The Third Annual Test of OCR Accuracy

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

1 Introduction

ISRI has conducted its third annual test of the accuracy of OCR systems. Vendors submitted their latest technology for recognizing machine-printed English text from page images. This year's test re-used the 460-page sample from U.S. Department of Energy (DOE) documents that was used a year ago [Rice 93a]. In addition, a new 200-page sample, randomly selected from popular magazines, was utilized.

Eleven vendors elected to participate and submitted a version by the deadline, January 18, 1994. One of these, TRW, provided a voting system which took part, along with the ISRI Voting Machine, in a special evaluation presented in Section 7. Of the other ten systems, four possessed de£ciencies that made it necessary to exclude them from the test. The six systems that were tested are listed in Table 1. All were "pre-releases" or "beta" versions at the time of submission.

Vendor	Version Name	Version #	Platform
Caere Corporation Los Gatos, CA	Caere OCR	132	SPARC
Calera Recognition Systems, Inc. Sunnyvale, CA	Calera WordScan	4	SPARC
Electronic Document Technology Singapore	EDT ImageReader	2.0	PC
ExperVision, Inc. San Jose, CA	ExperVision RTK	3.0	PC
Recognita Corp. of America Sunnyvale, CA	Recognita Plus DTK	2.00.D12	PC
Xerox Imaging Systems, Inc. Peabody, MA	XIS OCR Engine	10	SPARC

Table 1: OCR Systems Tested

The following systems were excluded from the test:

- CTA, Inc. of New Haven, CT submitted a version of TextPert DTK which did not function. A corrected version was received after the deadline.
- International Neural Machines, Inc. of Waterloo, Ontario provided a version of NeuroTalker which would frequently hang.
- Ligature, Inc. of Burlington, MA submitted a version of CharacterEyes which appeared unable to handle skewed input. It produced little or no output for page images containing moderately skewed text.
- OCRON, Inc. of Santa Clara, CA provided a version of Recore which was unable to process landscape pages.

2 Test Data and Methodology

Two sets of test data were utilized: the "DOE sample" and the "Magazine sample." The DOE sample consists of the same 460 pages that were used in last year's test. These pages were selected at random from a collection of approximately 2,500 scienti£c and technical documents (about 100,000 pages). This collection is described in [Nartker 92].

New this year is the Magazine sample. The 100 magazines having the largest paid circulation in 1992 were identi£ed from data reported in *Advertising Age* magazine [AdAge 93]. One recent issue (from 1992 or 1993) of each of these magazines was acquired, from which two pages were selected at random, thereby producing a 200-page sample.

This sample provides quite a contrast to the DOE sample:

1. It has a greater variety of typefaces and type sizes; however, nearly all of its fonts are proportional pitch, whereas the DOE sample contains both £xed and proportional pitch fonts.
2. It consists of clean, original pages, while the DOE sample includes many photocopies exhibiting a wide range of defects.
3. The DOE sample is almost entirely comprised of black characters printed on white paper, whereas the magazine pages contain many examples of colored text on colored backgrounds, including white characters on dark backgrounds ("reverse text").
4. In general, the magazine pages have complex, multi-column layouts, while the DOE pages have simpler layouts which include objects such as tables, graphs and equations that are not commonly found on magazine pages.

The pages of the DOE sample were scanned at 300 dots per inch (dpi) using a Fujitsu M3096E+ scanner; the scanner's default threshold was used to produce the binary images. The magazine pages were scanned at 200, 300 and 400 dpi using a Fujitsu M3096G scanner; the binary images were generated using a £xed threshold of 127 (out of 255) by this 8-bit gray scale scanner. The 200 and 400 dpi images were used only in the test of the effect of resolution described in Section 3.4.

Each page image was manually zoned, i.e., rectangular regions containing text were delineated, and correct text ("ground-truth") was carefully prepared corresponding to each zone. All text on a page was zoned, including tables, captions, page headers and footers, with a few exceptions; for example, mathematical equations were excluded from the DOE sample, and text within advertisements was omitted from the Magazine sample. Also, any degraded text that is essentially unreadable by humans was excluded. A total of 1,313 zones were de£ned for the DOE sample, containing a total of 817,946 characters. For the Magazine sample, 1,414 zones were de£ned containing 666,134 characters. See Table 2 for a breakdown by zone type. For more information on ISRI "zoning rules" and ground-truth representation, see [Rice 93c].

Each OCR system processed the same zoned portions of the same binary images, and was operated in an entirely automated manner. All processing, including the tabulation of errors, was carried out under computer control. The software tools that were used are part of the ISRI OCR Experimental Environment [Rice 93b].

Zone Type	DOE Sample # Zones	DOE Sample # Characters	Magazine Sample # Zones	Magazine Sample # Characters
“Main body” Text	512	667,161	1,072	630,441
Table	133	99,839	8	5,462
Caption	125	18,042	153	25,403
Footnote	67	13,981	2	655
Header/Footer	448	7,053	179	4,173
Other Text	28	11,870	0	0
Total	1,313	817,946	1,414	666,134

Table 2: Sample Data by Zone Type

3 Character Accuracy

Each character insertion, substitution or deletion required to correct the text generated by an OCR system is counted as an error.[1] *Character accuracy* is given by

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

where n is the total number of characters in the ground-truth text. Table 3 lists for each OCR system, the number of errors made, and the corresponding character accuracy, for the DOE and Magazine samples.

Last year we observed that the vendors achieved a remarkable 25 to 50% reduction in errors compared with the previous year [Nartker 94]. Using the 460-page DOE sample as the yardstick, the rate of reduction over this past year has been less dramatic, about 10%. Table 4 compares the number of errors made last year with this year, for the £ve vendors who participated both years.

[1] This metric is attributed to Levenshtein [Levenshtein 66].

	DOE Sample		Magazine Sample	
	# Errors	% Accuracy	# Errors	% Accuracy
Caere OCR	21,693	97.35	21,121	96.83
Calera WordScan	12,459	98.48	19,507	97.07
EDT ImageReader	36,658	95.52	29,815	95.52
ExperVision RTK	13,130	98.39	16,556	97.51
Recognita Plus DTK	36,381	95.55	27,291	95.90
XIS OCR Engine	15,324	98.13	24,321	96.35

Table 3: Character Accuracy

	1993	1994	
	# Errors	# Errors	% Change
Caere	24,074	21,693	–9.9
Calera	16,013	12,459	–22.2
ExperVision	15,186	13,130	–13.5
Recognita	36,250	36,381	+0.4
XIS	16,750	15,324	–8.5

Table 4: Progress in One Year, DOE Sample

3.1 Confusions

Table 5 lists the 20 most common errors, or "confusions," made by the OCR systems in processing the DOE and Magazine samples. This ranking is based on the median number of occurrences for the six participants. The leading source of error is the introduction and dropping of space characters. If a space is generated where there should be none, a word is split (e.g., "Neva da"); if a space is not generated where there should be one, two words are joined (e.g., "Universityof").

	DOE Sample		Magazine Sample	
	Correct	Generated	Correct	Generated
1		space	space	
2	space			space
3	,	.	e	c
4	e	c	.	
5	i	l	,	
6	O	0	,	.
7	1	I	l	i
8	1	l	i	l
9	0	O	t	l
10	.	,	c	e
11	y	v	a	o
12		.	y	v
13	l	i	1	I
14	.		n	ll
15	i	t	–	
16	c	e	b	h
17		’	rn	m
18	i	I	l	I
19	,			.
20	l	I	s	S

Table 5: Most Common Confusions

3.2 Effect of Font Features

Each page of the DOE sample was assigned to one of two groups depending on whether it contains mostly proportional or £xed pitch text. This resulted in a proportional group consisting of 156 pages (338,319 characters), and a £xed group containing 304 pages (479,627 characters). As Table 6 indicates, each OCR system achieved a higher character accuracy when processing the £xed pitch pages.

Similarly, the DOE sample was subdivided depending on whether a page contains mostly serif or sans serif text. This produced a serif group having 296 pages (622,331 characters) and a sans serif group with 164 pages (195,615 characters). Table 7 shows that higher accuracies were obtained on the serifed text.

	Proportional % Accuracy	Fixed % Accuracy
Caere OCR	96.50	97.95
Calera WordScan	98.04	98.78
EDT ImageReader	94.81	96.02
ExperVision RTK	98.20	98.53
Recognita Plus DTK	94.41	96.36
XIS OCR Engine	97.48	98.58

Table 6: Proportional vs. Fixed Pitch, DOE Sample

	Serif % Accuracy	Sans Serif % Accuracy
Caere OCR	97.56	96.68
Calera WordScan	98.55	98.25
EDT ImageReader	96.06	93.79
ExperVision RTK	98.61	97.72
Recognita Plus DTK	95.93	94.34
XIS OCR Engine	98.28	97.64

Table 7: Serif vs. Sans Serif, DOE Sample

3.3 Effect of Skew

Pages with skewed text present an added challenge for OCR systems. Figure 1 illustrates different skew angles. For each page of the DOE sample, the skew angle was estimated. Pages skewed less than $\frac{1}{2}$ degree were placed in one group (335 pages, 590,033 characters), pages skewed between $\frac{1}{2}$ and one degree were placed in a second group (83 pages, 143,395 characters), and pages with more than one degree of skew were placed in a third group (42 pages, 84,518 characters). The maximum skew angle in the DOE sample is 2.4 degrees.

Graph 1a shows the effect of skew on character accuracy. It appears that less than one degree of skew does not pose a problem, but there is a signi£cant decline in accuracy for some OCR systems when processing the pages with more than one degree of skew. We suspect that the systems least affected by skew are performing a de-skewing operation as a pre-processing step.

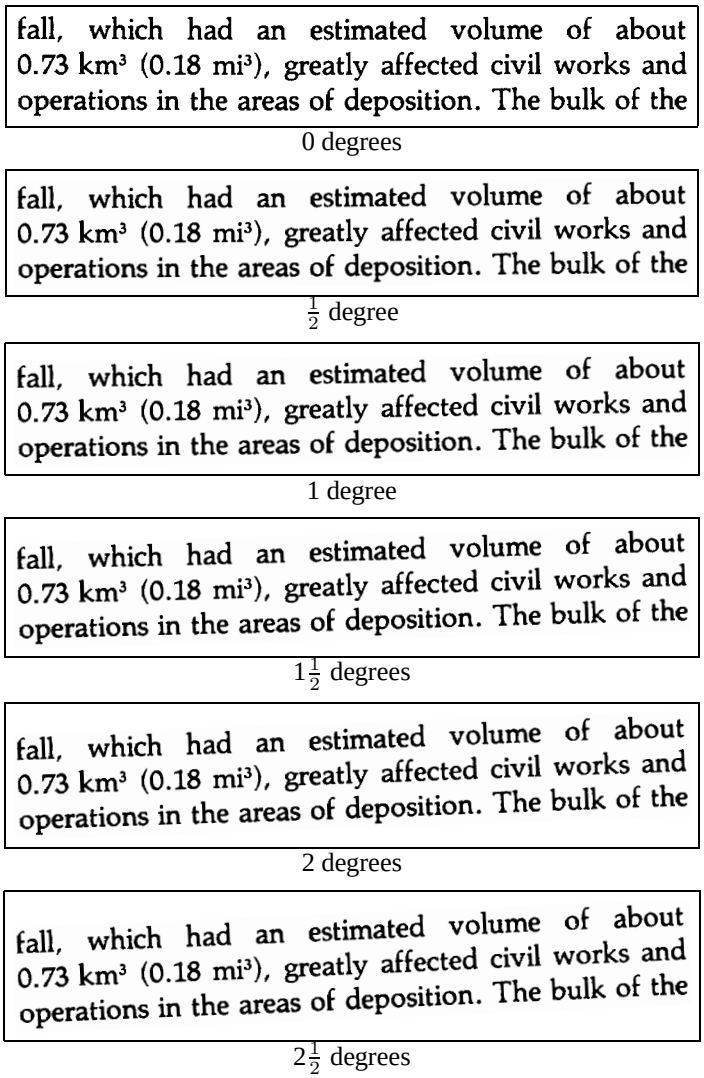


Figure 1: Skew Angles

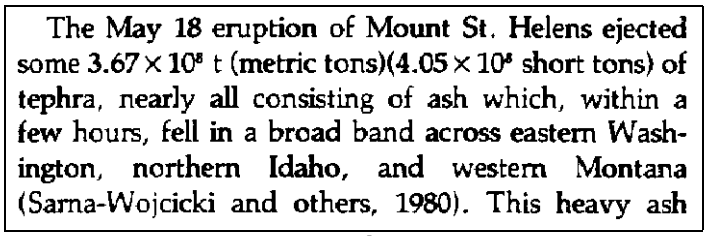
3.4 Effect of Resolution

To determine the effect of resolution on accuracy, pages of the Magazine sample were scanned at 200, 300 and 400 dpi. Figure 2 illustrates these three resolutions.

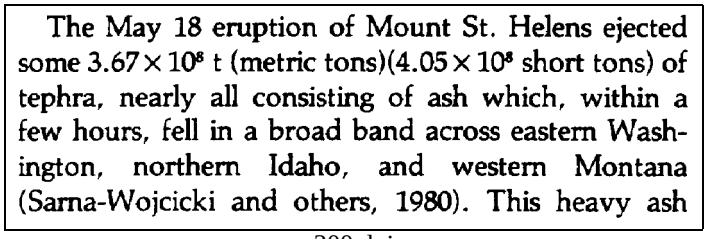
Table 8 shows the number of errors made by the OCR systems for each resolution,[2] and Graph 1b displays the corresponding character accuracies. Between 20 and 110% more errors were made on the 200 dpi images than on the 300 dpi images. ExperVision RTK was affected the least by the low resolution.

Surprisingly, the OCR systems made more errors on the 400 dpi images than on the 300 dpi images, with the exception of the XIS OCR Engine, which took advantage of the increased resolution to reduce its errors by almost 30%.

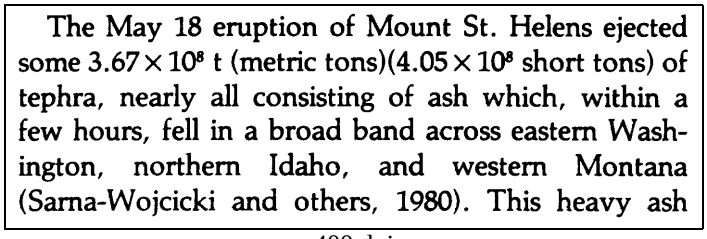
[2]Calera WordScan was unable to process the 400 dpi images.



200 dpi



300 dpi



400 dpi

Figure 2: Resolutions

	200 dpi # Errors	300 dpi # Errors	400 dpi # Errors
Caere OCR	41,950	21,121	24,718
Calera WordScan	27,394	19,507	—
EDT ImageReader	51,925	29,815	30,150
ExperVision RTK	20,109	16,556	19,210
Recognita Plus DTK	57,836	27,291	28,133
XIS OCR Engine	39,263	24,321	17,357

Table 8: Effect of Resolution, Magazine Sample

3.5 Effect of Page Quality

The median character accuracy achieved by several OCR systems when processing a given page is a good measure of the quality of the page, or at least its "OCR dif£culty." We use this measure to divide each sample into £ve "Page Quality Groups" of approximately equal size. Group 1 contains the pages with the highest median accuracy (least OCR dif£culty), and Group 5 contains the pages with the lowest median accuracy (greatest OCR dif£culty).

Tables 9 and 10 show the character accuracies achieved within each Page Quality Group for the DOE and Magazine samples, respectively. Graphs 2a and 2b present graphs of this data. A large percentage of the errors, between 65 and 75%, are made on the worst 20% of each sample, i.e., Group 5.

	Group 1	Group 2	Group 3	Group 4	Group 5
Caere OCR	99.86	99.58	99.26	98.14	89.92
Calera WordScan	99.92	99.59	99.39	98.68	94.83
EDT ImageReader	99.60	99.08	98.13	96.25	84.58
ExperVision RTK	99.92	99.70	99.48	98.64	94.25
Recognita Plus DTK	99.64	98.96	98.10	96.44	84.66
XIS OCR Engine	99.88	99.63	99.17	98.30	93.66

Table 9: Character Accuracy vs. Page Quality, DOE Sample

	Group 1	Group 2	Group 3	Group 4	Group 5
Caere OCR	99.71	99.37	98.90	97.64	88.74
Calera WordScan	99.74	99.34	98.82	97.78	89.86
EDT ImageReader	99.28	98.50	97.81	95.75	86.51
ExperVision RTK	99.82	99.27	99.33	98.19	91.13
Recognita Plus DTK	99.39	99.13	98.22	96.21	86.80
XIS OCR Engine	99.58	96.93	98.56	96.14	90.64

Table 10: Character Accuracy vs. Page Quality, Magazine Sample

Figures 3 and 4 show examples of dif£cult images from the DOE and Magazine samples, respectively. Broken and touching characters are the biggest source of error for the DOE sample, although skew is a contributing factor. While the Magazine sample contains some broken and touching characters, a greater challenge is presented by reverse text and text printed on shaded backgrounds; in fact, some of these regions were not readable by any of the OCR systems tested.

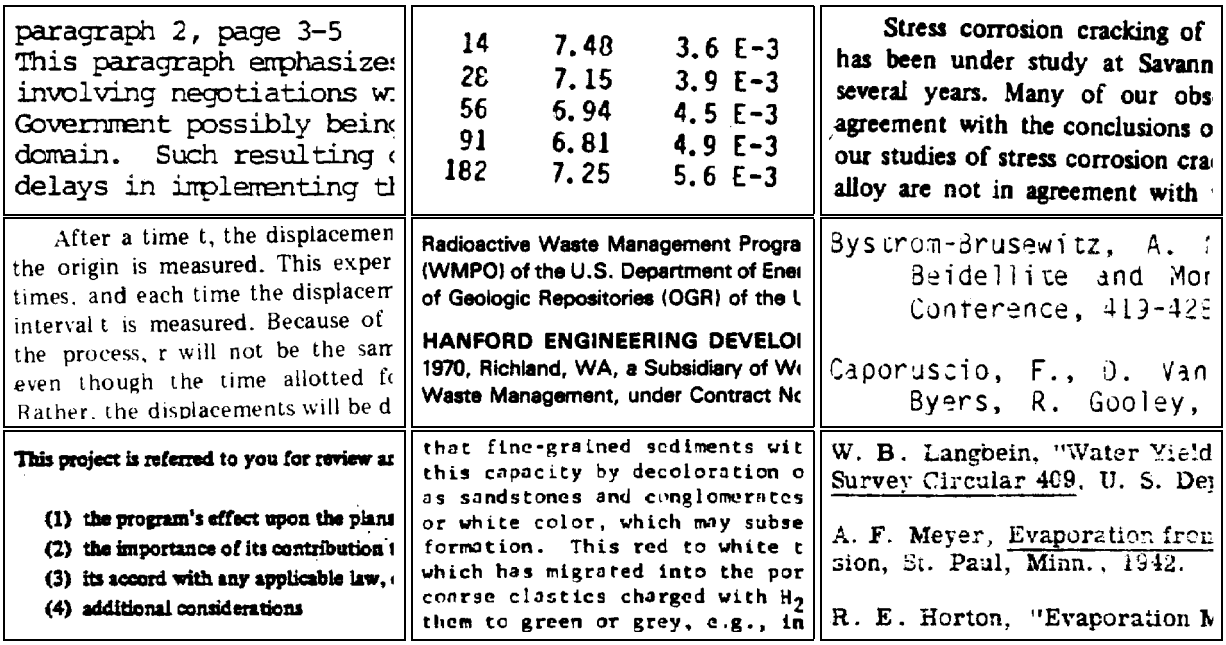


Figure 3: Examples from Page Quality Group 5, DOE Sample

Figure 4: Examples from Page Quality Group 5, Magazine Sample

4 Word Accuracy

In a text retrieval application, documents are retrieved from a database by matching search terms with words in the documents. If the database was loaded with OCR-generated text containing errors, the *word accuracy* of this data, i.e., the percentage of words correctly recognized by the OCR system, is of considerable interest.

We de£ne a word to be any sequence of one or more letters. Since full-text searching is usually 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., "UnIverSity").

The DOE and Magazine samples contain 119,497 and 114,361 words, respectively. Table 11 shows the number of misrecognized words, and the corresponding word accuracy, for each sample.

	DOE Sample		Magazine Sample	
	# Misrec.	% Accuracy	# Misrec.	% Accuracy
Caere OCR	7,028	94.12	5,406	95.27
Calera WordScan	3,215	97.31	4,546	96.02
EDT ImageReader	11,199	90.63	9,153	92.00
ExperVision RTK	3,925	96.72	3,575	96.87
Recognita Plus DTK	12,385	89.64	7,690	93.28
XIS OCR Engine	4,693	96.07	6,486	94.33

Table 11: Word Accuracy

Stopwords are common words such as "the," "of," "and," "in," etc., which are normally not indexed by text retrieval systems. Thus, it is *non-stopword accuracy*, i.e., the percentage of correctly recognized *non-stopwords*, that is especially relevant to text retrieval.

Using the default set of 110 stopwords provided by the BASISPLUS text retrieval product [IDI 90], it was determined that 35% of the words in the DOE sample are stopwords vs. 41% for the Magazine sample. The difference between the two samples is expected due to the technical nature of the DOE documents.

Table 12 shows the stopword and non-stopword accuracy for each sample; not surprisingly, the stopword accuracy

is considerably higher. Graphs 3a and 3b display the non-stopword accuracy for each Page Quality Group; note the signi£cant drop in non-stopword accuracy as page quality declines. In Graphs 4a and 4b, we see the effect of word length on non-stopword accuracy, for word lengths from one to 12 characters.

In text retrieval, users search not only for words, but also phrases. We de£ne a phrase of length n to be any sequence of n consecutive words. *Phrase accuracy* is the percentage of correctly recognized phrases. Graphs 5a and 5b show the phrase accuracy for lengths one to eight.[3]

Phrase accuracy provides a measure of "error bunching." Errors that are widely dispersed in generated text are more costly to correct than if the same number of errors were bunched. Given two OCR systems with the same word accuracy, the one with the higher phrase accuracy has produced errors that are more closely bunched, and thus, are more easily corrected.

[3]The phrase accuracy for length one is equivalent to the word accuracy.

	DOE Sample		Magazine Sample	
	Stopword	Non-stopword	Stopword	Non-stopword
Caere OCR	96.87	92.62	97.00	94.07
Calera WordScan	98.70	96.55	97.22	95.19
EDT ImageReader	94.86	88.33	95.00	89.91
ExperVision RTK	98.65	95.66	98.00	96.09
Recognita Plus DTK	93.98	87.27	95.09	92.01
XIS OCR Engine	98.32	94.85	96.31	92.95

Table 12: Stopword and Non-stopword Accuracy

5 Marked Character Ef£ciency

A *reject character* is a special character (usually ∼) that is generated by an OCR system when it is unable to recognize a character. If the OCR system is able to recognize the character, but has low con£dence in its decision, it may mark the generated character as suspect by preceding it with a special character called a *suspect marker* (often ∧). We refer to reject characters, and characters marked as suspect, as *marked characters*.

The purpose of marked characters is to attract the attention of the end-user to potential errors. We de£ne a *marked error* to be any error that can be identi£ed by examining marked characters. Clearly, the cost of correcting a marked error is substantially less than the cost of correcting an unmarked error; therefore, it is desirable for the OCR system to mark as many of its errors as possible. However, at the same time, it must minimize the number of *false marks*, which are correctly-generated characters that are marked as suspect. It takes time to verify the correctness of these characters, which adds to the overall cost of correction.

Table 13 illustrates these de£nitions.

	# Marked Characters	# Marked Errors	# Unmarked Errors	# False Marks
Nevada	0	0	0	0
Neva∼a	1	1	0	0
Nevacla	0	0	2	0
Nevac∧la	1	2	0	0
Neva∧da	1	0	0	1

Table 13: Marked Character Examples

While all six OCR systems produce reject characters, there are two that do not support suspect markers: EDT ImageReader and Recognita Plus DTK. The other four actually support more than one "level" of suspect markers, where the higher the level, the more characters are marked. We tested up to three of these levels.

Graphs 6a and 6b show how examining marked characters and correcting marked errors increases the character accuracy of the text. The X-axis indicates the number of marked characters examined (as a percentage of the total number of characters), and the Y-axis gives the character accuracy after the marked errors have been corrected. A high slope indicates a very ef£cient correction process, whereas a ¤at curve implies that most of the marked characters are false marks, and much time will be spent examining characters that are already correct.

Starting with the base character accuracy, each curve rises sharply to indicate the processing of the reject characters. This is a very ef£cient operation since a false mark can occur only when the correct character happens to be a tilde (∼). The slope then changes to re¤ect the rate of correction when processing the £rst level of suspect markers, and then changes again for both the second and third levels (if supported). In general, the slope decreases with each level, and thus, the higher levels are less useful.

Suppose a human editor examines up to 5,000 marked characters for each of our samples and corrects the identi£ed errors. Since this is less than 1% of the total number of characters in each sample, this is a reasonable task. The editor would £rst examine the reject characters, then the £rst level of suspect markers, followed by the second level, and so on, until 5,000 marked characters have been examined. Table 14 gives the number of errors in the text before and after this process, and the percentage of errors that are corrected. The editor is able to locate and correct almost one-third of the errors by examining less than 1% of the characters. This is very productive, but if an additional 5,000 marked characters were examined for each sample, only another 10% of the errors would be corrected.

	DOE Sample			Magazine Sample		
	Before	After	% Corrected	Before	After	% Corrected
Caere OCR	21,693	15,394	29	21,121	11,657	45
Calera WordScan	12,459	9,791	21	19,507	15,288	22
EDT ImageReader	36,658	31,573	14	29,815	24,314	18
ExperVision RTK	13,130	9,357	29	16,556	10,418	37
Recognita Plus DTK	36,381	26,338	28	27,291	19,584	28
XIS OCR Engine	15,324	10,388	32	24,321	17,453	28

Table 14: Error Correction Example

6 Automatic Zoning

All of the tests to this point have been performed on manually-zoned page images, i.e., each OCR system was given the coordinates of speci£c text regions to process. In this section we present the results of an automatic zoning test, in which the OCR systems were required to locate the text regions on each sample page, and determine their correct reading order.

The cost of correcting OCR-generated text is higher when automatic zoning is used because the text contains automatic zoning errors as well as recognition errors. But if the cost of correcting the automatic zoning errors is less than the cost of manual zoning, it is worthwhile to use this feature.

We estimate the cost of correcting the automatic zoning errors using a metric introduced a year ago [Kanai 93]. This metric determines the number of character insertions and “block move” operations required to correct the generated text. If the OCR system fails to locate a text region, insertions are needed to enter the missing block of text. Move operations are needed to correct the reading order of the text; many are required when the OCR system de-columnizes a table, or fails to de-columnize a multi-column page. Using a conversion factor, each move operation is expressed as an equivalent number of insertions so that the overall cost of correction is given solely in terms of insertions.

The metric is £rst applied to the text generated when automatic zoning is enabled; this determines the total cost of correcting both the recognition errors and the automatic zoning errors. The cost of correcting only the recognition errors is computed by applying the metric to the text generated when processing the manually-de£ned zones. Subtracting this cost from the total yields the cost of correcting the automatic zoning errors [Kanai 93].

Two of the six OCR systems did not take part in this test: Calera WordScan and EDT ImageReader. Calera WordScan does not support automatic zoning; EDT ImageReader claims to support this feature, but we did not observe any effect when it was enabled. The cost of correcting the automatic zoning errors is plotted for the other four OCR systems in Graphs 7a and 7b for the DOE and Magazine samples, respectively. Since the conversion factor is application-dependent, the cost is given for a range of conversion factors, from 0 to 100.

In the test conducted a year ago, it was noted that the greatest source of automatic zoning error in the DOE sample was the de-columnization of tables. Distinguishing between multi-column text, which requires de-columnization, and tables, which do not, is a dif£cult problem. Graphs 8a and 8b show that de-columnization of tables continues to be the leading source of error for the DOE sample. Conversely, the correct de-columnization of multi-column pages is the main challenge posed by the Magazine sample.

When distinguishing between tables and multi-column text, it appears that Recognita Plus DTK often chooses not

to de-columnize. While this strategy works very well for the DOE sample, it resulted in a high cost of correction for the Magazine sample. Overall, ExperVision RTK did the best on this test, managing to perform well on both samples.

7 Voting Systems

In this section we present the results of a test of two voting systems: the ISRI Voting Machine (Version 3.0), and TRW Ensemble (Version 1.0) from TRW of Redondo Beach, CA. A voting system processes a page image using OCR systems from several vendors. It takes the text generated by the OCR systems and partitions it into substrings on which the systems agree and disagree. To resolve disagreements, a voting algorithm is applied, and characters receiving the most votes are output. The goal of the voting system is to produce fewer errors than the best of the participating OCR systems. Past voting experiments have demonstrated that it is possible to correct more than 40% of the errors made by the best participant [Handley 91, Bradford 91, Rice 92].

The ISRI Voting Machine makes use of the latest versions submitted by Caere, Calera, ExperVision, Recognita and XIS. TRW Ensemble utilizes an earlier version from each of these vendors, plus a version from OCRON. Tables 15 and 16 give the number of errors and character accuracy for the DOE and Magazine samples, respectively.

	# Errors	% Accuracy
ISRI Voting Machine	7,204	99.12
TRW Ensemble	11,699	98.57
Calera WordScan	12,459	98.48
ExperVision RTK	13,130	98.39
XIS OCR Engine	15,324	98.13
Caere OCR	21,693	97.35
Recognita Plus DTK	36,381	95.55
EDT ImageReader	36,658	95.52

Table 15: Character Accuracy of Voting Systems, DOE Sample

	# Errors	% Accuracy
ISRI Voting Machine	13,496	97.97
ExperVision RTK	16,556	97.51
Calera WordScan	19,507	97.07
TRW Ensemble	20,584	96.91
Caere OCR	21,121	96.83
XIS OCR Engine	24,321	96.35
Recognita Plus DTK	27,291	95.90
EDT ImageReader	29,815	95.52

Table 16: Character Accuracy of Voting Systems, Magazine Sample

Graphs 9a and 9b show the character accuracy of the voting systems for each Page Quality Group. In Table 17, we indicate the number of errors made in each group of the DOE Sample. In addition, we determine the error reduction achieved relative to the participant making the fewest errors on this sample. For the ISRI Voting Machine, this is Calera WordScan Version 4. Since we did not evaluate the actual versions participating in TRW Ensemble, we compare its performance with ExperVision RTK Version 2.0, which made the fewest errors on this sample in last year's test. Similarly, Table 18 presents the results for the Magazine sample; however, the error reduction for TRW Ensemble is not given because we have no basis for comparison.

In general, the voting systems perform best when the text obtained from the OCR systems is of high accuracy. When a text region is degraded and dif£cult to read, there is usually much disagreement among the participants, which is dif£cult for a voting system to resolve. Indeed, if none, or perhaps only one, of the OCR systems is able to read a given text region, a voting system has little or no chance of producing accurate output.

The ISRI Voting Machine produces two levels of suspect markers based on the amount of disagreement among the participants. Graphs 10a and 10b show the effect of examining these marked characters and correcting the marked errors. The £rst level appears to be quite useful, while the second level is less valuable. Following the example presented in Section 5, an editor who examines 5,000 marked characters would correct about 25% of the errors.

	Group 1	Group 2	Group 3	Group 4	Group 5	Total
Calera WordScan 4	129	674	995	2,160	8,501	12,459
ISRI Voting Machine	27	114	445	950	5,668	7,204
Error Reduction	*79%*	*83%*	*55%*	*56%*	*33%*	*42%*
ExperVision RTK 2.0	163	477	1,072	2,211	11,263	15,186
TRW Ensemble	323	304	469	1,517	9,086	11,699
Error Reduction	—	*36%*	*56%*	*31%*	*19%*	*23%*

Table 17: Effect of Page Quality on Error Reduction, DOE Sample

	Group 1	Group 2	Group 3	Group 4	Group 5	Total
ExperVision RTK 3.0	244	958	900	2,382	12,072	16,556
ISRI Voting Machine	122	328	534	1,350	11,162	13,496
Error Reduction	*50%*	*66%*	*41%*	*43%*	*8%*	*18%*
TRW Ensemble	204	453	755	5,486	13,686	20,584

Table 18: Effect of Page Quality on Error Reduction, Magazine Sample

8 Interpretation

In this section we present the strengths of each OCR system based on our interpretation of the test results.

8.1 Caere OCR

Caere OCR is among the leaders in accuracy. It performs especially well on good quality pages. Its marked characters are very useful for error correction.

8.2 Calera WordScan

Calera WordScan is a very accurate system. It is relatively insensitive to skew, and performs well on degraded text. It does particularly well on £xed pitch text.

8.3 EDT ImageReader

This is the £rst year that we have tested a version from EDT. Unlike the four excluded systems described in Section 1, EDT ImageReader proved to be robust enough to participate in this year's test.

8.4 ExperVision RTK

Overall, ExperVision RTK performed the best in this year's test. It demonstrated consistently high accuracy. It performs especially well on proportional pitch text, and is least affected by low resolution (200 dpi). It also provides an excellent automatic zoning capability.

8.5 Recognita Plus DTK

Recognita Plus DTK provides an automatic zoning capability that is careful to avoid the de-columnization of tables, which is of considerable value when processing scienti£c and technical documents.

8.6 XIS OCR Engine

XIS OCR Engine is among the leaders in accuracy. It performs particularly well on £xed pitch text. Its accuracy when processing 400 dpi images is especially high.

8.7 TRW Ensemble

TRW is the £rst vendor to submit a voting system for our annual test. When processing good quality pages, TRW Ensemble produces about one-third fewer errors than the best participating OCR system.

9 Conclusion

Using contrasting test samples, we evaluated the latest OCR technology for recognizing machine-printed, English-language text. Character accuracy was viewed from several perspectives; the effects of font features, skew, resolution, and page quality were investigated. Measures of special signi£cance to text retrieval were presented, including word, non-stopword and phrase accuracy. We measured the degree to which marked characters facilitate the correction of recognition errors, and we estimated the cost of correcting automatic zoning errors. Finally, we demonstrated that voting systems are able to correct a large percentage of OCR errors, and we showed their limitations when processing poor quality pages.

It is important to note that the relative performance of these OCR systems may differ when processing other types of documents. Also, there are many aspects of these systems that were not evaluated, such as speed, purchase price, variety of output formats, quality of the user interface, etc. These factors should be considered when selecting the right OCR system for a given task.

Lastly, we wish to emphasize that ISRI does not endorse any particular OCR system or systems. The purpose of this test is to provide end-users with a current, independent assessment of OCR technology, and to identify problems at the state-of-the-art, thereby assisting vendors and researchers to advance the technology.

References

[AdAge 93] "Ad Age 300," *Advertising Age*, June 14, 1993.

[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.

[Kanai 93] J. Kanai, S. V. Rice, and T. A. Nartker, *A Preliminary Evaluation of Automatic Zoning*, Technical Report ISRI TR-93-02, University of Nevada, Las Vegas, April 1993.

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

[Nartker 94] T. A. Nartker, S. V. Rice, and J. Kanai, "OCR Accuracy: UNLV's Second Annual Test," *INFORM Magazine*, Association for Information and Image Management, January 1994.

[Rice 92] 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.

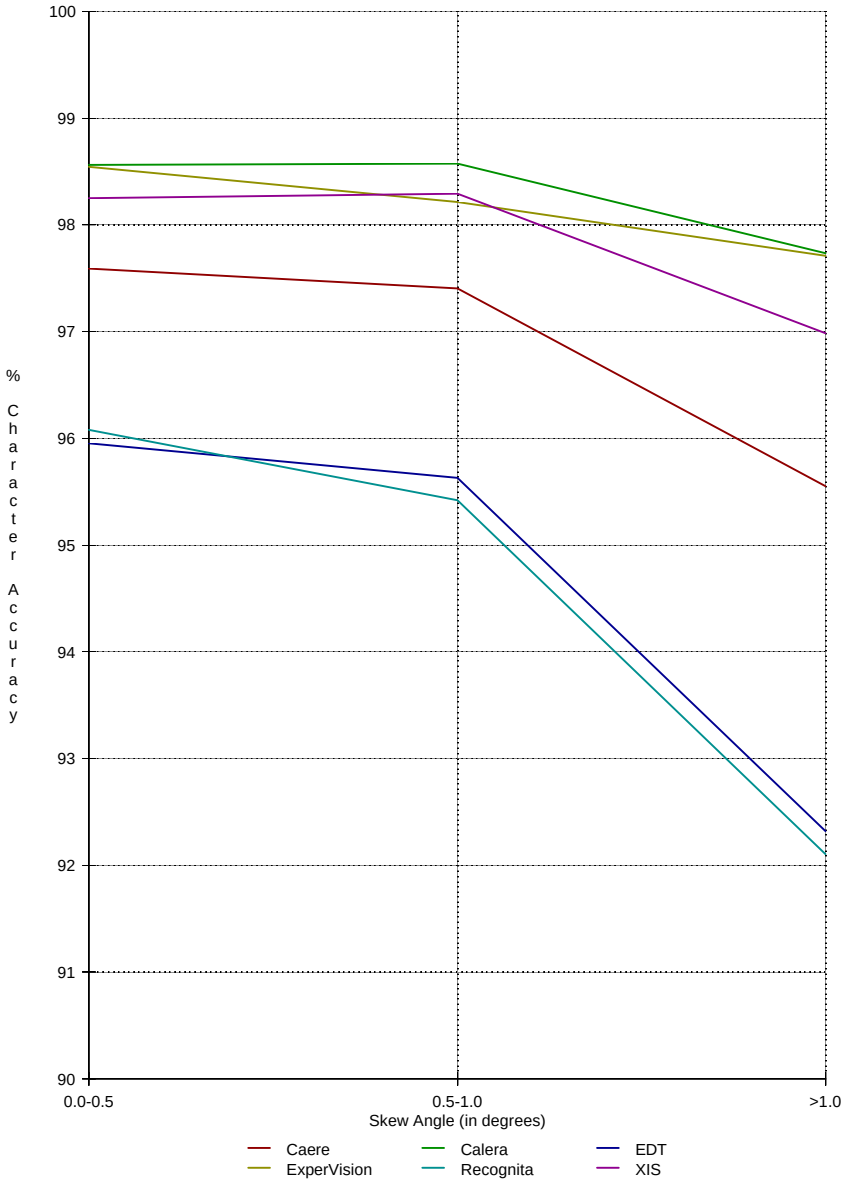
[Rice 93a] S. V. Rice, J. Kanai, and T. A. Nartker, *An Evaluation of OCR Accuracy*, Technical Report ISRI TR-93-01, University of Nevada, Las Vegas, April 1993.

[Rice 93b] S. V. Rice, *The OCR Experimental Environment, Version 3*, Technical Report ISRI TR-93-04, University of Nevada, Las Vegas, April 1993.

[Rice 93c] S. V. Rice, J. Kanai, and T. A. Nartker, *Preparing OCR Test Data*, Technical Report ISRI TR-93-08, University of Nevada, Las Vegas, June 1993.

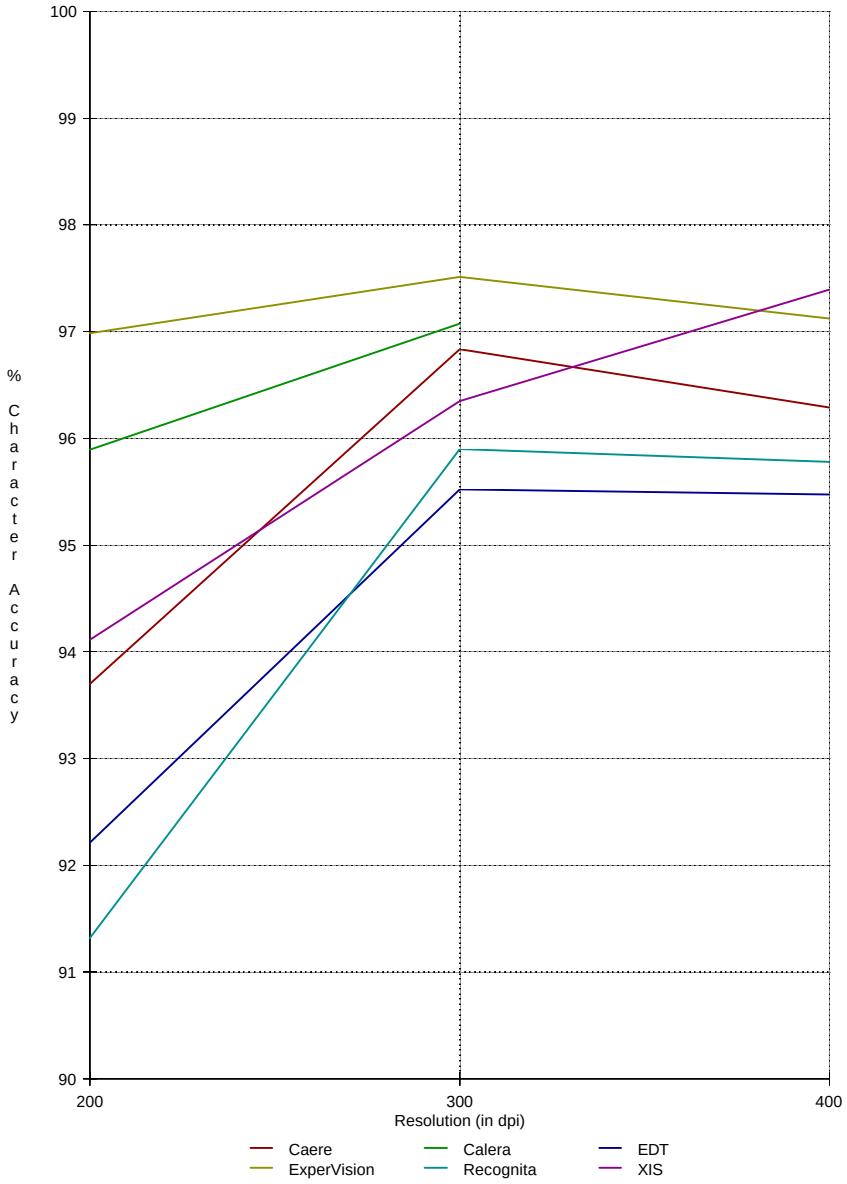
Graph 1a: Effect of Skew

DOE Sample



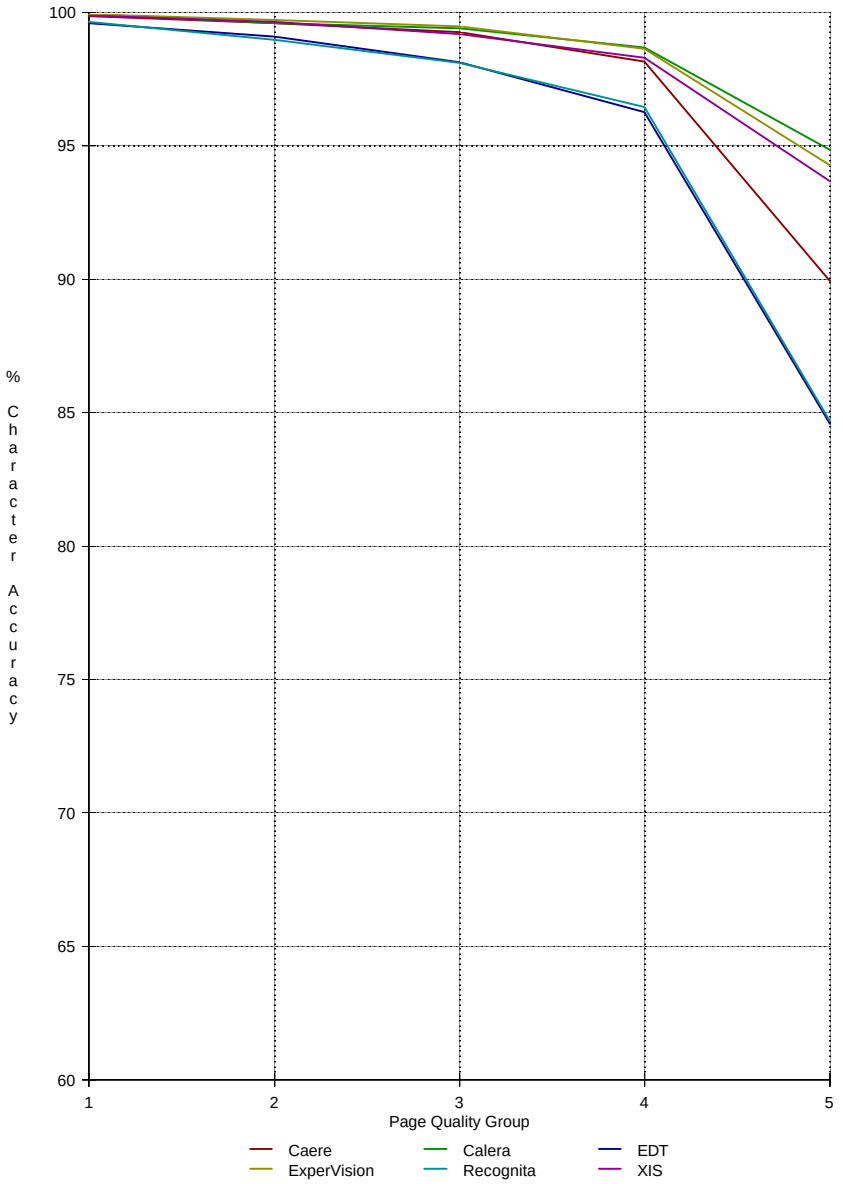
Graph 1b: Effect of Resolution

Magazine Sample



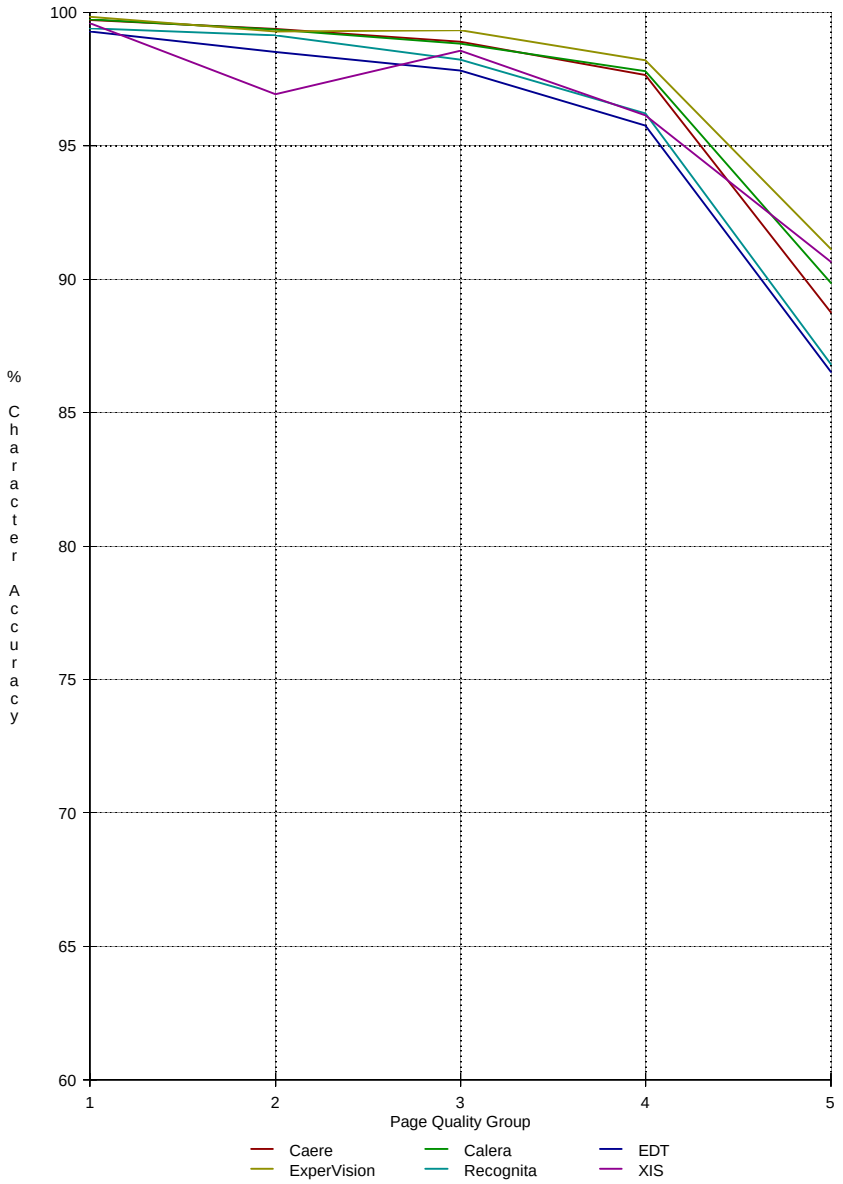
Graph 2a: Character Accuracy

DOE Sample



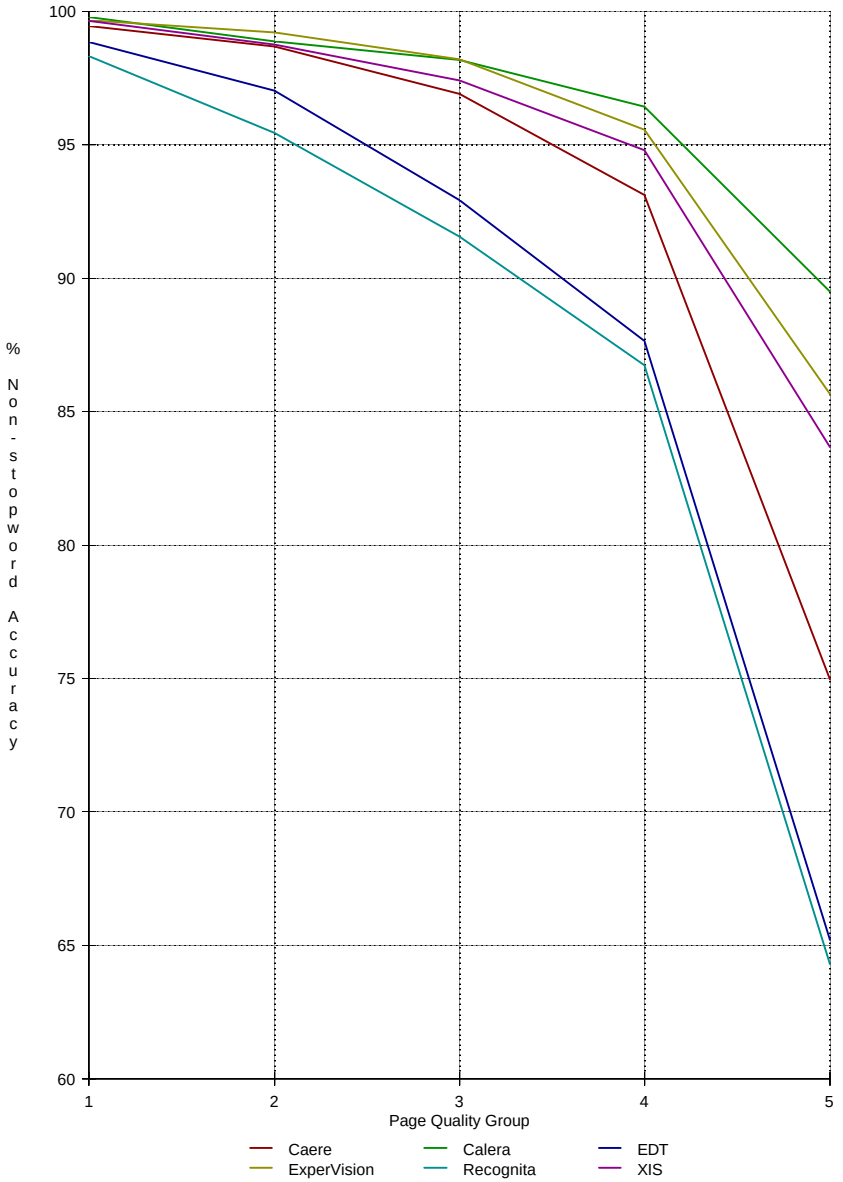
Graph 2b: Character Accuracy

Magazine Sample



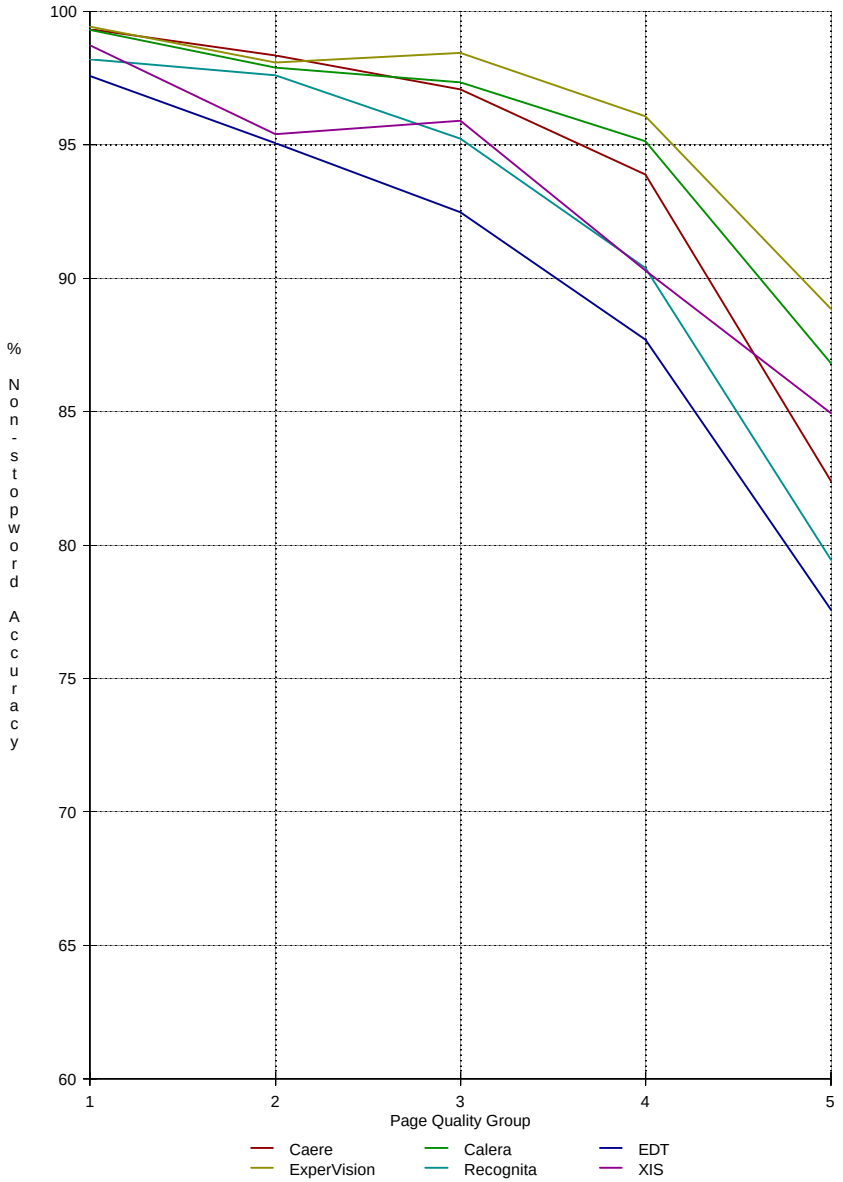
Graph 3a: Non-stopword Accuracy

DOE Sample



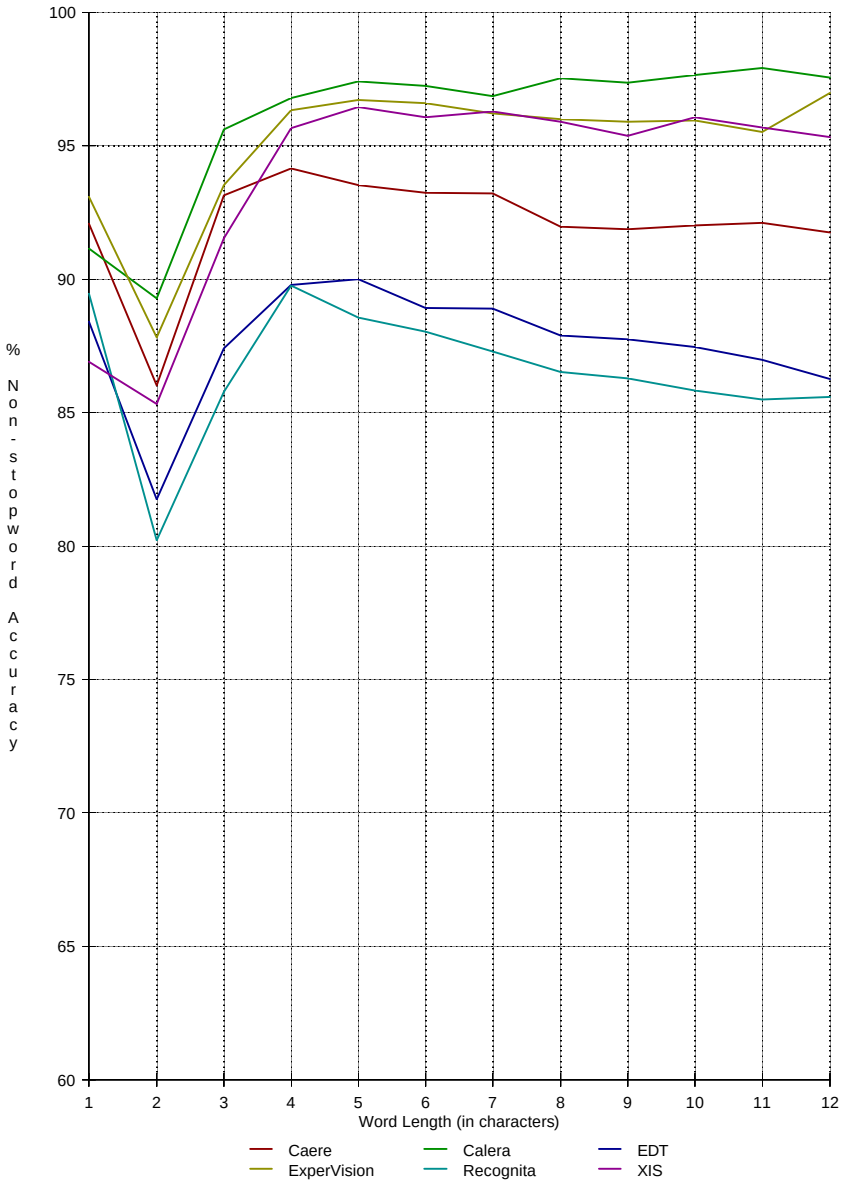
Graph 3b: Non-stopword Accuracy

Magazine Sample



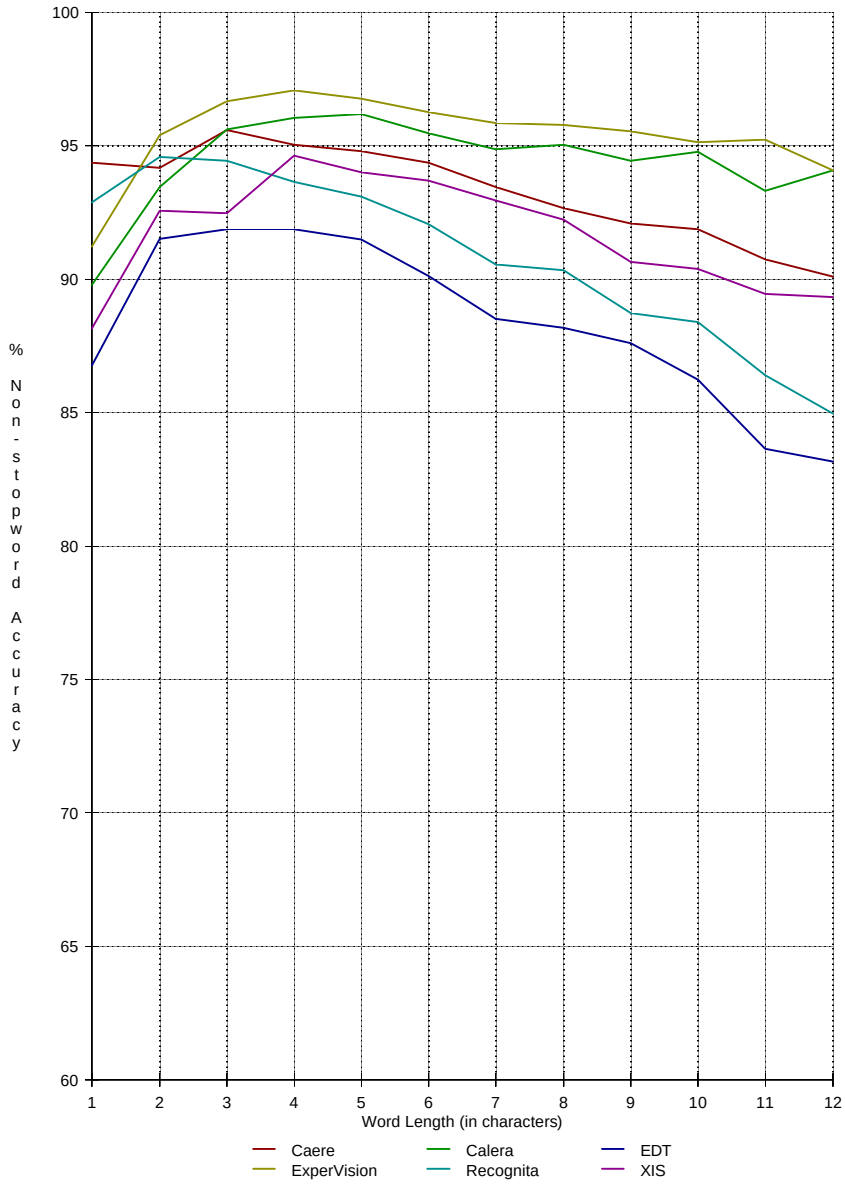
Graph 4a: Effect of Word Length

DOE Sample



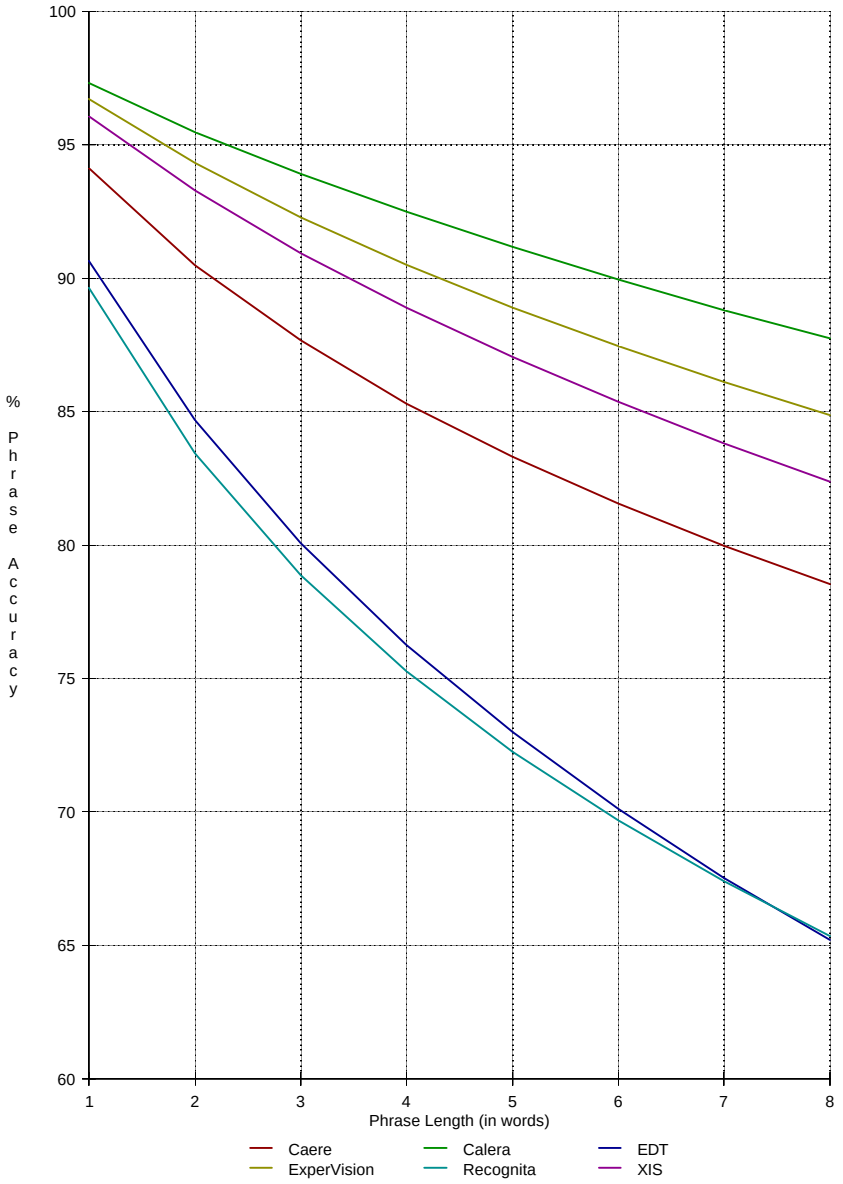
Graph 4b: Effect of Word Length

Magazine Sample



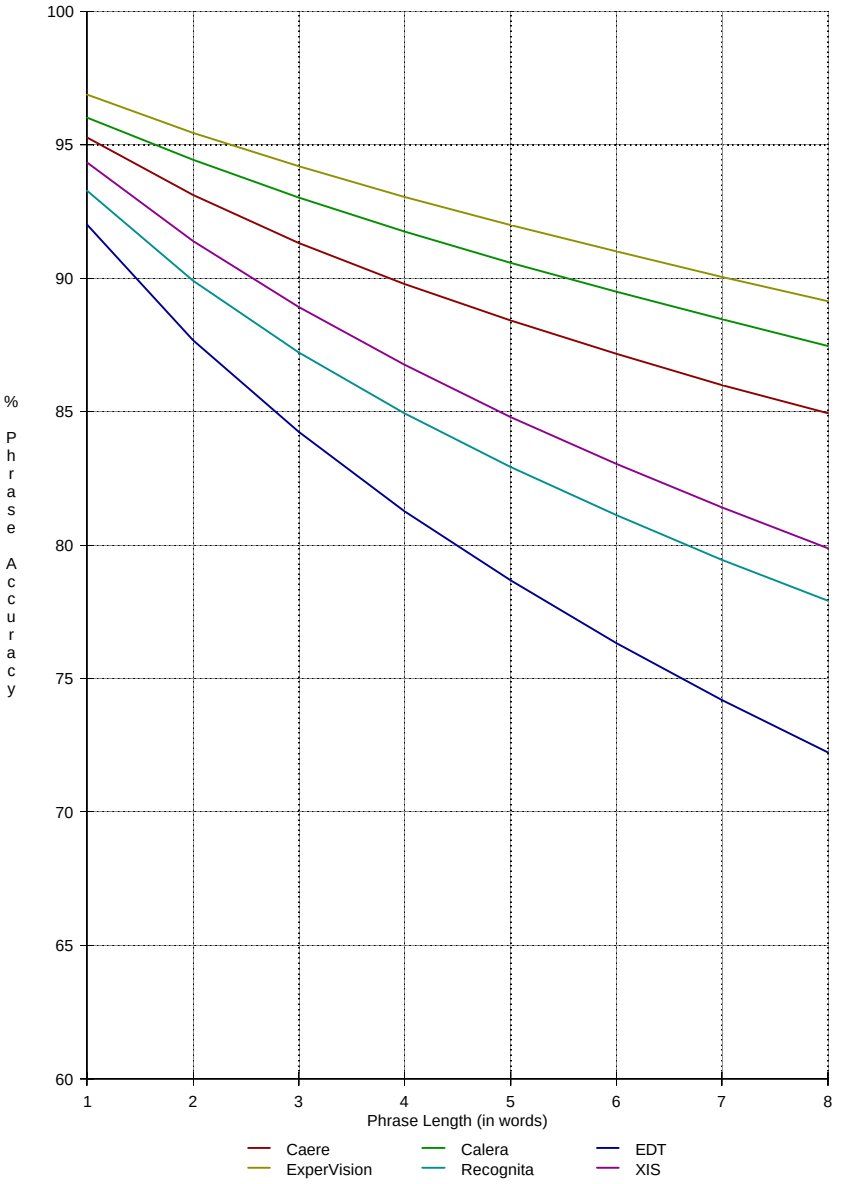
Graph 5a: Phrase Accuracy vs. Length

DOE Sample



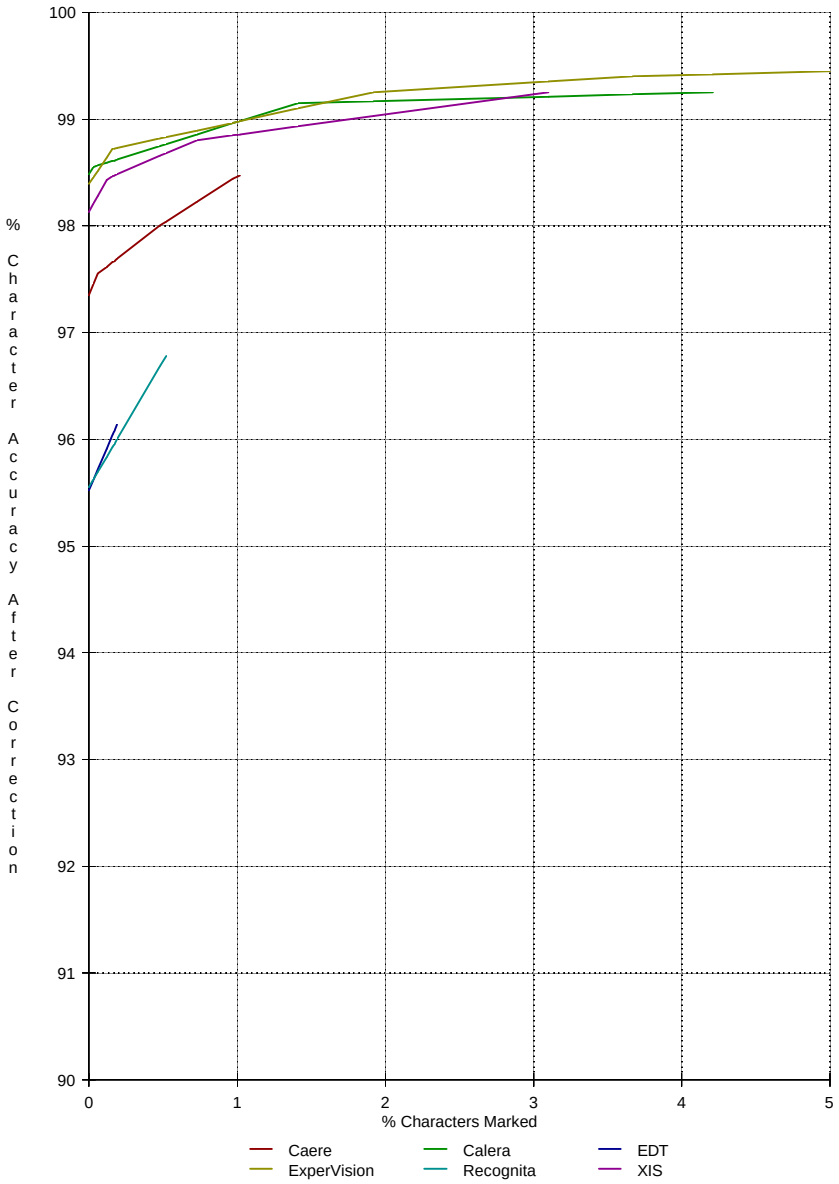
Graph 5b: Phrase Accuracy vs. Length

Magazine Sample



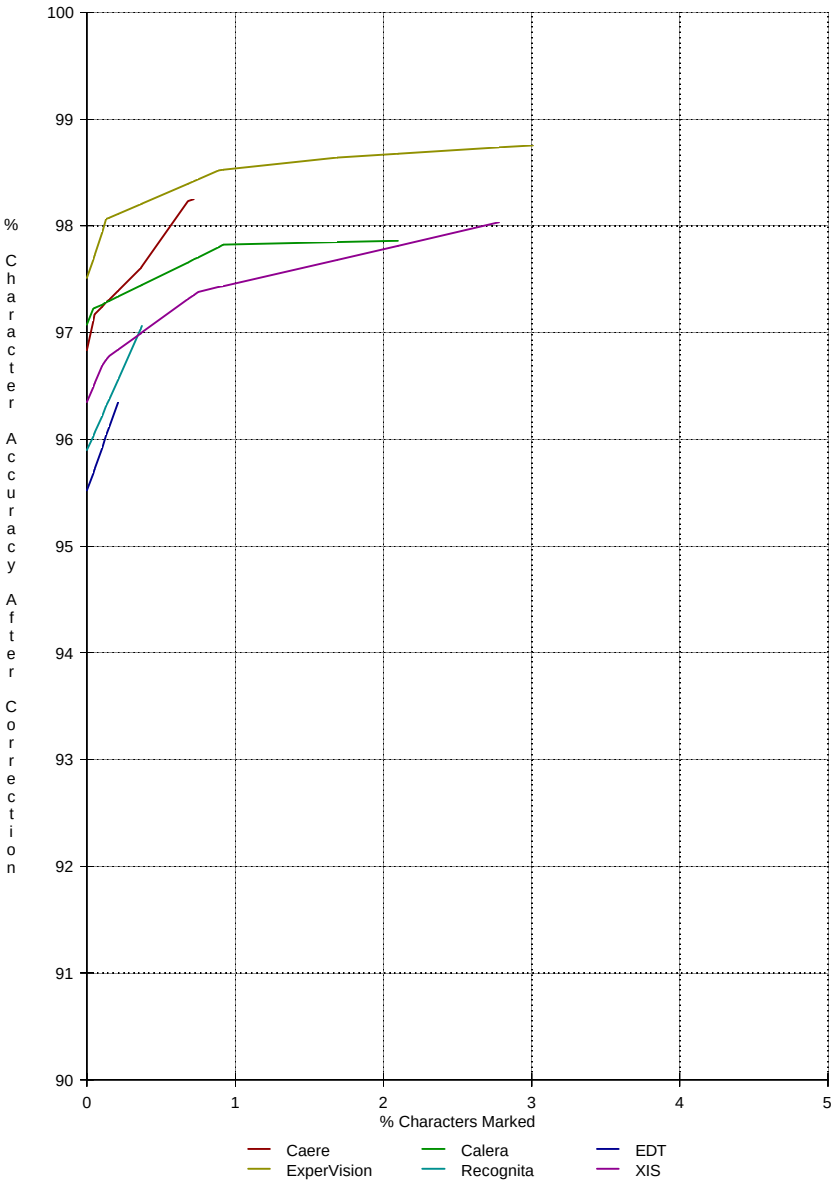
Graph 6a: Marked Character Efficiency

DOE Sample



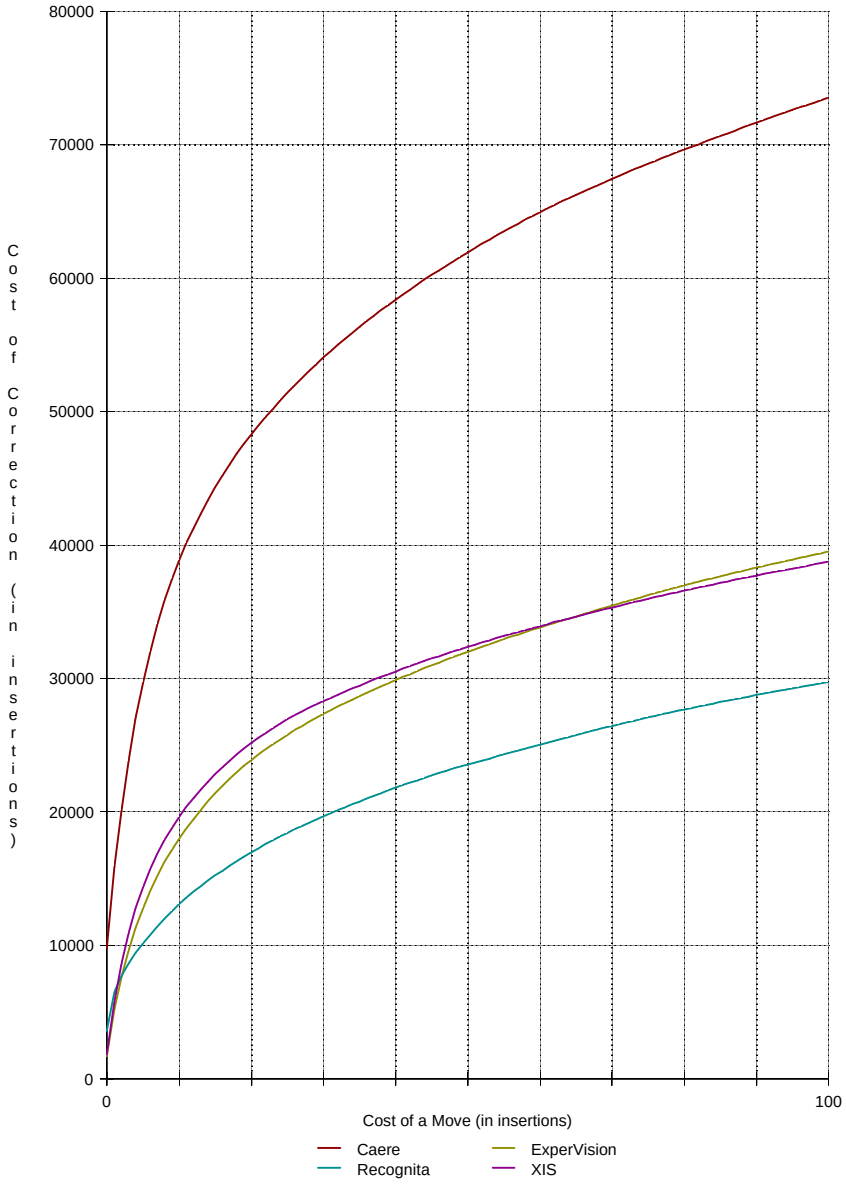
Graph 6b: Marked Character Efficiency

Magazine Sample



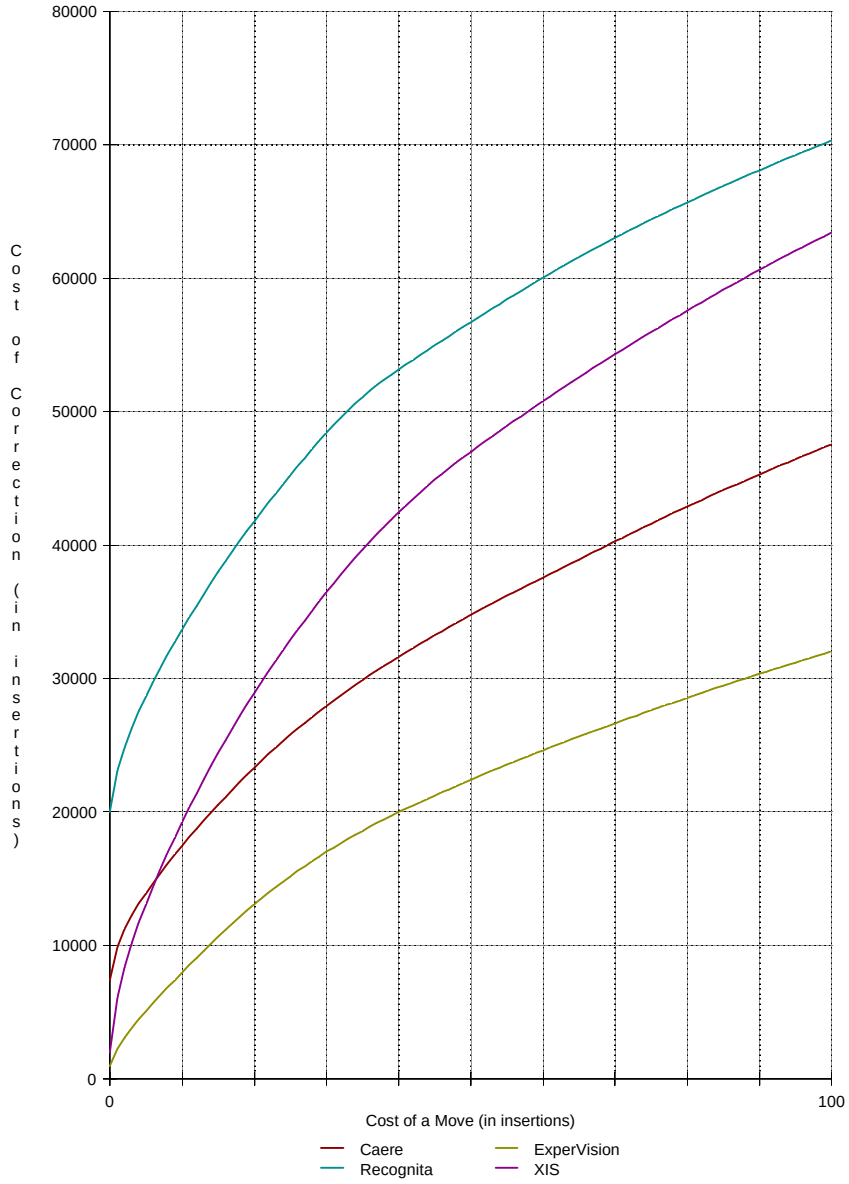
Graph 7a: Automatic Zoning

DOE Sample

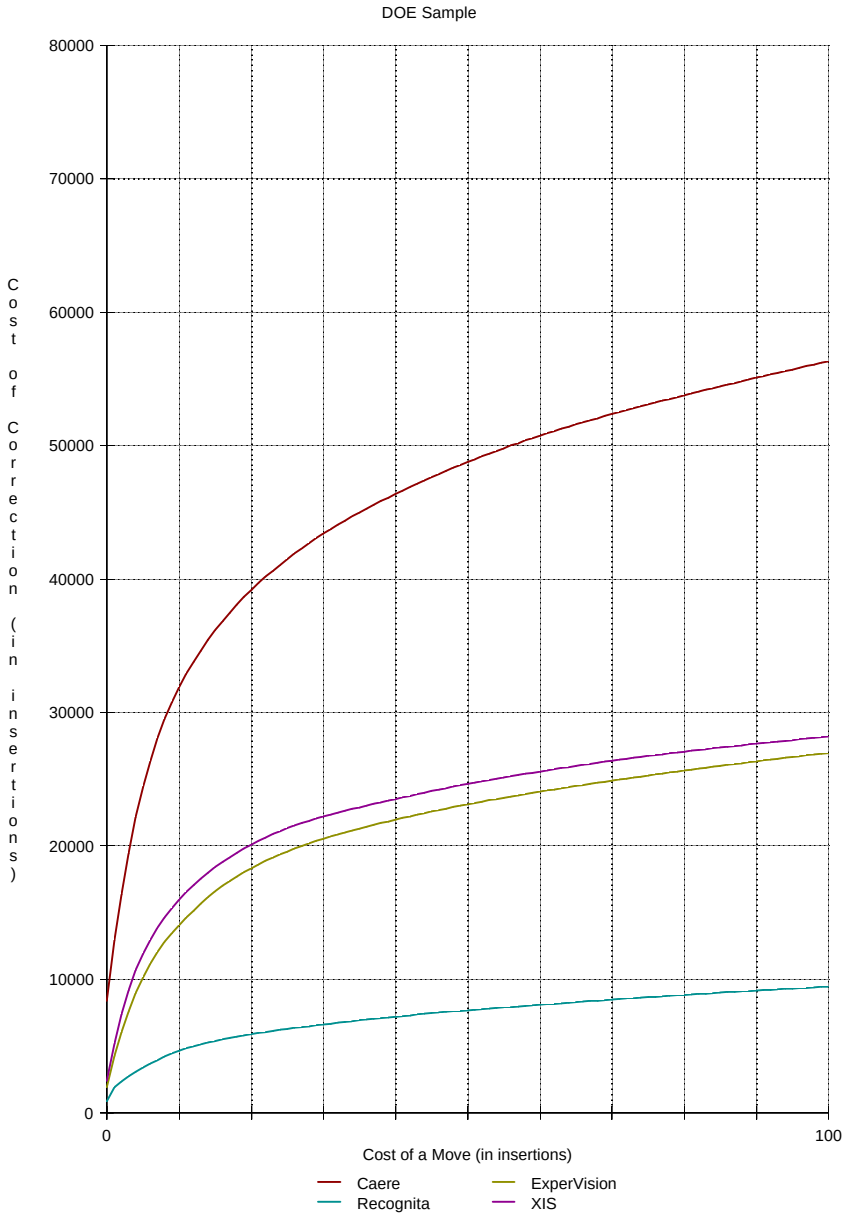


Graph 7b: Automatic Zoning

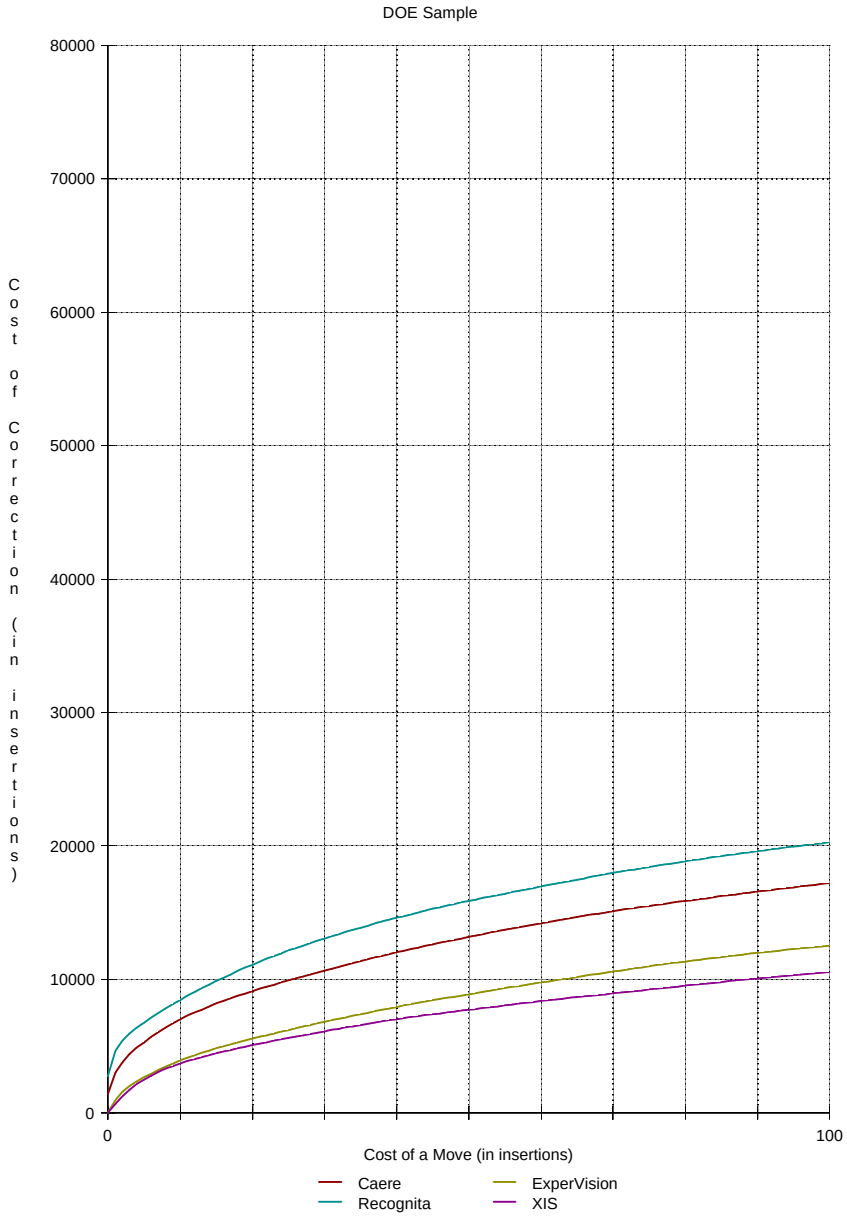
Magazine Sample



Graph 8a: Automatic Zoning, Table Pages

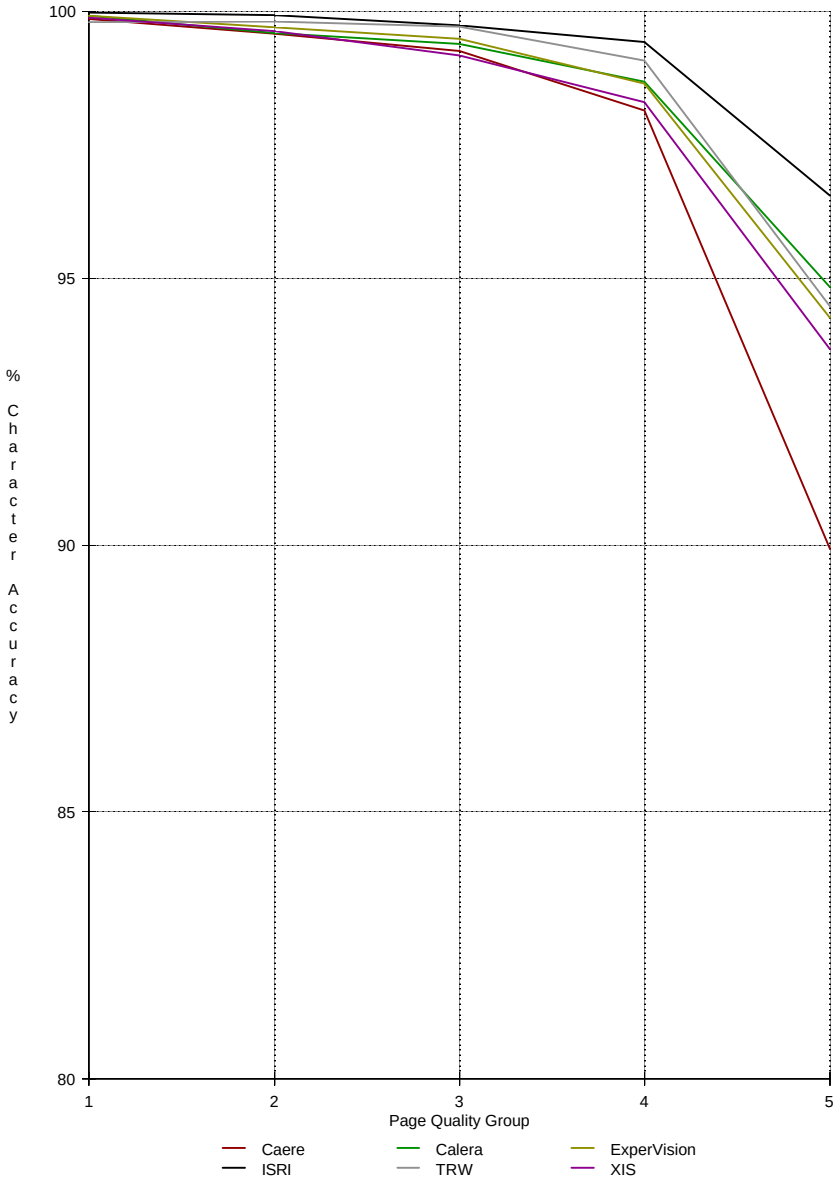


Graph 8b: Automatic Zoning, Other Pages



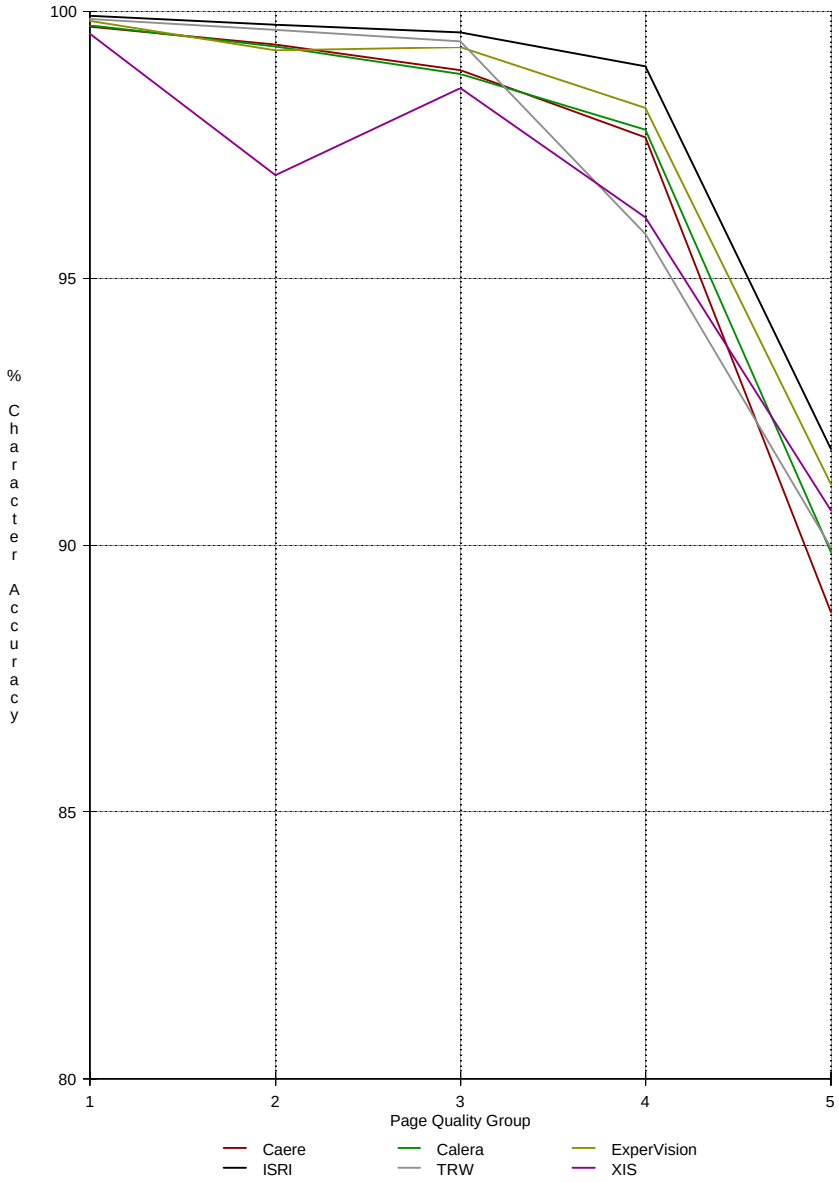
Graph 9a: Character Accuracy of Voting

DOE Sample



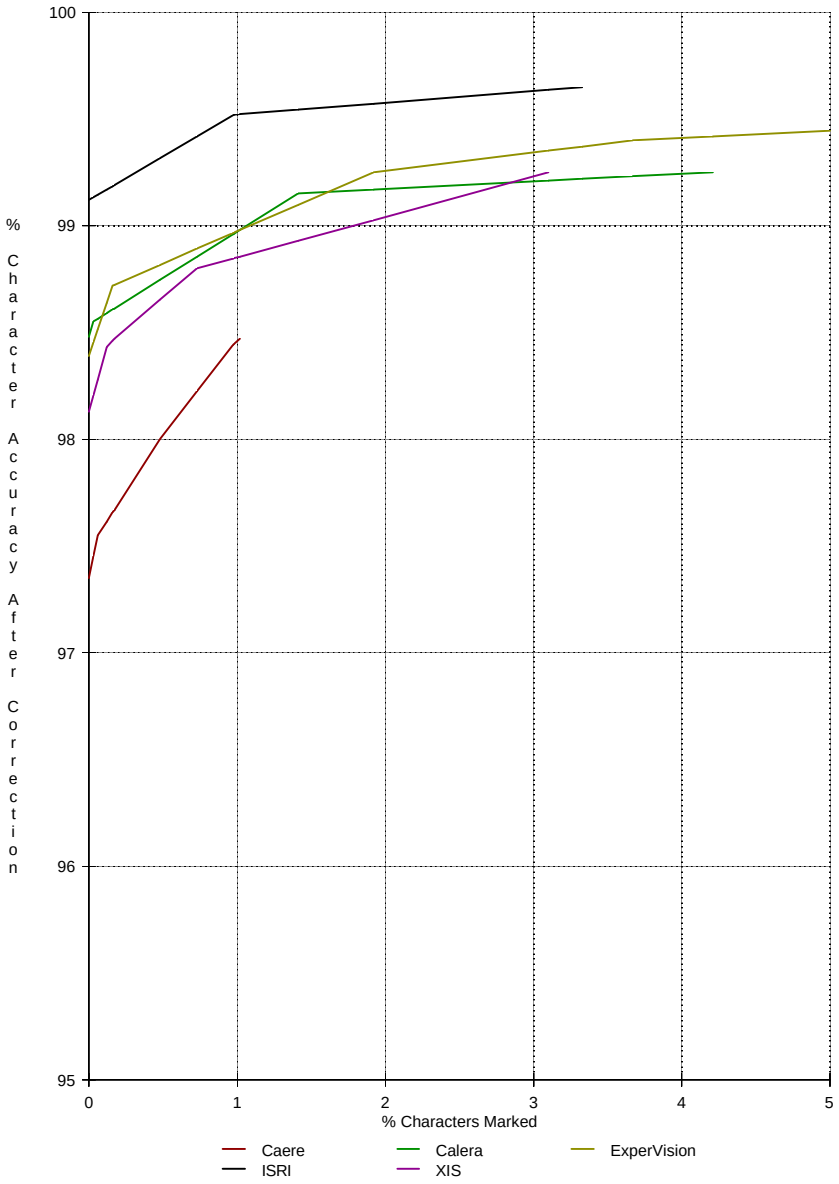
Graph 9b: Character Accuracy of Voting

Magazine Sample



Graph 10a: Efficiency of Voting Markers

DOE Sample



Graph 10b: Efficiency of Voting Markers

Magazine Sample