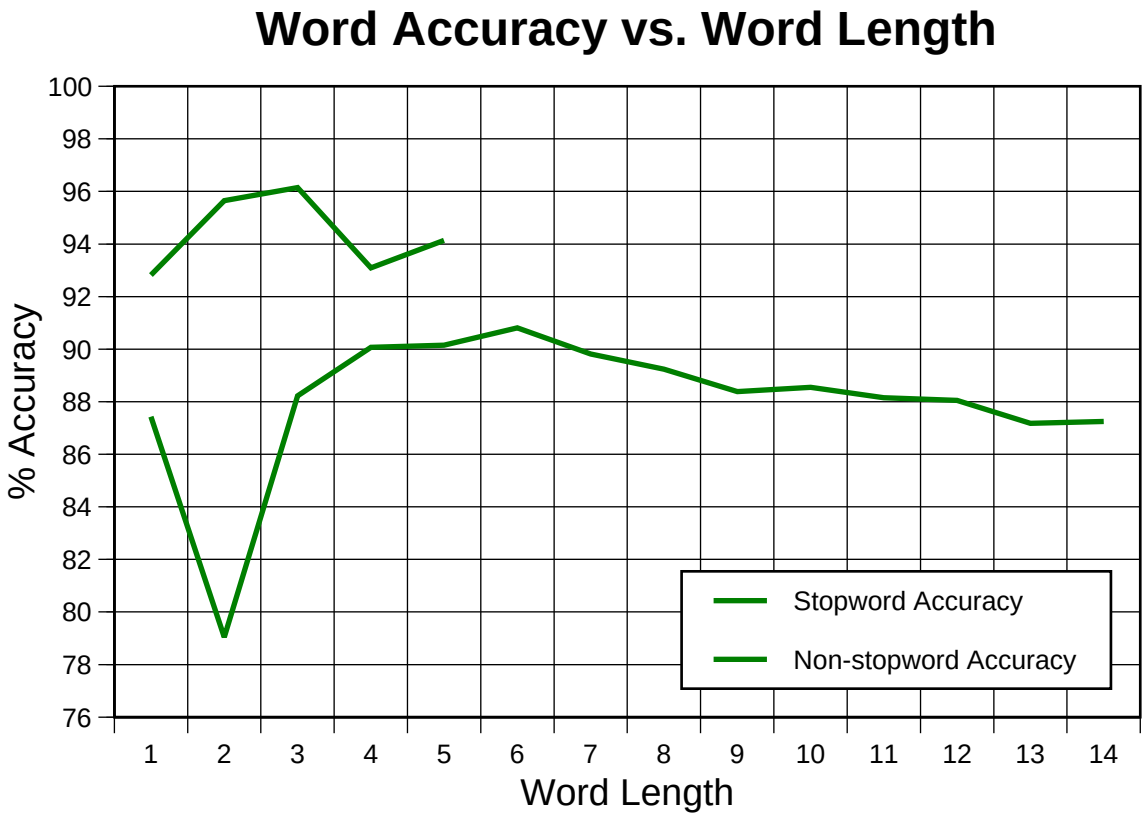
Graph 4



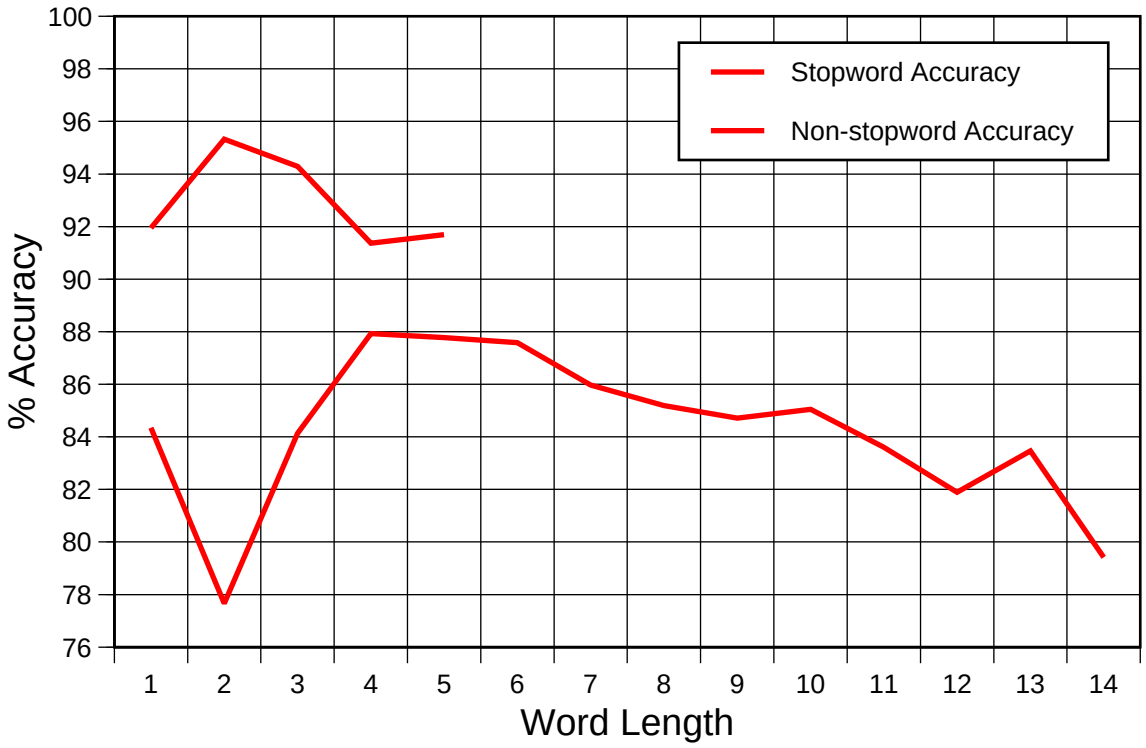
Graph 5a: Caere OCR



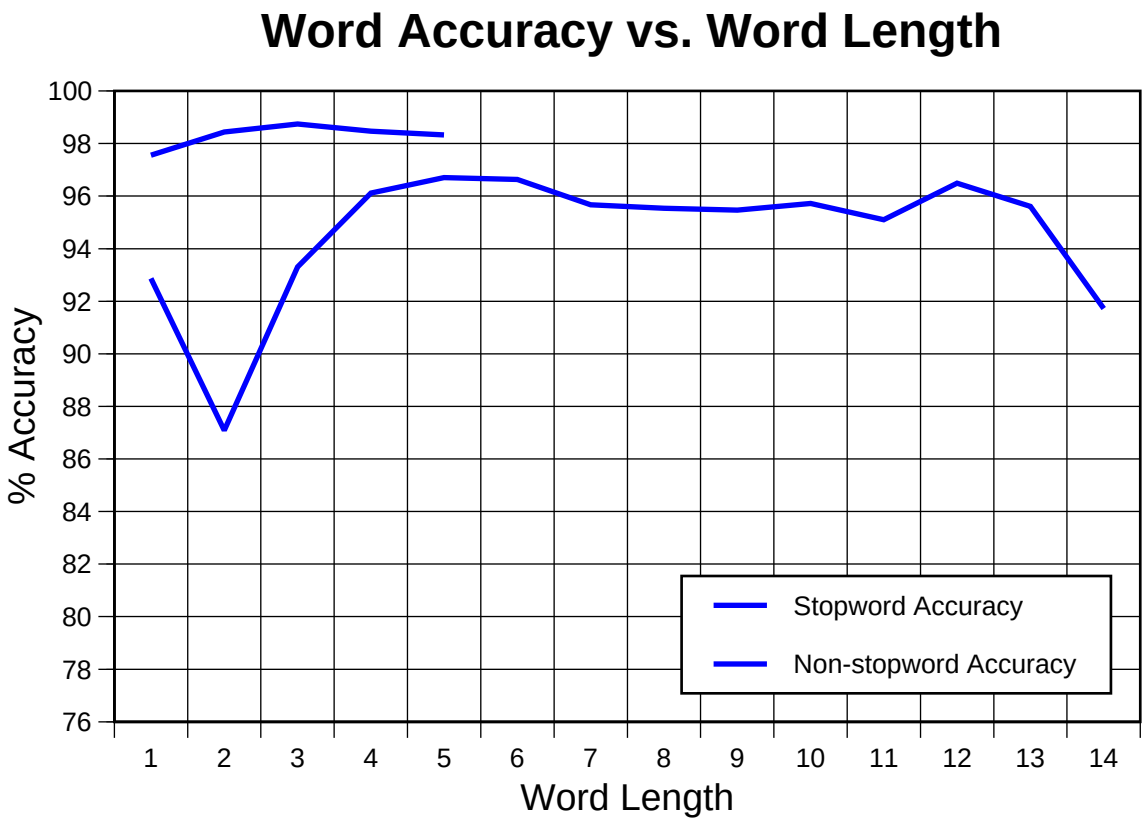
Graph 5b: Calera MM600



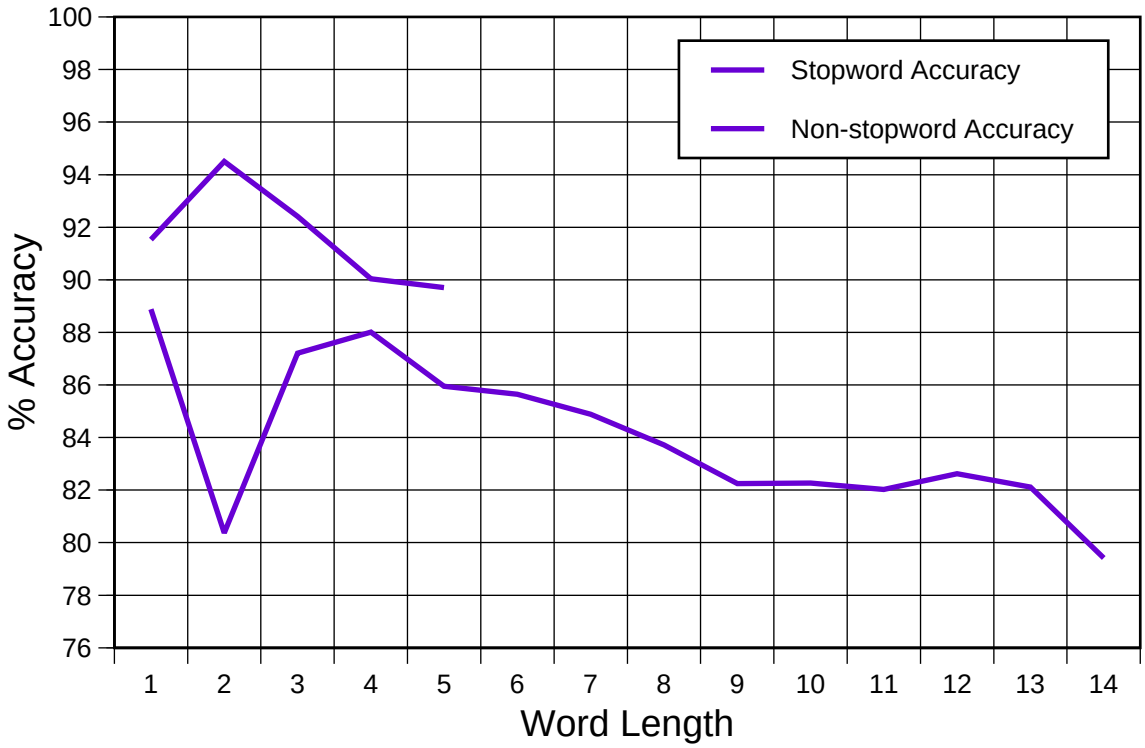
Graph 5c: Cognitive Cuneiform



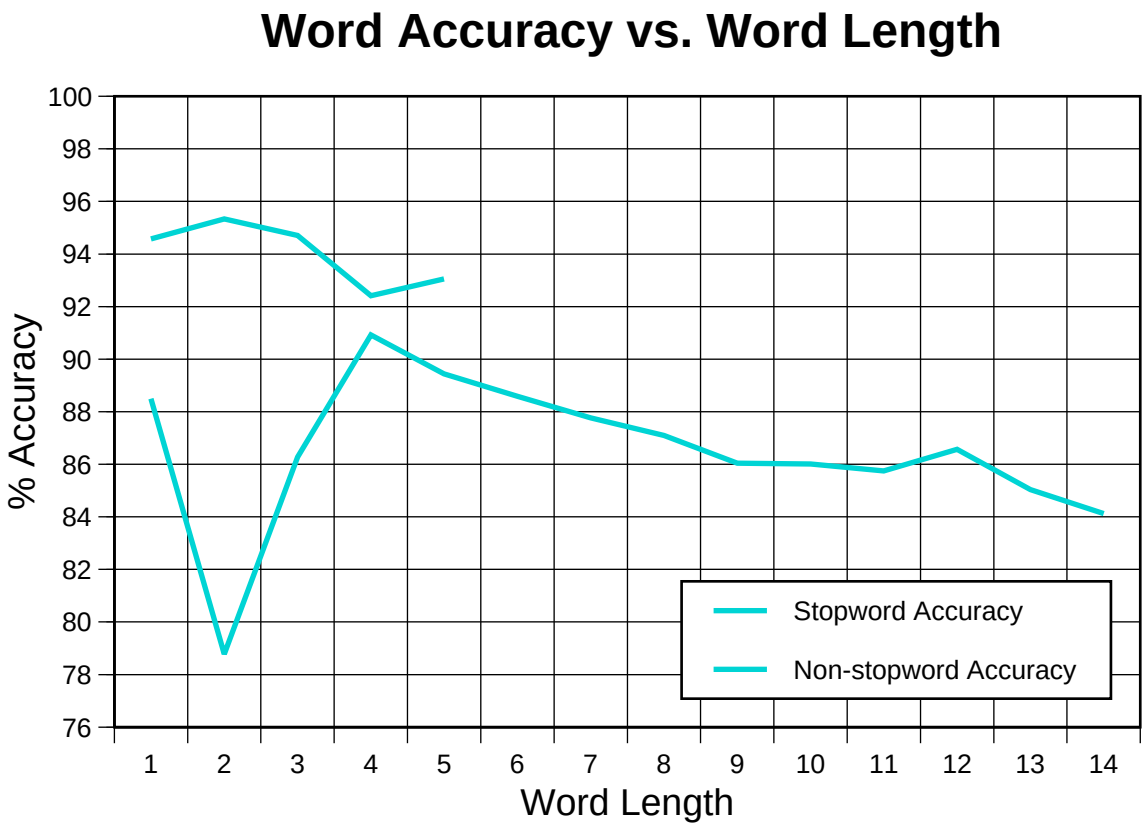
Graph 5d: CTA TextPert DTK



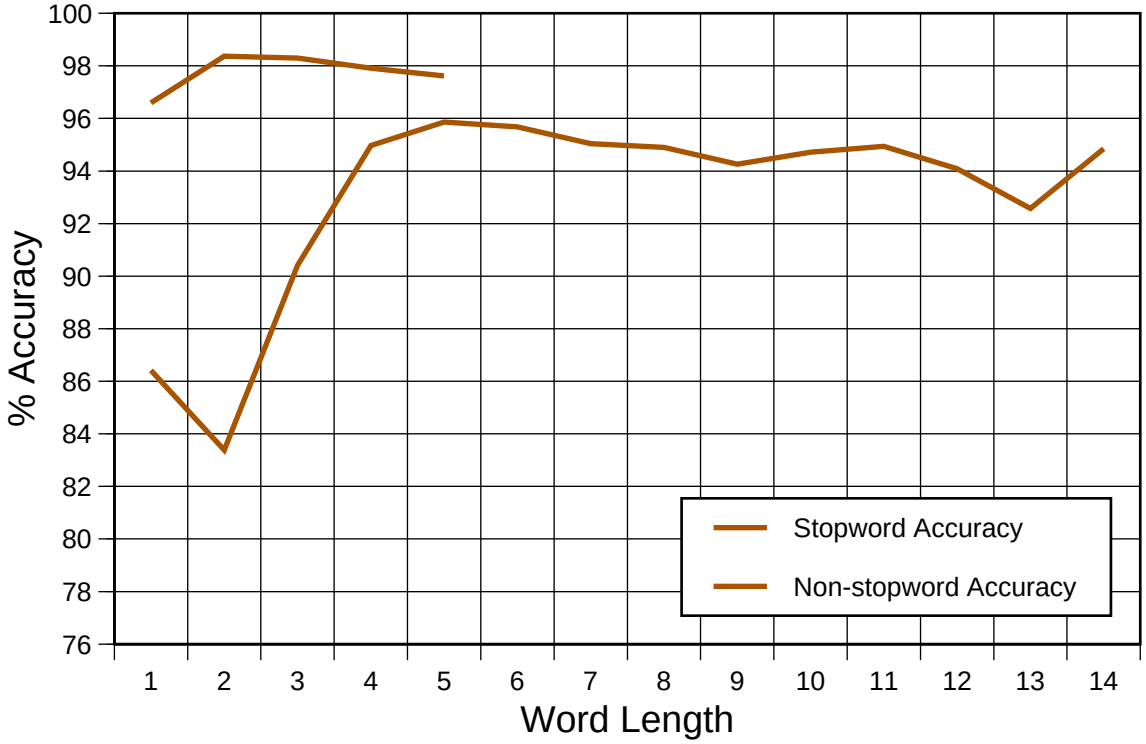
Graph 5e: ExperVision RTK



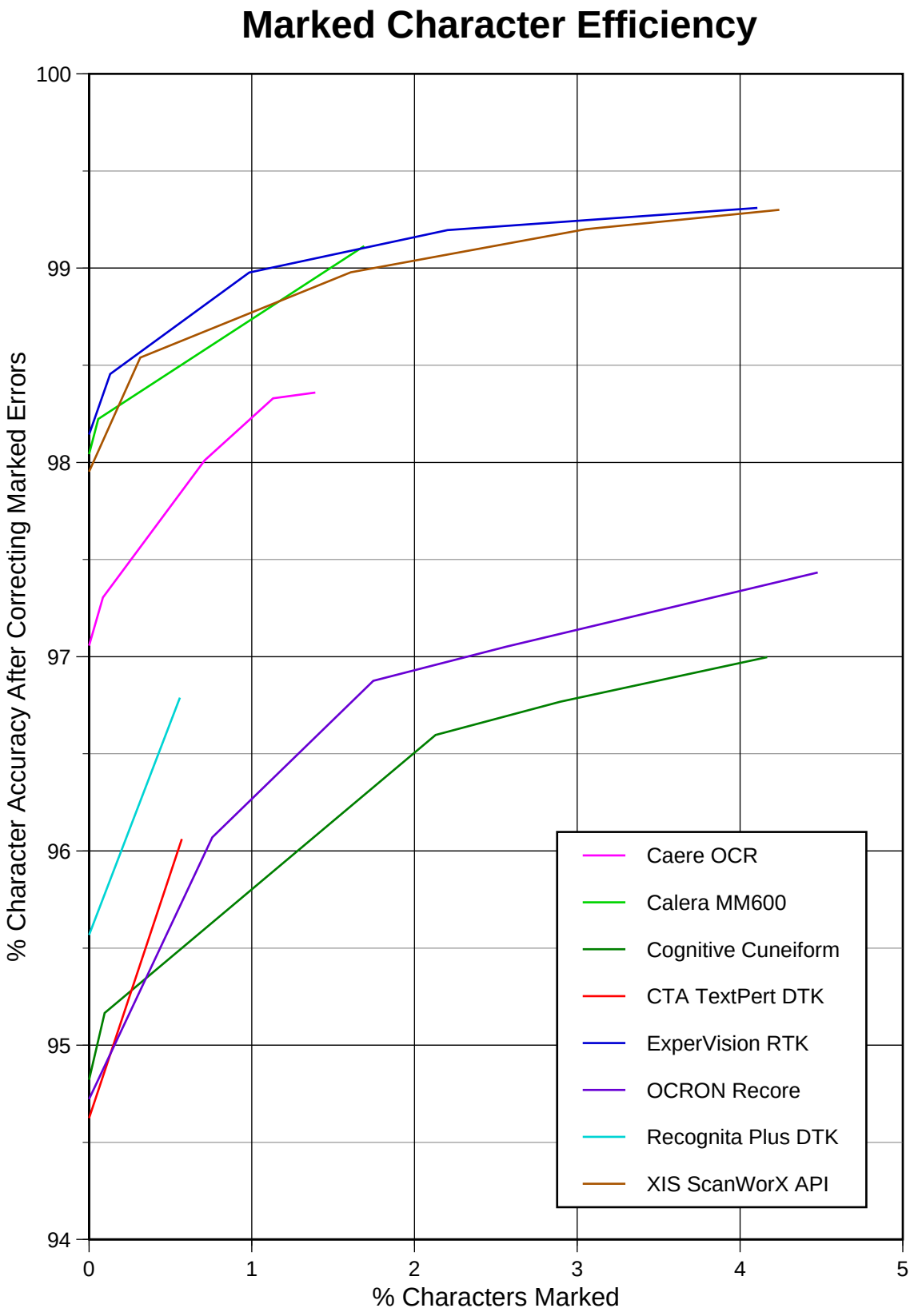
Graph 5f: OCRON Recore



Graph 5g: Recognita Plus DTK

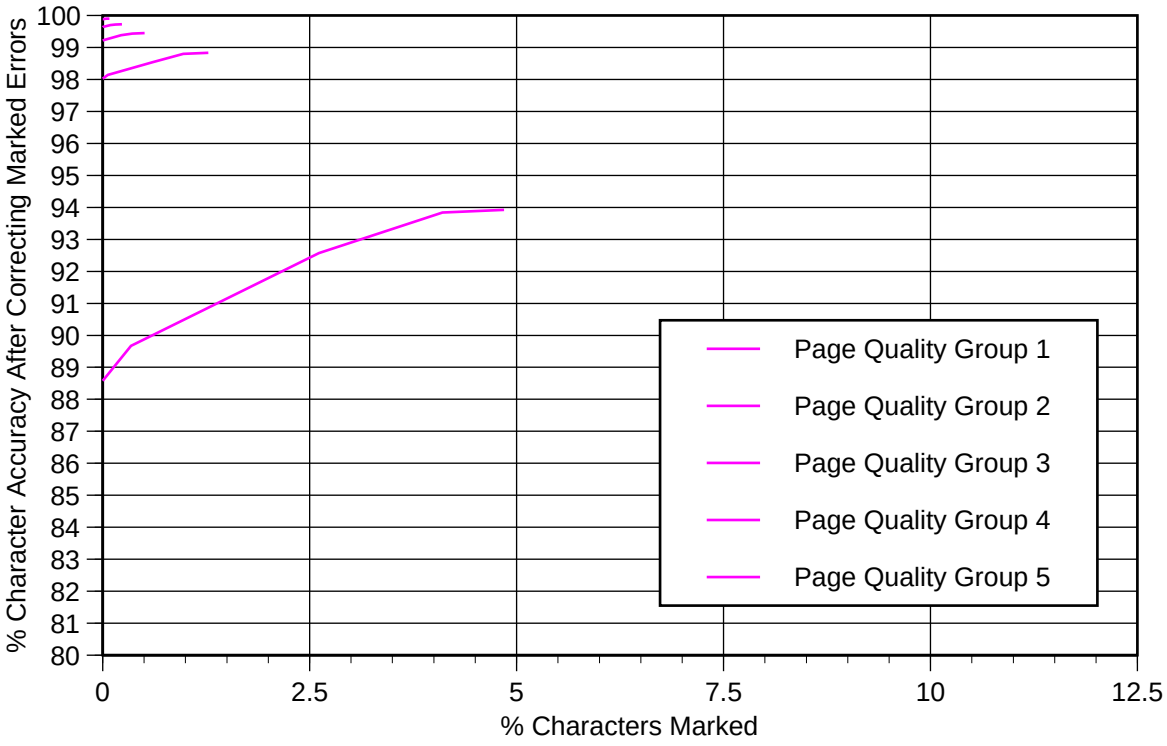


Graph 5h: XIS ScanWorX API

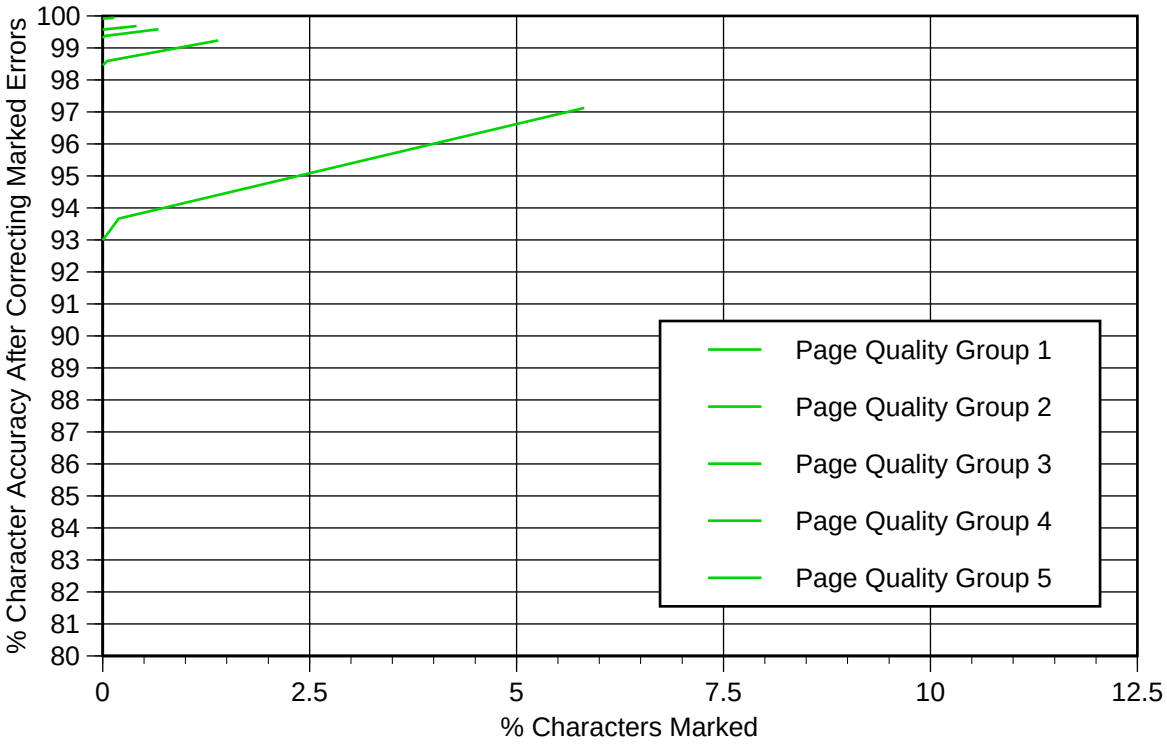

Marked Character Efficiency
100
99
98
97
96
95
94
0
1
2
3
4
5
% Character Accuracy After Correcting Marked Errors
% Characters Marked
Caere OCR
Calera MM600
Cognitive Cuneiform
CTA TextPert DTK
ExperVision RTK
OCRON Recore
Recognita Plus DTK
XIS ScanWorX API

Graph 6

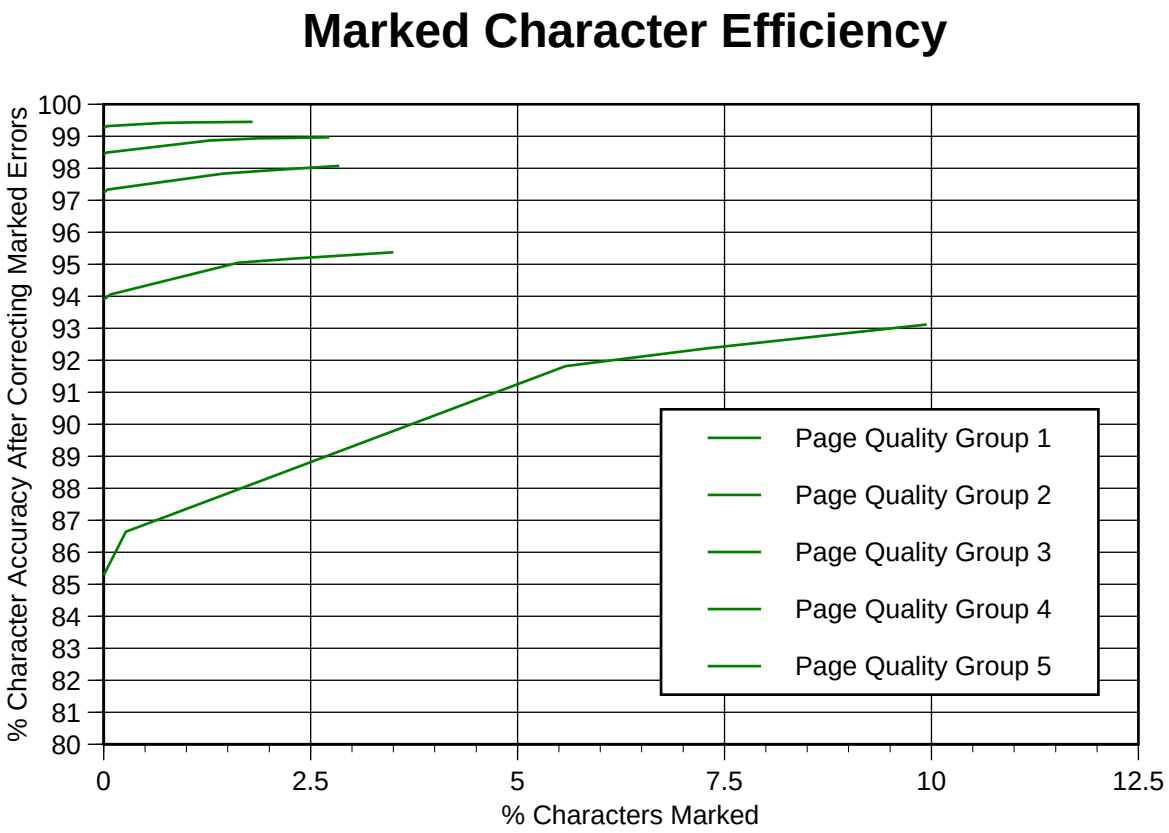
Marked Character Efficiency



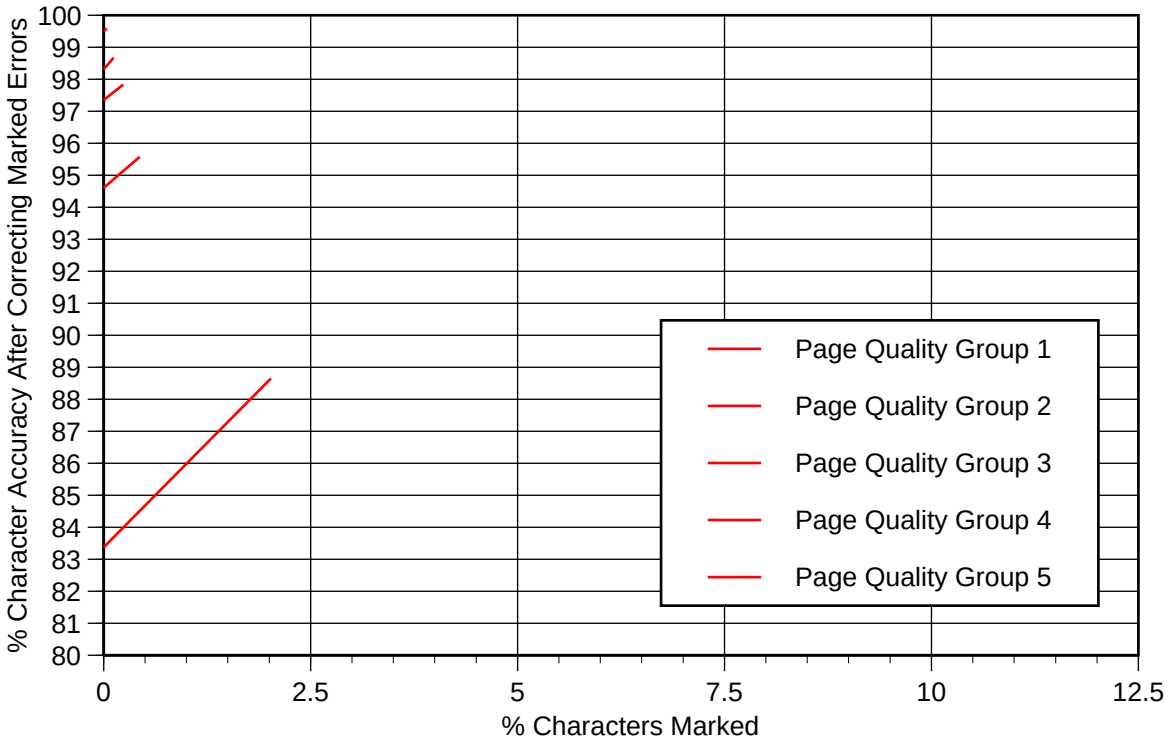
Graph 7a: Caere OCR



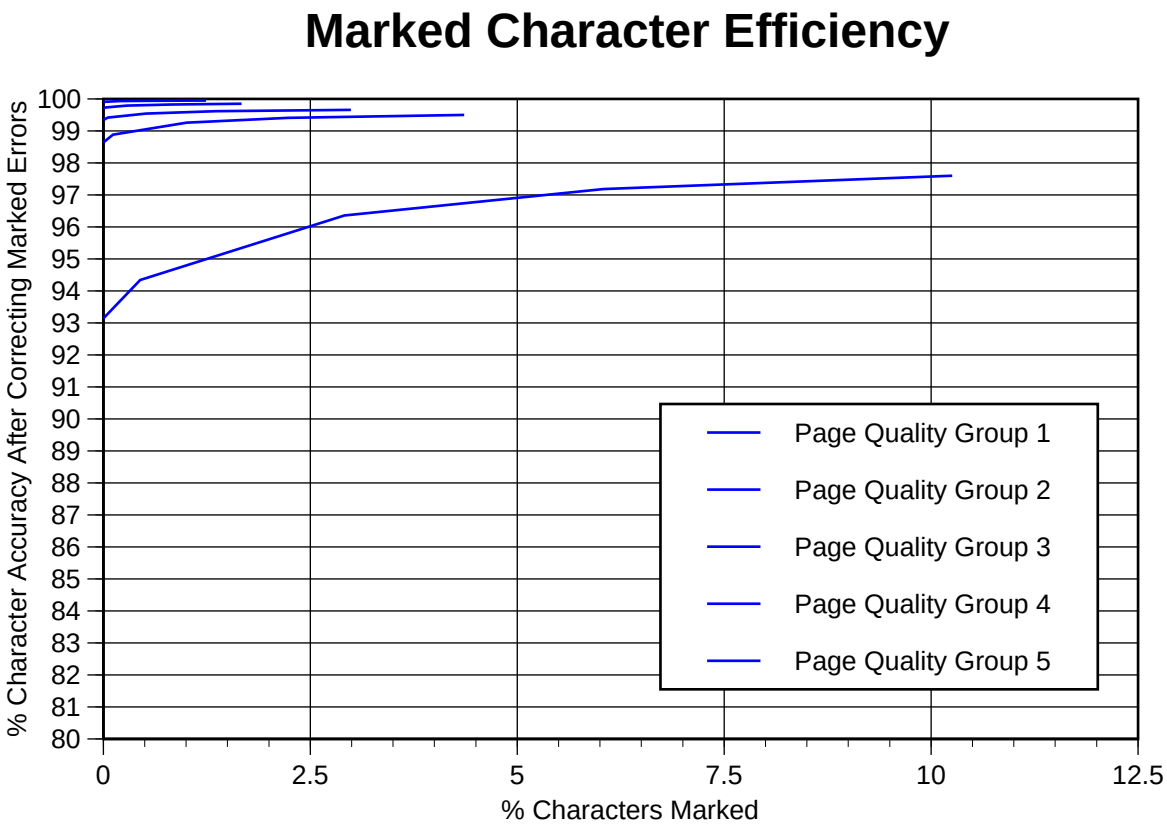
Graph 7b: Calera MM600



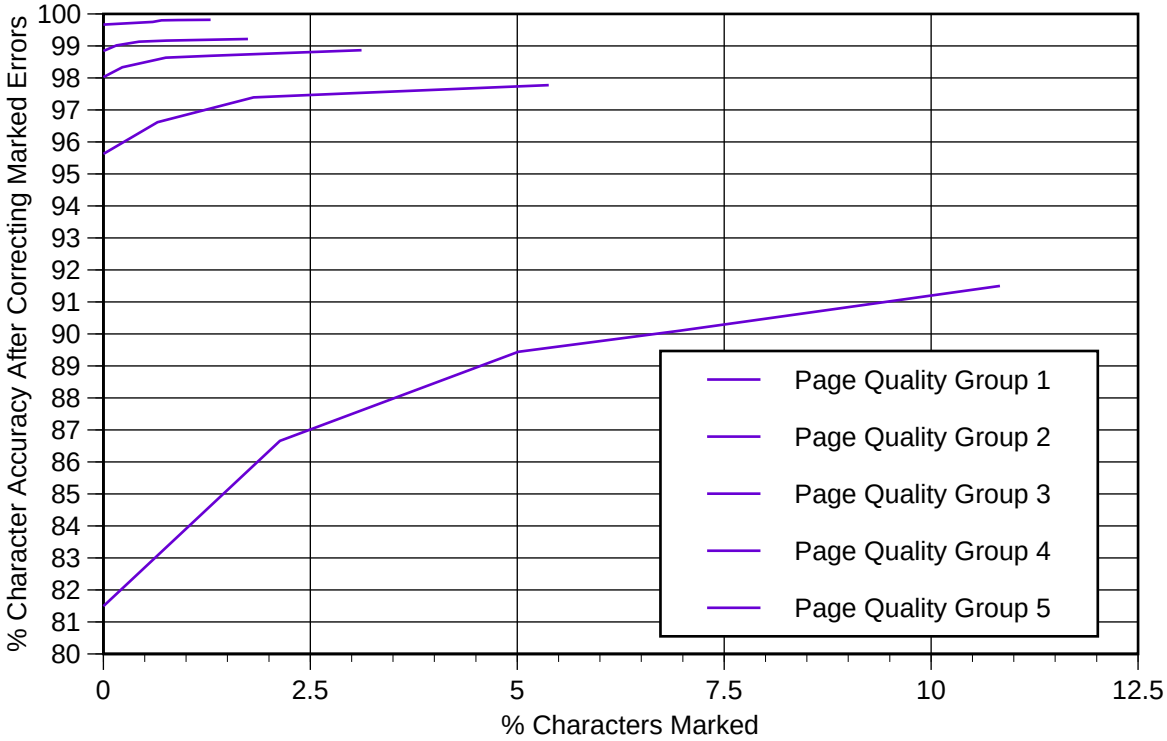
Graph 7c: Cognitive Cuneiform



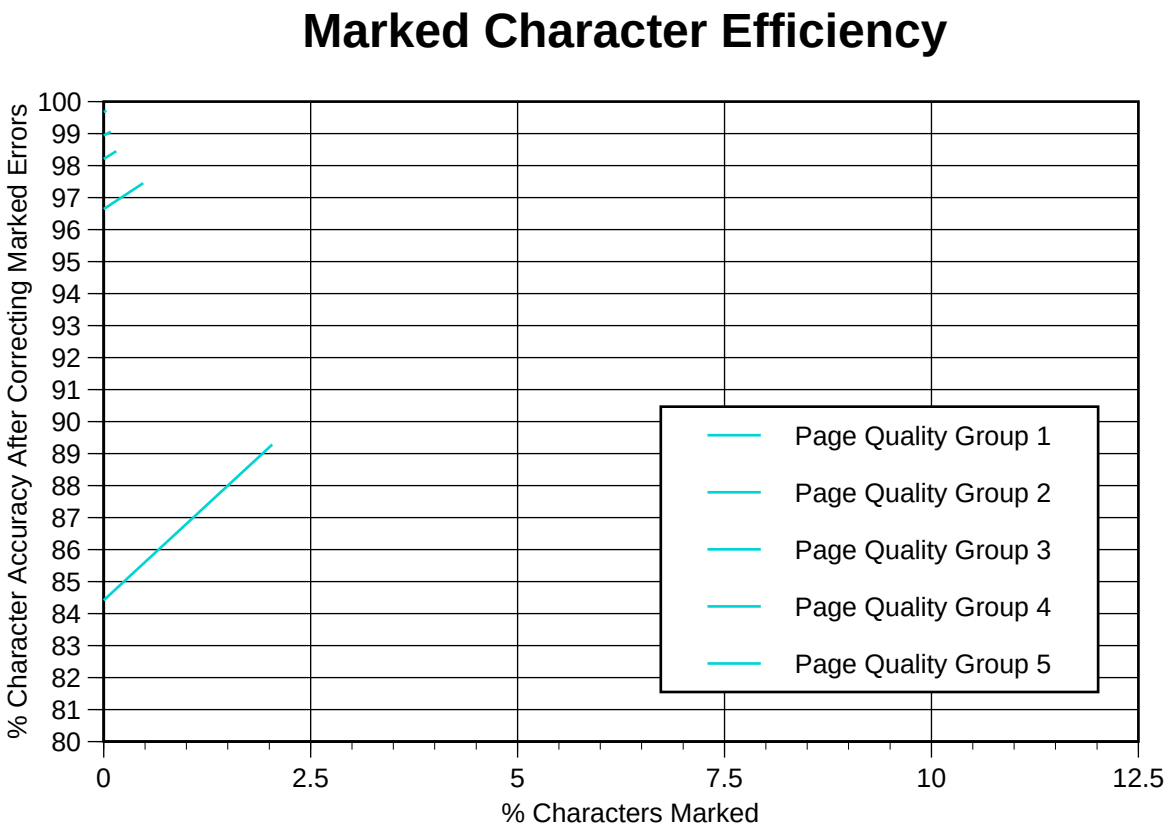
Graph 7d: CTA TextPert DTK



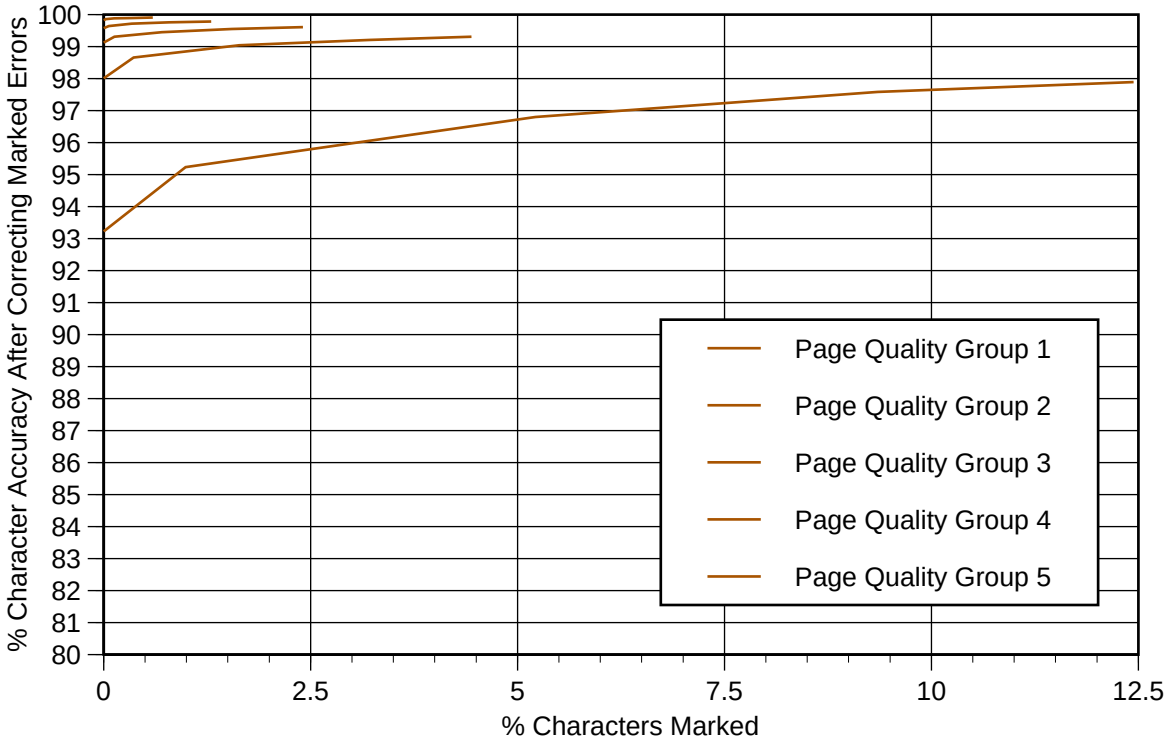
Graph 7e: ExperVision RTK



Graph 7f: OCRON Recore



Graph 7g: Recognita Plus DTK



Graph 7h: XIS ScanWorX API