The Fourth Annual Test of OCR Accuracy

Stephen V. Rice, Frank R. Jenkins, and Thomas A. Nartker

1 Introduction

For four years, ISRI has conducted an annual test of optical character recognition (OCR) systems known as "page readers." These systems accept as input a bitmapped image of any document page, and attempt to identify the machine-printed characters on the page. In the annual test, we measure the accuracy of this process by comparing the text that is produced as output with the correct text. The goals of the test include:

1. to provide a current, independent assessment of system performance,
2. to measure the advances in the technology from year to year,
3. to gain insight into the complex nature of OCR, and
4. to identify problems at the state-of-the-art.

The scope of the test has increased greatly over the past four years. In the first test [Rice 92], six OCR systems processed binary images of 132 pages containing a total of 278,000 characters. These pages were randomly selected from a U.S. Department of Energy (DOE) database of scientific and technical documents. In the second year, new measures of performance were introduced in evaluating eight OCR systems using a larger DOE sample (460 pages and 817,000 characters) [Rice 93a, Kanai 93, Nartker 94a]. The third annual test re-used this DOE sample and featured a 200-page sample of articles from popular U.S. magazines scanned at three different resolutions. Six OCR systems were thus tested on pages containing nearly 1.5 million characters [Rice 94, Nartker 94b].

In this report, we present the results of the fourth annual test, our largest and most comprehensive to date. The test samples contain more than three million characters from business letters, DOE documents, and articles from magazines and newspapers. Each page has been scanned four times to produce binary images at three different resolutions, plus one gray scale image. Furthermore, fax images at two different resolutions have been obtained for each business letter page. We introduce our first non-English sample, which is a collection of Spanish-language newspaper articles, and for the first time, we report the speed of the OCR systems.

1.1 Participants

Any organization may participate in the annual test provided:

1. it submits a version of an OCR system by the established deadline (December 15, 1994 for the fourth annual test),
2. the version runs on a PC or Sun SPARCstation, and
3. the version can process specific regions of a TIFF image in a fully automatic (non-interactive) way.

Furthermore, only one entry is allowed per organization.

There are many features of OCR systems that are evaluated in this test. Submitted versions need not support all of these features. For example, if a version does not support automatic zoning or Spanish OCR, then it will simply be excluded from that portion of the test.

Table 1 lists the eight organizations that participated in this year's test, and the versions they submitted. Hewlett Packard Laboratories submitted a research prototype that operates on only an HP workstation. This was allowed because HP provided the hardware, and facilitated the interface, well in advance of the deadline.

1.2 Test Data

Five test samples were used in this year's test.

1. The *Business Letter Sample* contains a variety of letters received by businesses and individuals and donated to ISRI.
2. The *DOE Sample* is the third and largest sample we have prepared by randomly selecting pages from a DOE collection of scientific and technical documents.
3. The *Magazine Sample*, which was used in the third annual test, consists of pages selected at random from the 100 U.S. magazines having the largest circulation.
4. The *English Newspaper Sample* contains articles selected at random from the 50 U.S. newspapers having the largest circulation.
5. The *Spanish Newspaper Sample* contains articles selected at random from 12 popular newspapers from Argentina, Mexico, and Spain.

For the newspaper samples, only articles from the first section of the newspaper were selected, and each article was clipped from the newspaper.

Each test page was placed manually on the platen of a Fujitsu M3096G scanner, and then digitized four times to produce binary images at 200, 300, and 400 dots per inch (dpi), and an 8-bit gray scale image at 300 dpi. A global threshold of 127 (out of 255) was used to create the binary images for the Business Letter, DOE, and Magazine Samples. A different threshold was chosen for the newspaper samples: 75 for the English articles, and 95 for the Spanish articles.

Table 1: Participating Organizations

Organization	Version Name	Version No.	Platform	Version Type
Caere Corp. Los Gatos, California	Caere OCR	138.1	Sun SPARCstation	pre-release
Electronic Document Technology Pte. Ltd. Singapore	EDT ImageReader	3.0	PC DOS	commercial release
Hewlett Packard Laboratories Bristol, England	HP Labs OCR	7.0	HP workstation	research prototype
International Neural Machines Inc. Waterloo, Ontario	INM NeuroTalker	2.52	PC DOS	beta release
Ligature Ltd. Jerusalem, Israel	Ligature CharacterEyes Pro	2.6	PC Windows	beta release
MAXSOFT-OCRON, Inc. Fremont, California	MAXSOFT-OCRON Recore	3.2	PC Windows	beta release
Recognita Corp. Budapest, Hungary	Recognita OCR	3.0	PC Windows	beta release
Xerox Imaging Systems, Inc. Peabody, Massachusetts	XIS OCR Engine	10.5	Sun SPARCstation	beta release

Table 2: Test Data

	Pages	Zones	Words	Characters
Business Letter Sample	200	1,419	51,460	319,756
DOE Sample	785	2,280	213,552	1,463,512
Magazine Sample	200	1,414	114,361	666,134
English Newspaper Sample	200	781	84,026	492,080
Spanish Newspaper Sample	144	558	57,670	348,091
Total	1,529	6,452	521,069	3,289,573

We created fax images of the Business Letter Sample by transmitting each page locally, using standard and fine modes, from a Xerox 7024 fax machine to a fax modem. The standard-mode fax images have a resolution of 204 dpi in the X-direction, and 98 dpi in the Y-direction. The fine-mode images are also 204 dpi in the X-direction, but are 196 dpi in the Y-direction.

We manually "zoned" each page, i.e., we delineated and ordered the text regions of the page. The OCR systems processed only these "zones." Some text was deemed to be unsuitable for the test, and was excluded; examples include equations, advertisements, text that is part of a figure (such as the labelling of a graph or map), and text that is considered to be unreadable by humans.

We carefully prepared the correct text, or "ground-truth," corresponding to each zone. To ensure the highest possible accuracy, the text for each zone was entered four times, by different typists working independently. The four versions were reconciled with the help of a difference algorithm.

Table 2 gives the number of pages, zones, words, and characters in each test sample.

1.3 Test Operation

Version 5.0 of the OCR Experimental Environment was used to conduct the fourth annual test. This is a suite of software tools developed by ISRI for large-scale, automated testing and experimental research in OCR. An earlier version of this software is described in [Rice 93b].

This software runs on Sun SPARCstations and provides remote control of PCs. Each OCR system is operated in a fully automatic manner, i.e., without human interaction. The comparison of OCR-generated text with correct text, and the tabulation of accuracy statistics, are performed entirely under computer control.

Pages were not re-scanned for each OCR system, nor were page images re-zoned; all OCR systems processed the same zoned portions of the same page images. Unless otherwise noted, tests were run using the 300 dpi binary and gray scale images. Exceptions are the tests involving the fax business letters, and the test of the effect of resolution, which also utilized the 200 and 400 dpi binary images.

Caere OCR and the XIS OCR Engine were operated under SunOS 4.1.3 on a single-processor Sun SPARCstation 10 with 64 megabytes of memory. The five PC-based OCR systems performed on identically-configured 486DX/33 machines with 8 megabytes of memory, running under MS-DOS 5.0, and for three of these, MS Windows 3.1. HP Labs OCR was operated under HP-UX A.09.01 on an HP 9000 Model 735 with 32 megabytes of memory.

Each machine was unburdened when timing figures were recorded, i.e., only the OCR system was running on the machine. Each OCR system processed one page image per invocation; thus, the timing figures include the modest overhead of initializing the OCR system for each image.

2 Character Accuracy

While there are many ways of quantifying the deviation between OCR-generated and correct text, in our most fundamental measure, we reflect the effort required by a human editor to correct the OCR-generated text. Specifically, we compute the minimum number of edit operations (character insertions, deletions, and substitutions) needed to fully correct the text. We refer to this quantity as the number of *errors* made by the OCR system. Expressing this as a percentage of the total number of characters, we obtain the *character accuracy*:

$$\frac{\#characters - \#errors}{\#characters}.$$

In the past, we utilized an algorithm that tended to over-estimate the minimum number of edit operations by 5 to 10%. In the interest of precise reporting, this year we have switched to an algorithm that computes the minimum number exactly. It is an optimized version of an algorithm by Ukkonen [Ukkonen 85].

Tables 3a-3f give the character accuracy results for each test sample. Some entries are missing due to unsupported features. Only Caere OCR and HP Labs OCR accept gray scale input. EDT ImageReader, HP Labs OCR, and INM NeuroTalker do not support Spanish OCR.

2.1 Failures

A failure is detected when an OCR system "crashes" or "hangs" when processing a page image, or when it returns an error status upon termination. An entry of *none* in the *Failures* column indicates that no failures were detected. Otherwise, the number of failed pages is specified, followed by the number of characters on those pages, expressed as a percentage of the total number of characters in the sample. If the latter exceeds one percent, then the failures are deemed to be excessive, and the accuracy results are not reported; otherwise, errors are charged equal to the number of characters on the failed pages.

Not all failures are detected. Bugs could cause extraneous characters to be output, or could prevent the generation of correct characters. Failures of this type cannot be distinguished in an automatic and reliable way from recognition errors. They go undetected, and errors are charged in proportion to the editing effort needed to correct the damage.

If the character accuracy of an OCR system is less than 90% for a particular sample, then we note only that the accuracy falls below this threshold. We do not report further the performance of this system on this sample.

2.2 Confidence Intervals

Graphs 1a-1g show approximate 95% confidence intervals for character accuracy. These intervals were computed using a technique from statistics known as the *jackknife estimator* [Dudewicz 88]. In applying this technique, we have made the assumption that the pages within a sample are independent, but we have not assumed that the characters within a page are independent.

An OCR system that performs consistently within a sample is represented by a narrow interval, whereas a wide interval indicates considerable variability. When comparing the performance of two systems, non-overlapping intervals imply that there is a statistically significant difference between the systems.

Character Accuracy

Table 3a: Original Business Letters

	300 dpi Binary			300 dpi 8-bit Gray Scale		
	Errors	% Accuracy	Failures	Errors	% Accuracy	Failures
Caere OCR	4,459	98.61	none	3,102	99.03	none
EDT ImageReader	13,162	95.88	1 / 0.30	---	---	---
HP Labs OCR	5,959	98.14	none	4,850	98.48	none
INM NeuroTalker	---	< 90.00	none	---	---	---
Ligature CharacterEyes Pro	---	---	1 / 1.07	---	---	---
MAXSOFT-OCRON Recore	8,377	97.38	none	---	---	---
Recognita OCR	11,280	96.47	none	---	---	---
XIS OCR Engine	5,473	98.29	none	---	---	---

Table 3b: Fax Business Letters

	Standard-mode Fax			Fine-mode Fax		
	Errors	% Accuracy	Failures	Errors	% Accuracy	Failures
Caere OCR	18,361	94.26	none	7,559	97.64	none
EDT ImageReader	---	< 90.00	1 / 0.70	15,345	95.20	none
HP Labs OCR	---	< 90.00	none	8,815	97.24	none
INM NeuroTalker	---	< 90.00	none	24,552	92.32	none
Ligature CharacterEyes Pro	---	---	---	15,689	95.09	none
MAXSOFT-OCRON Recore	---	< 90.00	none	9,403	97.06	none
Recognita OCR	---	---	---	10,193	96.81	none
XIS OCR Engine	17,541	94.51	none	7,453	97.67	none

Table 3c: DOE Sample

	300 dpi Binary			300 dpi 8-bit Gray Scale		
	Errors	% Accuracy	Failures	Errors	% Accuracy	Failures
Caere OCR	37,503	97.44	2 / 0.50	32,791	97.76	1 / 0.33
EDT ImageReader	94,234	93.56	1 / 0.13	---	---	---
HP Labs OCR	36,349	97.52	none	33,390	97.72	none
INM NeuroTalker	---	< 90.00	none	---	---	---
Ligature CharacterEyes Pro	---	---	7 / 1.28	---	---	---
MAXSOFT-OCRON Recore	56,746	96.12	none	---	---	---
Recognita OCR	57,713	96.06	none	---	---	---
XIS OCR Engine	34,644	97.63	none	---	---	---

Table 3d: Magazine Sample

	300 dpi Binary			300 dpi 8-bit Gray Scale		
	Errors	% Accuracy	Failures	Errors	% Accuracy	Failures
Caere OCR	14,483	97.83	none	8,568	98.71	none
EDT ImageReader	---	---	2 / 2.02	---	---	---
HP Labs OCR	15,043	97.74	none	10,425	98.43	none
INM NeuroTalker	---	< 90.00	none	---	---	---
Ligature CharacterEyes Pro	41,563	93.76	none	---	---	---
MAXSOFT-OCRON Recore	23,312	96.50	none	---	---	---
Recognita OCR	26,474	96.03	none	---	---	---
XIS OCR Engine	16,784	97.48	none	---	---	---

Table 3e: English Newspaper Sample

	300 dpi Binary			300 dpi 8-bit Gray Scale		
	Errors	% Accuracy	Failures	Errors	% Accuracy	Failures
Caere OCR	5,079	98.97	none	7,478	98.48	none
EDT ImageReader	---	---	3 / 1.74	---	---	---
HP Labs OCR	6,432	98.69	none	5,125	98.96	none
INM NeuroTalker	47,773	90.29	none	---	---	---
Ligature CharacterEyes Pro	11,230	97.72	none	---	---	---
MAXSOFT-OCRON Recore	7,002	98.58	none	---	---	---
Recognita OCR	10,495	97.87	none	---	---	---
XIS OCR Engine	5,513	98.88	none	---	---	---

Table 3f: Spanish Newspaper Sample

	300 dpi Binary			300 dpi 8-bit Gray Scale		
	Errors	% Accuracy	Failures	Errors	% Accuracy	Failures
Caere OCR	5,394	98.45	none	---	---	1 / 1.44
Ligature CharacterEyes Pro	13,512	96.12	1 / 0.25	---	---	---
MAXSOFT-OCRON Recore	10,012	97.12	none	---	---	---
Recognita OCR	8,929	97.43	none	---	---	---
XIS OCR Engine	7,213	97.93	none	---	---	---

2.3 Speed and Throughput

For most applications, speed is less important than accuracy. Indeed, there is little use for a fast OCR system that produces mostly gibberish as output. But given OCR systems of comparable accuracy, speed becomes an important factor.

We are generally opposed to reporting raw speed figures without consideration of accuracy. Thus, we introduce the following *throughput* function which reports speed while penalizing for errors:

$$\frac{\#characters - P \times \#errors}{\#seconds}.$$

P represents the penalty assigned to each error. When $P = 0$, the function gives the raw speed in terms of characters per second. Some authors have defined throughput to mean "correct characters" per second, which corresponds to $P = 1$. We do not feel that this is a sufficient penalty for errors; hence, in Graphs 2a-2g, we present throughput for $P = 0$ to 10.

2.4 Accuracy by Character Class

In accuracy by character class, we divide the ground-truth characters into classes, and determine the percentage of characters in each class that were correctly recognized. The following classes were used.

1. *Spacing*: the blank and end-of-line characters,
2. *a - z*: the lowercase letters,
3. *A - Z*: the uppercase letters,
4. *0 - 9*: the decimal digits, and
5. *Special*: punctuation and other special symbols.

For the Spanish Newspaper Sample, a *Spanish* class was added containing the Spanish accented letters and punctuation symbols. Graphs 3a-3g display the results for each test sample.

The largest of these classes is the *a - z* class; depending on the sample, 68 to 75% of the ground-truth characters belong to this class. The second largest class is the *Spacing* class, accounting for 16 to 17% of the characters.

The *A - Z*, *0 - 9*, and *Special* classes contain 3-7%, 1-6%, and 3-5% of the characters, respectively. The *Spanish* class contains 2% of the characters in the Spanish Newspaper Sample. The OCR systems were less accurate on these smaller classes.

2.5 Effect of Resolution

Graphs 4a-4e show how character accuracy is affected by decreasing the resolution of binary images from 300 to 200 dpi, and by increasing it to 400 dpi. Graph 4a also includes the fax images. If the data point for a particular resolution is missing, then the OCR system had difficulty processing the images scanned at this resolution: either it made excessive failures or its accuracy was less than 90%.

Decreasing the resolution from 300 to 200 dpi caused a substantial increase in the number of errors: approximately a 50% increase for the Business Letter and DOE Samples, and a 75% increase for the Magazine and English Newspaper Samples. The number of errors jumped by 200% for the Spanish Newspaper Sample.

One would expect the number of errors to decrease by increasing the resolution from 300 to 400 dpi. In some cases it did by a small amount, but just as often, the number of errors increased. The higher resolution provided little or no advantage.

The fine-mode fax images have essentially the same resolution as the 200 dpi images. But when processing the former, the OCR systems made 5 to 15% fewer errors than they did on the latter. Upon inspection, we observed that the images created by the fax machine are "darker," and contain fewer broken characters, than the images produced using the Fujitsu scanner.

The standard-mode fax images presented a very difficult test for the OCR systems. More than twice the number of errors were made on these images than on the fine-mode images, and only two systems, Caere OCR and the XIS OCR Engine, attained a character accuracy above 90%. Two organizations, Ligature and Recognita, chose not to participate in this test.

2.6 Page Quality Groups

If we process a given page using several OCR systems, and determine the character accuracy of each system on this page, then we can compute the median of these accuracies to obtain a measure of the quality, or "OCR difficulty," of the page. We use this measure to divide the pages of each sample into five "Page Quality Groups" of approximately equal size. Group 1 contains the pages with the highest median accuracy (best page quality), and Group 5 contains the pages with the lowest median accuracy (worst page quality). In Graphs 5a-5g, the character accuracy within each group is plotted to show the effect of page quality.

A large percentage of the errors, about 50 to 60%, are made on the worst 20% of each sample, i.e., Group 5. For the DOE Sample, this percentage is even higher, roughly 70 to 80%. We can gain some insight into what makes OCR difficult by examining the images of pages belonging to Group 5. In Figures 1-5, we present snippets taken from these images. Each was reproduced from the 300 dpi binary image, and enlarged by 50% to make it easier to see the degradation.

Broken and touching characters, also known as "splits" and "joins," are a very common source of error, and occur in each test sample. When processing the business letters, the OCR systems encountered some difficulty reading letterheads, which are often printed in a stylistic manner; also, creases in the hard copy affected the recognition of entire lines. The DOE Sample contains many challenging tables, and some pages with skewed text and/or curved baselines. Text printed on a shaded background is common in magazine articles, and is a significant source of error. Bleedthrough and other irregularities of newsprint caused some speckling of the newspaper images.

3 Word Accuracy

A popular application of OCR is to build a text database from a collection of hard-copy documents. Information retrieval techniques can then be applied to locate documents of interest. In this environment, the percentage of words that are correctly recognized, or *word accuracy* of the

Word Accuracy

Table 4a: Original Business Letters

	300 dpi Binary				300 dpi 8-bit Gray Scale			
	Misrec. Words	Word Accuracy	Stopword Accuracy	Non-stopword Accuracy	Misrec. Words	Word Accuracy	Stopword Accuracy	Non-stopword Accuracy
Caere OCR	1,144	97.78	98.94	96.96	795	98.46	99.35	97.83
EDT ImageReader	3,654	92.90	95.72	90.92	---	---	---	---
HP Labs OCR	1,631	96.83	98.40	95.73	1,495	97.09	98.47	96.14
MAXSOFT-OCRON Recore	1,990	96.13	98.12	94.75	---	---	---	---
Recognita OCR	2,621	94.91	96.97	93.46	---	---	---	---
XIS OCR Engine	1,578	96.93	98.75	95.66	---	---	---	---

Table 4b: Fax Business Letters

	Standard-mode Fax				Fine-mode Fax			
	Misrec. Words	Word Accuracy	Stopword Accuracy	Non-stopword Accuracy	Misrec. Words	Word Accuracy	Stopword Accuracy	Non-stopword Accuracy
Caere OCR	4,643	90.98	95.62	87.73	1,998	96.12	98.11	94.72
EDT ImageReader	---	---	---	---	4,527	91.20	94.67	88.78
HP Labs OCR	---	---	---	---	2,421	95.30	97.39	93.83
INM NeuroTalker	---	---	---	---	7,605	85.22	91.98	80.49
Ligature CharacterEyes Pro	---	---	---	---	4,876	90.52	94.41	87.81
MAXSOFT-OCRON Recore	---	---	---	---	2,858	94.45	97.17	92.54
Recognita OCR	---	---	---	---	3,069	94.04	96.90	92.03
XIS OCR Engine	4,909	90.46	94.91	87.35	2,229	95.67	98.23	93.88

Table 4c: DOE Sample

	300 dpi Binary				300 dpi 8-bit Gray Scale			
	Misrec. Words	Word Accuracy	Stopword Accuracy	Non-stopword Accuracy	Misrec. Words	Word Accuracy	Stopword Accuracy	Non-stopword Accuracy
Caere OCR	9,386	95.60	98.05	94.24	8,298	96.11	98.61	94.73
EDT ImageReader	23,350	89.07	93.47	86.62	---	---	---	---
HP Labs OCR	7,826	96.34	98.97	94.87	7,208	96.62	99.09	95.26
MAXSOFT-OCRON Recore	15,451	92.76	96.49	90.70	---	---	---	---
Recognita OCR	16,674	92.19	95.69	90.25	---	---	---	---
XIS OCR Engine	9,239	95.67	98.44	94.13	---	---	---	---

Table 4d: Magazine Sample

	300 dpi Binary				300 dpi 8-bit Gray Scale			
	Misrec. Words	Word Accuracy	Stopword Accuracy	Non-stopword Accuracy	Misrec. Words	Word Accuracy	Stopword Accuracy	Non-stopword Accuracy
Caere OCR	3,659	96.80	97.89	96.05	1,992	98.26	99.09	97.68
HP Labs OCR	4,566	96.01	97.47	94.99	3,458	96.98	98.19	96.13
Ligature CharacterEyes Pro	11,617	89.84	91.95	88.37	---	---	---	---
MAXSOFT-OCRON Recore	6,595	94.23	95.79	93.15	---	---	---	---
Recognita OCR	6,261	94.53	96.26	93.32	---	---	---	---
XIS OCR Engine	4,923	95.70	97.32	94.56	---	---	---	---

Table 4e: English Newspaper Sample

	300 dpi Binary				300 dpi 8-bit Gray Scale			
	Misrec. Words	Word Accuracy	Stopword Accuracy	Non-stopword Accuracy	Misrec. Words	Word Accuracy	Stopword Accuracy	Non-stopword Accuracy
Caere OCR	1,181	98.59	99.10	98.24	1,506	98.21	98.80	97.79
HP Labs OCR	1,946	97.68	98.73	96.94	1,505	98.21	98.96	97.67
INM NeuroTalker	13,989	83.35	88.47	79.71	---	---	---	---
Ligature CharacterEyes Pro	3,646	95.66	96.80	94.85	---	---	---	---
MAXSOFT-OCRON Recore	2,219	97.36	98.22	96.74	---	---	---	---
Recognita OCR	2,948	96.49	97.88	95.50	---	---	---	---
XIS OCR Engine	1,892	97.75	98.64	97.11	---	---	---	---

Table 4f: Spanish Newspaper Sample

	300 dpi Binary	
	Misrec. Words	Word Accuracy
Caere OCR	2,193	96.20
Ligature CharacterEyes Pro	5,785	89.97
MAXSOFT-OCRON Recore	5,015	91.30
Recognita OCR	3,400	94.10
XIS OCR Engine	2,966	94.86

OCR-generated text, is an important measure.

We define a word to be any sequence of one or more letters. A word is considered to be correctly recognized if all of its letters have been correctly identified. Since full-text searching is normally performed on a case-insensitive basis, a letter that is generated in the wrong case (e.g., *C* for *c*) is still considered to be correct.

Tables 4a-4f give the number of misrecognized words and the word accuracy for each test sample. Graphs 6a-6g show the word accuracy within each Page Quality Group.

3.1 Stopwords and Non-stopwords

Stopwords are common words such as *the*, *of*, *and*, *in*, etc. Since these words have little retrieval value, they are normally not indexed by text retrieval systems. Since users search only for *non-stopwords*, it is the percentage of non-stopwords that are correctly recognized, or *non-stopword accuracy*, that is especially of interest.

We utilize the default set of 110 stopwords from the BASISPLUS text retrieval system [IDI 90]. About 40% of the words in the test samples are stopwords. Tables 4a-4e show the stopword and non-stopword accuracy for each test sample. (These were not computed for the Spanish Newspaper Sample.)

3.2 Distinct Non-stopword Accuracy

Suppose a user wishes to find every document in the database that contains a specific term (non-stopword). If a document containing this term is to be located, then at least one occurrence of the term must have been correctly recognized by the OCR system. With this in mind, we introduce a measure called *distinct non-stopword accuracy.*

For a given page, we refer to each unique term occurring on the page as a *distinct non-stopword*, and we say that it is correctly recognized if at least one of its occurrences has been correctly identified. Distinct non-stopword accuracy is the percentage of distinct non-stopwords that are correctly recognized.

It has been argued that full-text searching is resilient to OCR errors because of the inherent redundancy of text. Because a search term may occur many times within a document, it is unlikely that the OCR system will misrecognize every occurrence. Under the assumption of independence, if OCR-generated text has an overall non-stopword accuracy of 90%, then for a term occurring n times, the probability that every occurrence has been misrecognized is 10^{-n}. A counter-argument can be put forward that an image defect that causes the OCR system to misrecognize one occurrence may be present in every occurrence, making it considerably more likely that every occurrence will be missed.

In Graphs 7a-7f, we show how often at least one occurrence is correctly recognized as the number of occurrences increases from one to four. Contrast these graphs with Graph 7g, which illustrates the expected shape of the curves under the independence assumption.

3.3 Phrase Accuracy

Users also search for documents containing specific phrases. We define a phrase of length n to be any sequence of n words. A phrase is correctly recognized if all of its words have been correctly identified. *Phrase accuracy* is the percentage of phrases that are correctly recognized. Graphs 8a-8g show the accuracy for phrases of lengths one through eight. Note that the phrase accuracy for length one is equal to the word accuracy.

Phrase accuracy provides a useful measure of "error bunching." Given two OCR systems with the same word accuracy, the one with the lower phrase accuracy has produced errors that are more widely dispersed throughout the text.

4 Marked Character Efficiency

Finding and correcting errors in OCR-generated text can be a tedious and expensive process for users. But an OCR system provides some assistance by flagging the generated characters that it believes are most likely in error. A *reject character* (~) is placed in the output when the OCR system is unable to recognize a character. Also, the system may place a *suspect marker* (∧) before each character that is generated with low confidence. We refer to reject characters, and characters marked as suspect, as *marked characters*.

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

In the above sentence, there are three *marked errors*: the two reject characters and the "l" in "contains." The second "a" in "markers" is an *unmarked error*. The "d" in "and" is a *false mark*, which is a correctly-generated character that is marked as suspect.

In marked character efficiency, we measure the utility of the marked characters produced by an OCR system. Graphs 9a-9c display curves that show how the character accuracy of the OCR-generated text increases as a human editor examines more and more marked characters and corrects the marked errors. Initially, this process is very efficient as the editor corrects the errors identified by reject characters, and then the first level of suspect markers. But as the number of suspect markers increases, the curves flatten due to an increasing percentage of false marks.

Marked characters make it possible for an editor to inspect only one-half of one percent of the OCR-generated text yet correct 20 to 45% of the errors in the text. The editor may examine more characters than this, but the operation becomes considerably less efficient.

5 Automatic Zoning

In every test discussed so far, the OCR systems were given the coordinates of the text regions to process. In the test of automatic zoning, this information was not provided, and each system was asked to locate the text regions, and determine their correct reading order.

To measure how well this task is performed, we apply an algorithm that estimates the number of character insertions and block move operations needed to correct the automatic zoning errors. If an OCR system does not find a text region, then insertions are required to enter the missing text. If text blocks are out of order, then move operations are needed to re-order them. Using a conversion factor to express each move operation in terms of an equivalent number of insertions, the cost of correction is ultimately given solely in terms of insertions. Details of this method are presented in [Kanai 95].

Graphs 10a-10c present the results of this test. The cost of correction is plotted for a range of conversion factors, and has been normalized using the number of characters in each sample to allow for comparison across samples. HP Labs OCR and INM NeuroTalker do not support automatic zoning, and are thus missing from the graphs. The other curves that are missing are due to excessive failures. The XIS OCR Engine performed the best overall on this test.

The cost of correction was least for the English Newspaper Sample. Since each article was clipped from the newspaper, the OCR systems contended primarily with a headline above one or more columns of text. But the magazine articles were not clipped, and are part of a page layout that may be very creative (i.e., complicated). The DOE pages presented the challenge of distinguishing tables from multi-column text; the OCR system should "de-columnize" the latter, but not the former.

6 Comparison of OCR Systems: Accuracy and Speed

6.1 Binary Input

The best overall accuracy on binary images was achieved by Caere OCR, HP Labs OCR, and the XIS OCR Engine. We observed no significant differences in accuracy among these systems, with one exception: HP Labs OCR was out-performed by the other two systems on the standard-mode fax images.

But significant differences in speed were evident. The XIS OCR Engine was 2.3 to 4.4 times faster than Caere OCR, depending on the sample. Although Caere OCR and HP Labs OCR operated on different platforms, it is apparent that the former was much faster than the latter. When HP Labs OCR was submitted for this test, an HP Labs representative indicated that this version had not been optimized and was not expected to be competitive in terms of speed.

MAXSOFT-OCRON Recore and Recognita OCR comprise a second tier of systems. Although these two systems were comparable in terms of accuracy, Recognita OCR performed with roughly twice the speed.

A third tier of systems consists of EDT ImageReader, INM NeuroTalker, and Ligature CharacterEyes Pro. An interesting feature of INM NeuroTalker is the ability to adjust the trade-off of speed versus accuracy. At the request of INM, we began to test this system using the setting for highest accuracy (and slowest speed). But after encountering too many failures for the system to remain in the test, at INM's request we changed the setting to one that decreased the accuracy, increased the speed, and circumvented the failures. (INM attributes the failures to a faulty DOS extender that was shipped with the system.)

Since the Magazine Sample was used in the third annual test, it can serve as a yardstick to measure progress in the past year. Four organizations participated in both the third and fourth annual tests: Caere, EDT, Recognita, and XIS. In this year's test, the systems from Caere and XIS made 27 to 28% fewer errors on this sample than their predecessors made a year ago. The Recognita version made 6% more errors this year than last year. No comparison was possible for the EDT version because of failures.

6.2 Gray Scale Input

Caere OCR and HP Labs OCR utilize fundamentally different approaches when processing a gray scale image. HP Labs OCR first “binarizes” the image, i.e., creates a binary image from the gray scale image, and then recognizes the characters on the binary image. On the other hand, Caere OCR recognizes the characters directly from the gray scale image.

It is clear that gray scale input offers advantages over binary input in terms of accuracy. Depending on the sample, these systems produced 10 to 40% fewer errors when given gray scale input. But there was one exception: Caere OCR made substantially more errors when processing the gray scale images of the English Newspaper Sample. This may have been caused by bleedthrough, which is often visible in the gray scale image of a newspaper, but mostly disappears after the binarization process.

It appears that gray scale input was of greatest value when recognizing text printed on a shaded background, which is commonly found in the Magazine Sample. The binary image of this text is usually speckled and problematic, but given a gray scale image, an OCR system has a much better chance of separating the text from its background.

Gray scale images require considerably more storage and take longer to process than binary images. On average, Caere OCR needed twice as much time, and HP Labs OCR needed 20% more time, to process the gray scale input.

7 Summary

Eight organizations submitted OCR systems for the fourth annual test. These systems processed bitmapped images of more than 1,500 pages of business letters, scientific documents, and magazine and newspaper articles. The analytic tools of the OCR Experimental Environment were used to compare the OCR-generated text with the correct text, and compute several measures of performance. These include character, word, non-stopword, and phrase accuracy, and three new measures: throughput, accuracy by character class, and distinct non-stopword accuracy. Also, the effect of page quality was observed, the utility of marked characters was gauged, and the cost of correcting automatic zoning errors was estimated.

The accuracy of the OCR systems declined dramatically when the resolution of the images was reduced from 300 to 200 dpi; however, little or no benefit was obtained by increasing the resolution to 400 dpi. Fax images, especially the standard-mode variety, presented a significant challenge. The improved accuracy obtained from gray scale input demonstrates that this is an important new direction.

With the assistance of confidence intervals, we partitioned the eight systems into three tiers based on the accuracy results for binary input. We then noted differences in speed within tiers. This “ranking” reflects the performance of these systems on this test; their relative performance may vary when processing other types of documents, or when processing similar documents under different test conditions. Finally, we wish to emphasize that ISRI does not endorse any particular OCR system or systems.

Acknowledgment

We acknowledge the help of Dr. Ashok Singh and Dr. George Nagy in establishing the confidence interval method for character accuracy. Also, "Accuracy by Character Class" was suggested by Junichi Kanai.

References

[Dudewicz 88] Edward J. Dudewicz and Satya N. Mishra. *Modern Mathematical Statistics*, pages 743-748. John Wiley & Sons, 1988.

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

[Kanai 93] Junichi Kanai, Stephen V. Rice, and Thomas A. Nartker. A preliminary evaluation of automatic zoning. Technical Report 93-02, Information Science Research Institute, University of Nevada, Las Vegas, April 1993.

[Kanai 95] Junichi Kanai, Stephen V. Rice, Thomas A. Nartker, and George Nagy. Automated evaluation of OCR zoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17(1):86-90, January 1995.

[Nartker 94a] Thomas A. Nartker, Stephen V. Rice, and Junichi Kanai. OCR accuracy: UNLV's second annual test. *Inform*, Association for Information and Image Management. 8(1):40+, January 1994.

[Nartker 94b] Thomas A. Nartker and Stephen V. Rice. OCR accuracy: UNLV's third annual test. *Inform*, Association for Information and Image Management. 8(8):30+, September 1994.

[Rice 92] Stephen V. Rice, Junichi Kanai, and Thomas A. Nartker. A report on the accuracy of OCR devices. Technical Report 92-02, Information Science Research Institute, University of Nevada, Las Vegas, March 1992.

[Rice 93a] Stephen V. Rice, Junichi Kanai, and Thomas A. Nartker. An evaluation of OCR accuracy. Technical Report 93-01, Information Science Research Institute, University of Nevada, Las Vegas, April 1993.

[Rice 93b] Stephen V. Rice. The OCR experimental environment, version 3. Technical Report 93-04, Information Science Research Institute, University of Nevada, Las Vegas, April 1993.

[Rice 94] Stephen V. Rice, Junichi Kanai, and Thomas A. Nartker. The third annual test of OCR accuracy. Technical Report 94-03, Information Science Research Institute, University of Nevada, Las Vegas, April 1994.

[Ukkonen 85] Esko Ukkonen. Algorithms for approximate string matching. *Information and Control* 64:100-118, 1985.

Figure 1: Examples from Page Quality Group 5, Business Letter Sample

EMBASSY OF
COMMERCIAL
CHICAG

Michael J. Deasy
Public Information Officer
Calif. State Department of
Transportation

quarter page adve
your support agai
your convenience.

at the Solutions Centre. E
additions to both the Tech
opening Conference Briefin

organization receiving
expert information on t
research and maintains

UNITED STATES DEP
National Institute of Sta
Gaithersburg, Maryland 2089

Please forward them t
returned to us together

I would highly re
organization. Sh

10140 MESA RIM ROAD
SAN DIEGO, CALIFORNIA 92121
800-334-9191 619-453-9191
FACSIMILE: 619-453-9294

Don't miss out on the latest
fyi/im newsletter and the So

Incentive Award - A mid
FAXed images with a sys

Service Representatives: sin

We have enclosed a C

P.S. This enroll
by June 15, 1993

$49.95! And this is **not**
"display quality"--then
Completely scalable, For

With Discover Card yo
Insurance⁺. Every ti
flight insurance auto
charging your airline

month after your account is opened, and may
waived first year; $40 each year thereafter for
minimum, $25 maximum. Transaction fee for
transaction fees for two special Premium Acce

Figure 2: Examples from Page Quality Group 5, DOE Sample

Operation of pressure
transducer tensio-
meters in an infil-

9. Precision

9.1 Criteria for judging the a
the maximum density and opti

Erosional cuts on
streams; sand dunes
at Kings Beach

Pinyon-juniper woodland is
Mountains or Slate Range; but th
7,000 feet in the Panamint Mount
Lake and in the northern Argus M
Lake. Associated with this zone

LEWELLEN, W. S., "The
of the Tornado Vortex," *Pr*
Symposium on Tornadoes:

Acanthite, summary of thermodyna
data, 195
Activity, of aqueous species, 36, 37,
93-96, 98, 114, 141, 143 145, 147

In reactor fuel elements a
approximately unity is not u
mixing is to be expected and.

6 to 60 are shown, as well as the p
and permeabilities for samples of si
varying according to equation 26 t
sample volume of 1035 cm cubed f

Activity provides sit
data that may have si
final design of repos

	12	3:46:56.12	CSNL
	20	13:15:47.30	
	20	13:18:44.50	
MAY	6	6:17:13.99	

1057	37.4	119.0
401	33.7	118.1
2300	37.6	118.9

SE of Hawthorne, Nev.	6.5	38.3
Parkfield	6.0	35.9
Southern California	5.1	34.1
N of Bishop	5.0	37.5

FIGURE 37.—Salt pool and collapse struc
northwest of Badwater. Drawn b

COMPOSITION

would bring about dehydrati
acid sites (i.e. surface alumini
ions). The sites which chemis

engineered barriers must be desi
The *ground water protection*
requirements (40 CFR 191.16) foc
the quality of any "special sourc

consideration for the
heterogeneities and 2
thermally inducted sa

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

ter. Still, much remains mysterious: W
cent? The patient had no French ances
ventured from his hometown of Worc

enough to justify preventive
ments. Even if these medicatio
sense in my case, though, I'd l

achieved the same results, bu
they do it?
Answer: In years past, the s
were fit with a considerable a

GALLONS USED
DAILY BY A FAMILY

1 Style. Liquid (**L**) or powder (**P**).
2 Calories. Per eight-fluid-ounce
3 Carbohydrates. Percent by we

the cheapest plastic lenses with
coating. The prices they were q
by as much as 75 percent from

Summer's fresh har
sauces and marina

PARKS
*# Children under 3 admitted free. * Prices do not inclu*
Walt Disney World Resort #
Lake Buena Vista, Florida; (407) 824-4321

TOP 10 TAPE RENTALS
UNFORGIVEN Clint Eastwood, *Warner*.........
THE BODYGUARD Kevin Costner, *Warner*.....

Years ago, in a group-th
session, I heard a story
have never forgotten.

on scientific research, said Dr.
ic McDuffie, former senior vic
dent for medical affairs for the
tis Foundation.

Opry, **Garth Brooks** met
backstage with cancer pa-
tient **Libby Sharp** of Gatlin-

independent front and rear suspension,
ic damping, and speed-sensitive rack-and
steering. These underpinnings allow Cad

WICKENBURG, ARIZONA
The Meadows
A leading treatment center for

Taste buds take note. *Tray Gourm*
Your Own Chef in the College C
(Lake Isle Press, $10.95) is the boo

cooked on the stovetop
kitchen comfortably coo

Figure 4: Examples from Page Quality Group 5, English Newspaper Sample

TALK OF THE DAY:
"Names change for po-
litical, geographical or

By SUSAN MILLIGAN
News Washington Bureau
WASHINGTON — With V

Most likely to be stolen:
Volkswagen Cabriolet
Ford Mustang convertible
Cadillac DeVille two-door

Previous positions: Chairman of
Cisneros Asset Management Co.
1989-92; Mayor of San Antonio,
1981-89; San Antonio City Council,

keeper predicted yesterday hi
will be indicted on embezz
charges, signaling a glitch in

The Inquirer wants its news report
be fair and correct in every respec
you have a question or comment
about news coverage, write to

them because of who he wants
The distinctive marks turn of
spective employers, threatenin
life he plans away from the st

Washington Post
THIMPU, Bhutan — Its cit
seldom write letters and there

and conversations
with world leader

neighbors and then leading the
the hill toward a new life.
Federal and state officials say
him a lot.

cholera in the crammed Rwan
camps of eastern Zaire is goi
United Nations said yesterda
But as the threat from ch

heretofore unlabeled wines: Cl
teau Margaux; Chateau Lafite; Ro
ert Mondavi cabernet sauvigno

Belize, a small Central Ame
country, use plant medicines,
losses of forests, and a lack of i
est among young people in beco

The village must rely on
the McHenry County Sher-
iff's Department in **Wood-**
stock to enforce the limit,

ing a "large number of casua
spokesman Maj. Dacre Hollow
In response, U.N. officials w

unemployment rate decreased to
the work force in March from 8.3
ruary, but rose from 7.1% in Ma
according to SCB, the national

Figure 5: Examples from Page Quality Group 5, Spanish Newspaper Sample

nández, exclamó indignado que la
dor bonaerense "es volver al '49",
constitucional sancionada ese año
dente Juan Domingo Perón, que lo

Yemen del Sur, aprovechan
una breve "tregua" para
fumar un cigarrillo. (Reuter)

PARANA (Enviado es-
pecial). — Hombres con
armas y largavistas apos-
tados en las cúpulas de la

El extenso debate por
Constituyente, ganad
la posición conjunt

Pensiones sociales

La Caja de Pensiones So
les-Ley 5110, delegación Rosa

En declaraciones radiales
senador porteño señaló que e
informe elaborado en el ám
de la Auditoría General de

–¿Declararán cuando se
llame a indagatorias?
–No. Existe el derecho a call
eso no tiene nada que ver con la

sospechoso. Desde esferas social
políticas se consideró que el a
había sido Guillermo Luque, hij
entonces diputado nacional por e

IXTAPALUCA (Notimex).
enfrentamiento entre un grupo
muneros y agentes de la Dire
General de Seguridad Públ

, el doctor Jorge Carpizo
a Porfirio Muñoz Ledo,
rencorosos del PRD por el

De acuerdo con el plan de r
la compañía, que aún no se d
totalidad, la intención es dive
servicios con la construcción

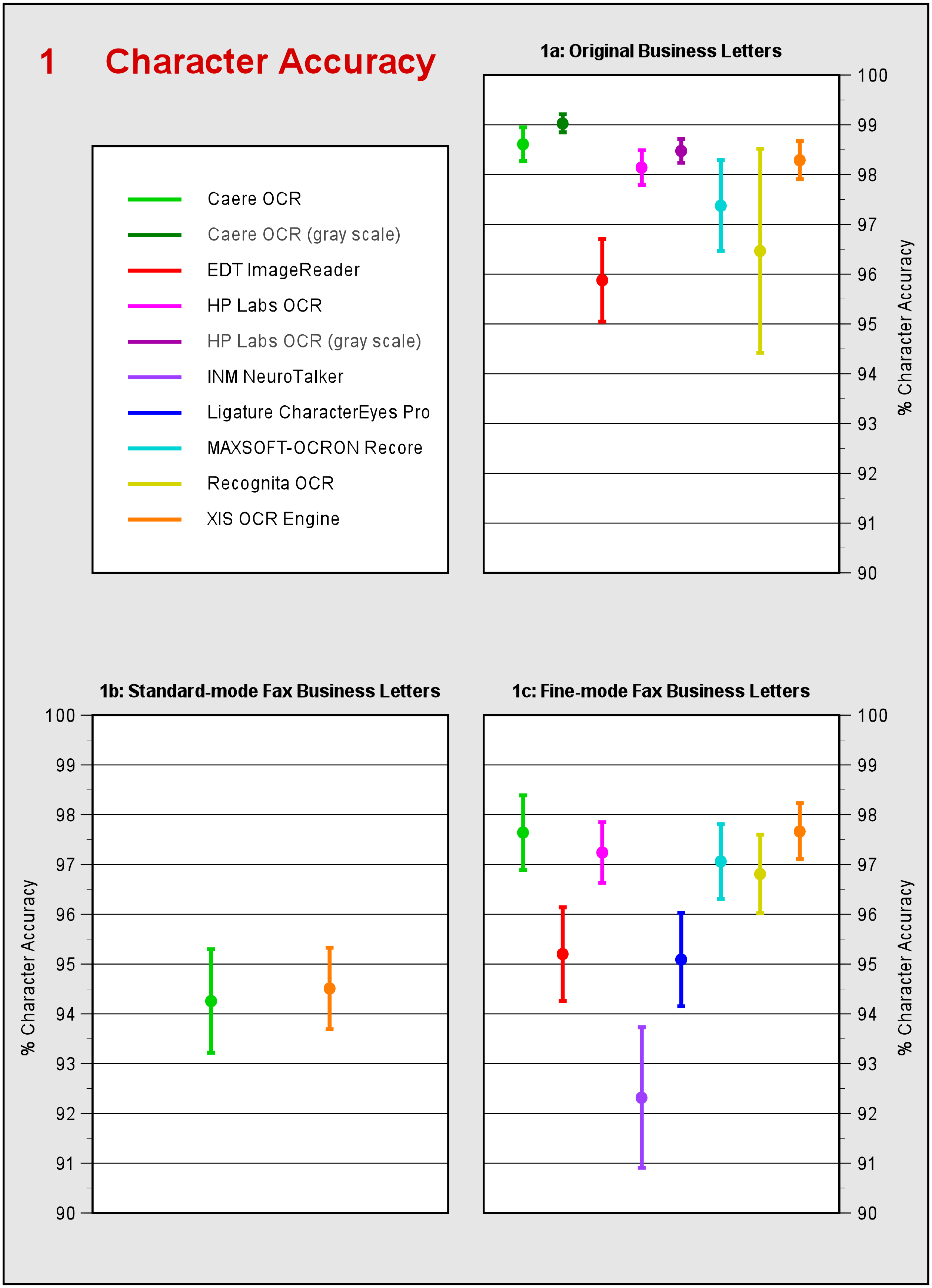
Presidencia de la Repúblic
Zedillo, en materia económi
está siendo difundida por as

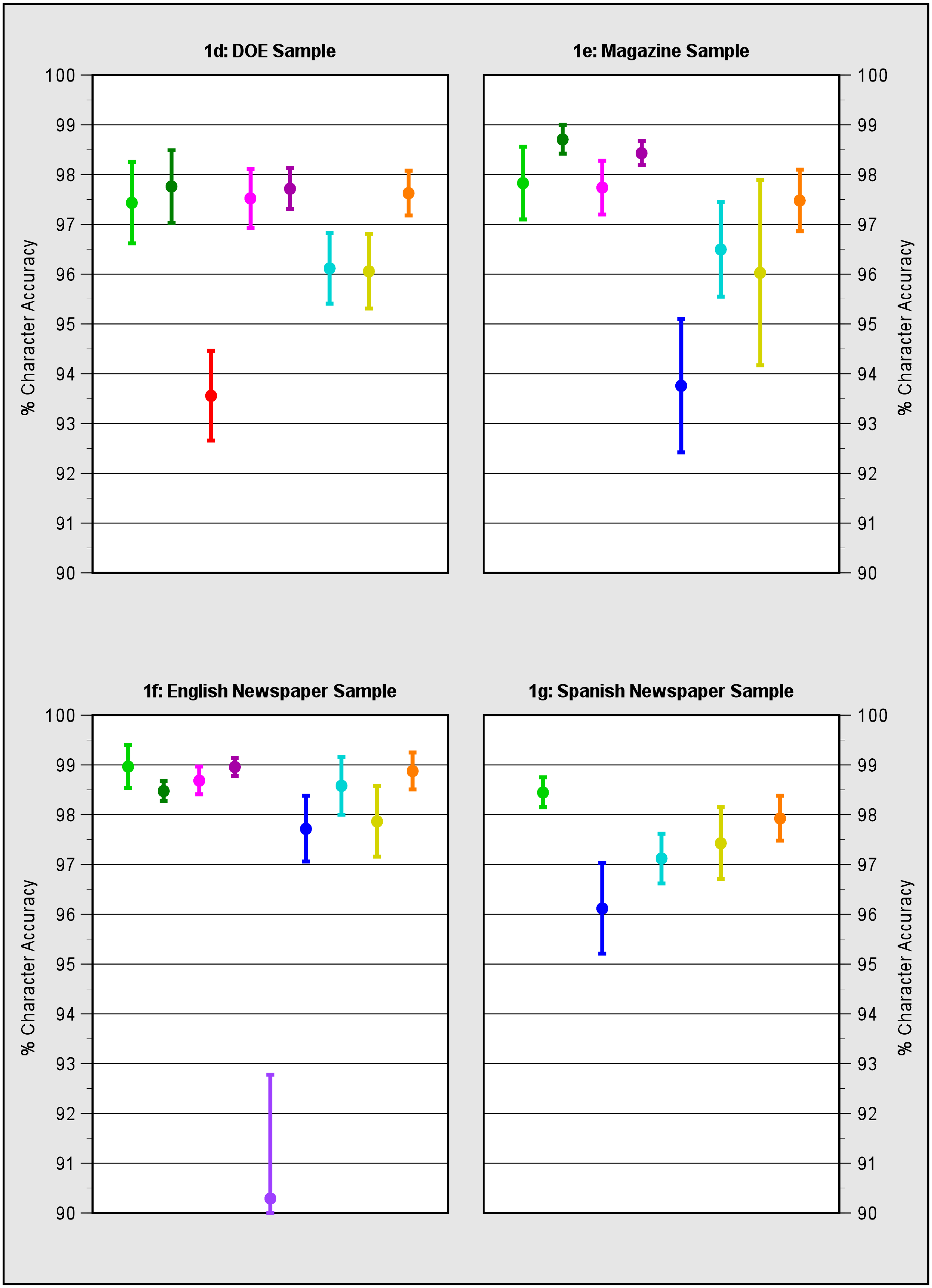
BARCELONA. (Agencias.) -
dirigentes de CiU y del PP
cambiaron ayer duras críticas y

Y TIEMPO LIBRE

El histórico via

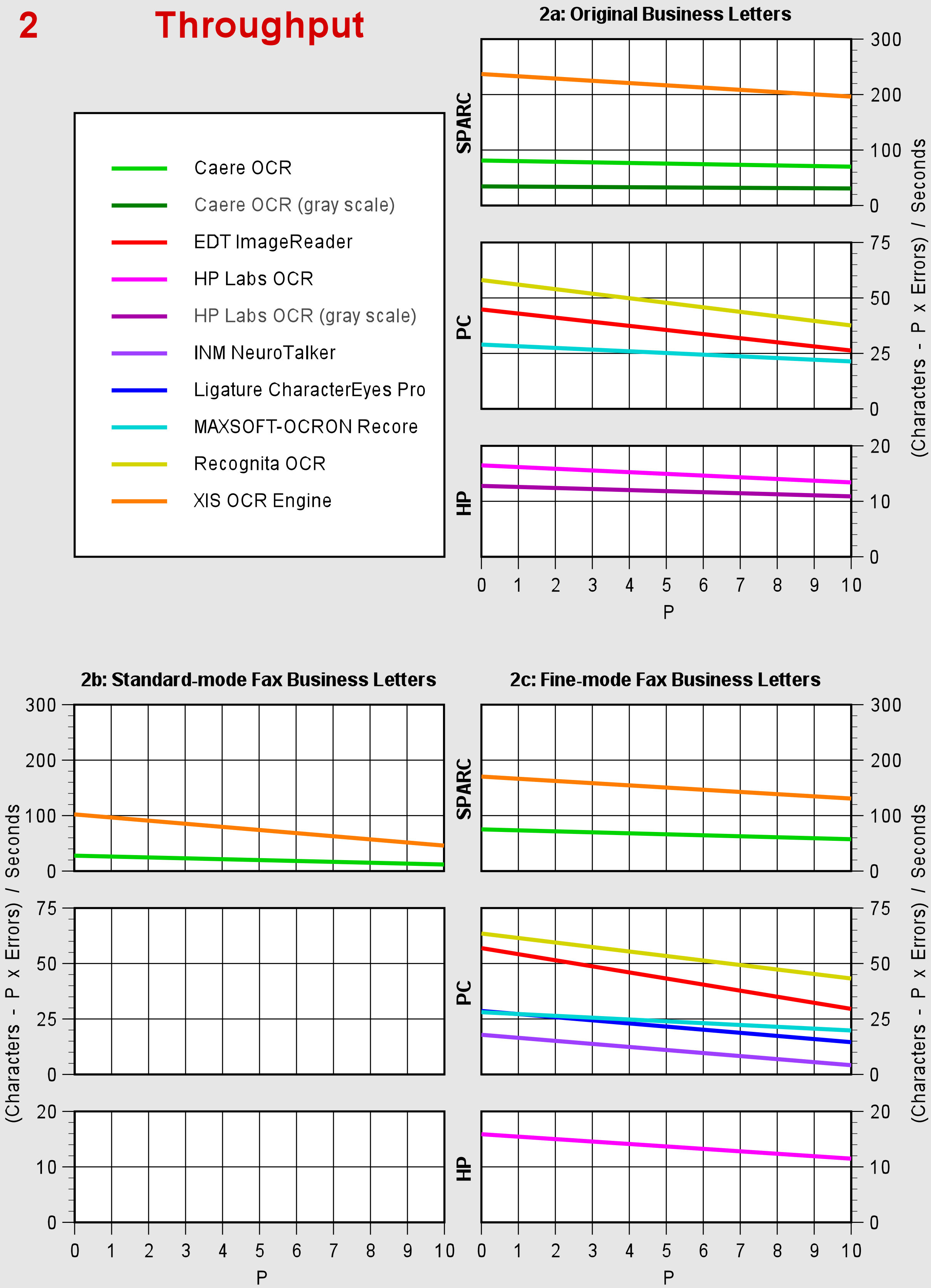
El famoso paseo a pie entre
Martí d'Empúries que cantó co
sep Pla es lo que debía ser, un p

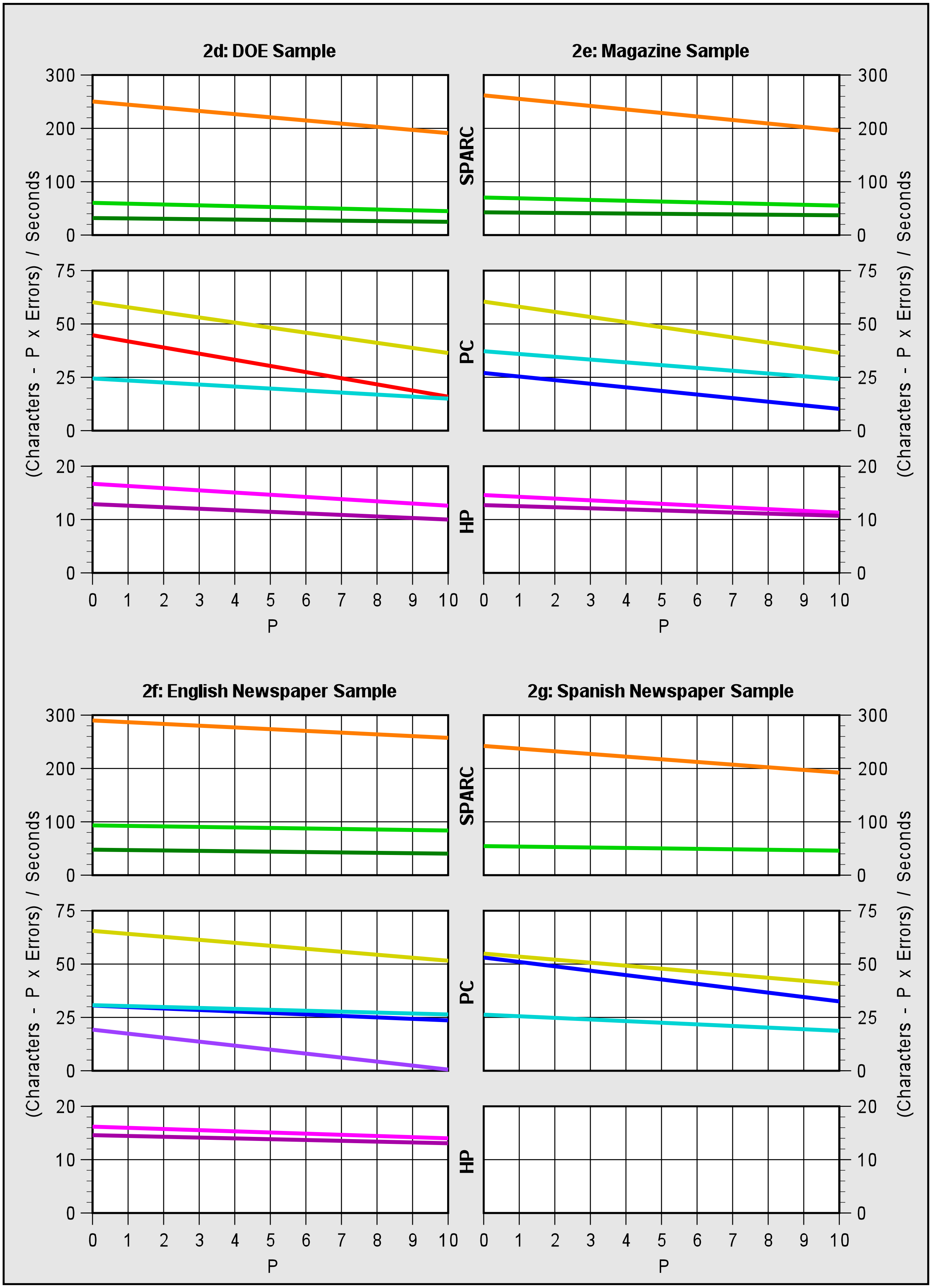

1 Character Accuracy
1a: Original Business Letters
1b: Standard-mode Fax Business Letters
1c: Fine-mode Fax Business Letters
% Character Accuracy
100
99
98
97
96
95
94
93
92
91
90
Caere OCR
Caere OCR (gray scale)
EDT ImageReader
HP Labs OCR
HP Labs OCR (gray scale)
INM NeuroTalker
Ligature CharacterEyes Pro
MAXSOFT-OCRON Recore
Recognita OCR
XIS OCR Engine



1d: DOE Sample
1e: Magazine Sample
1f: English Newspaper Sample
1g: Spanish Newspaper Sample
% Character Accuracy
100
99
98
97
96
95
94
93
92
91
90

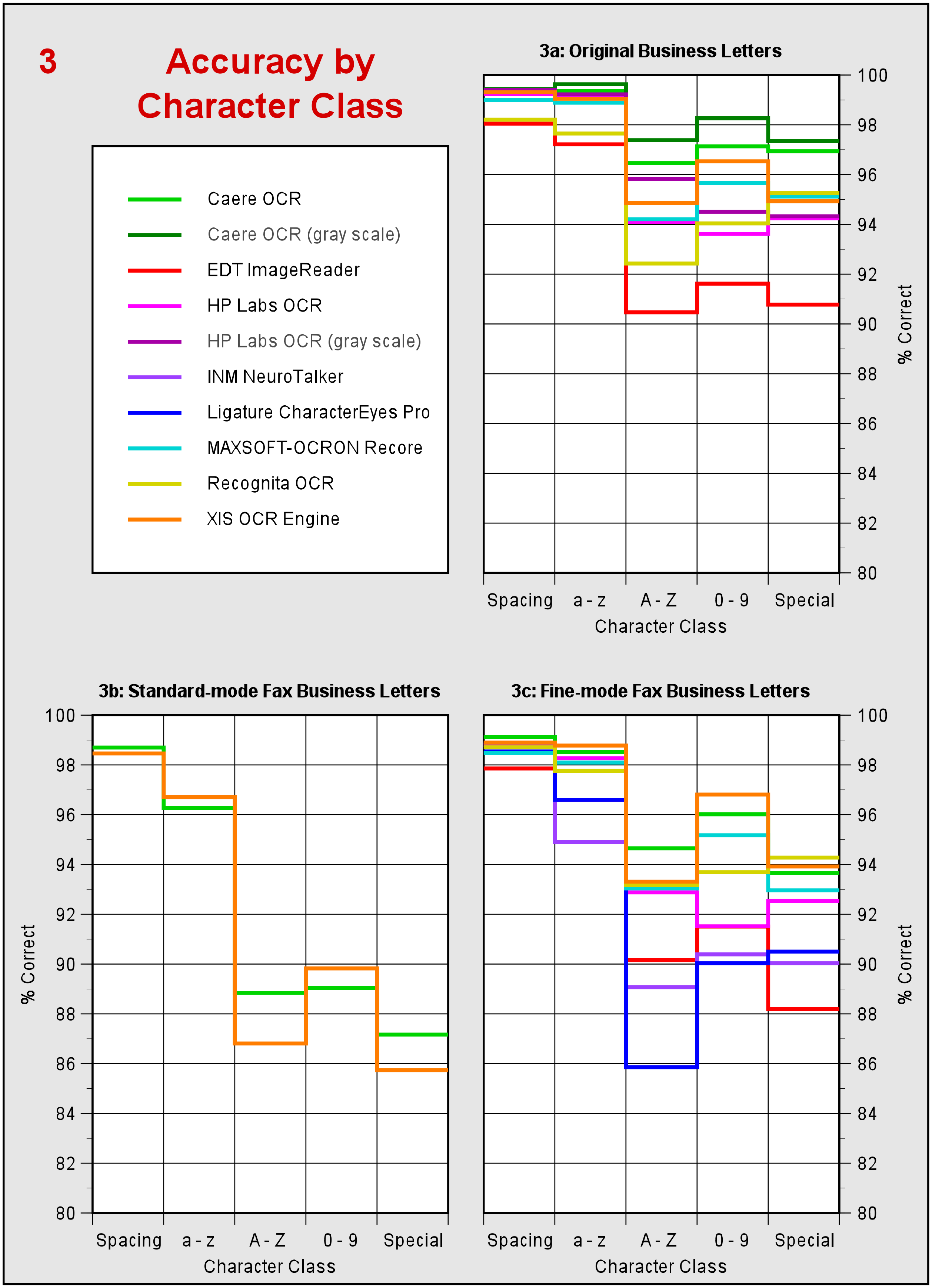
2 Throughput

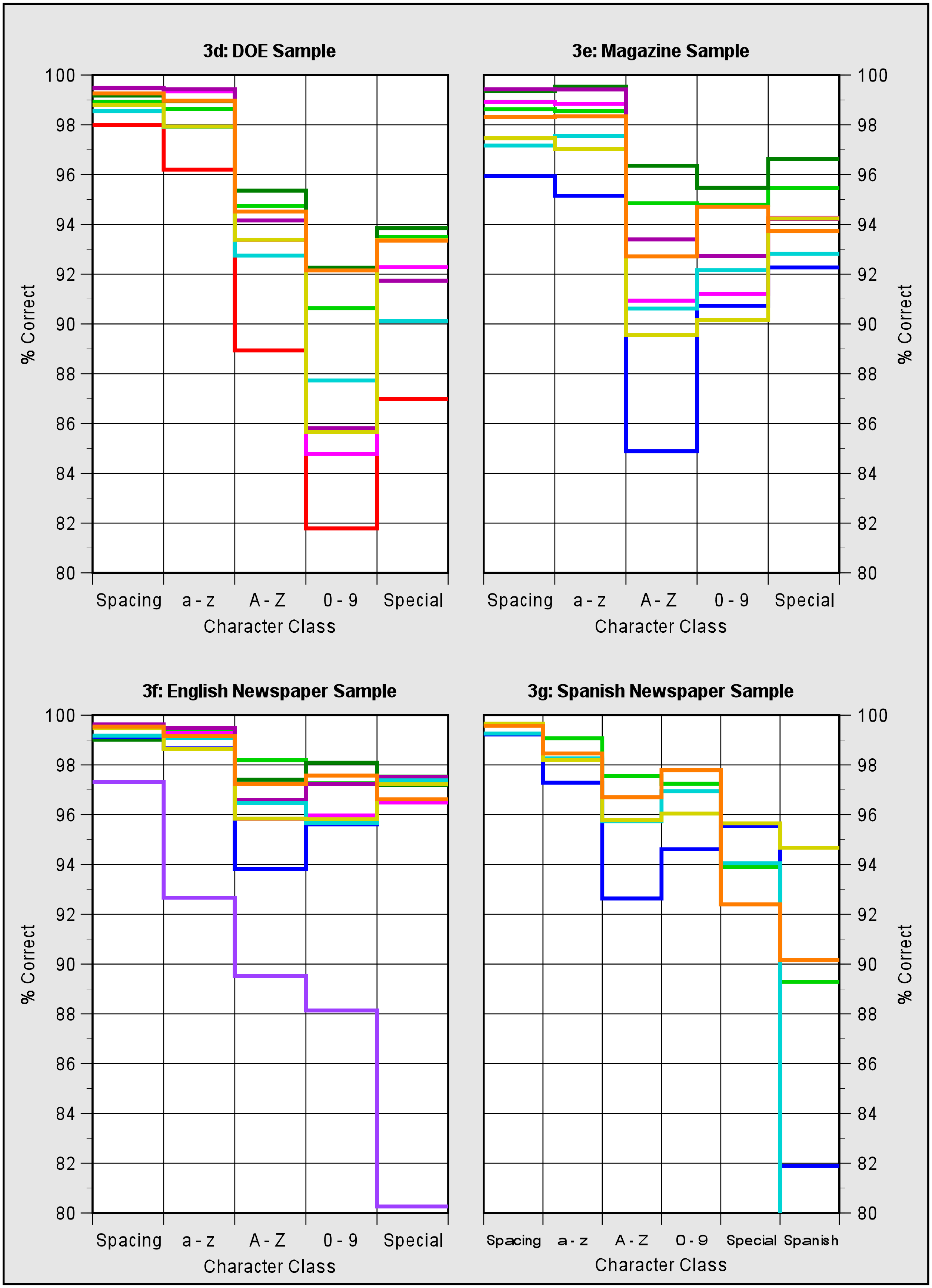




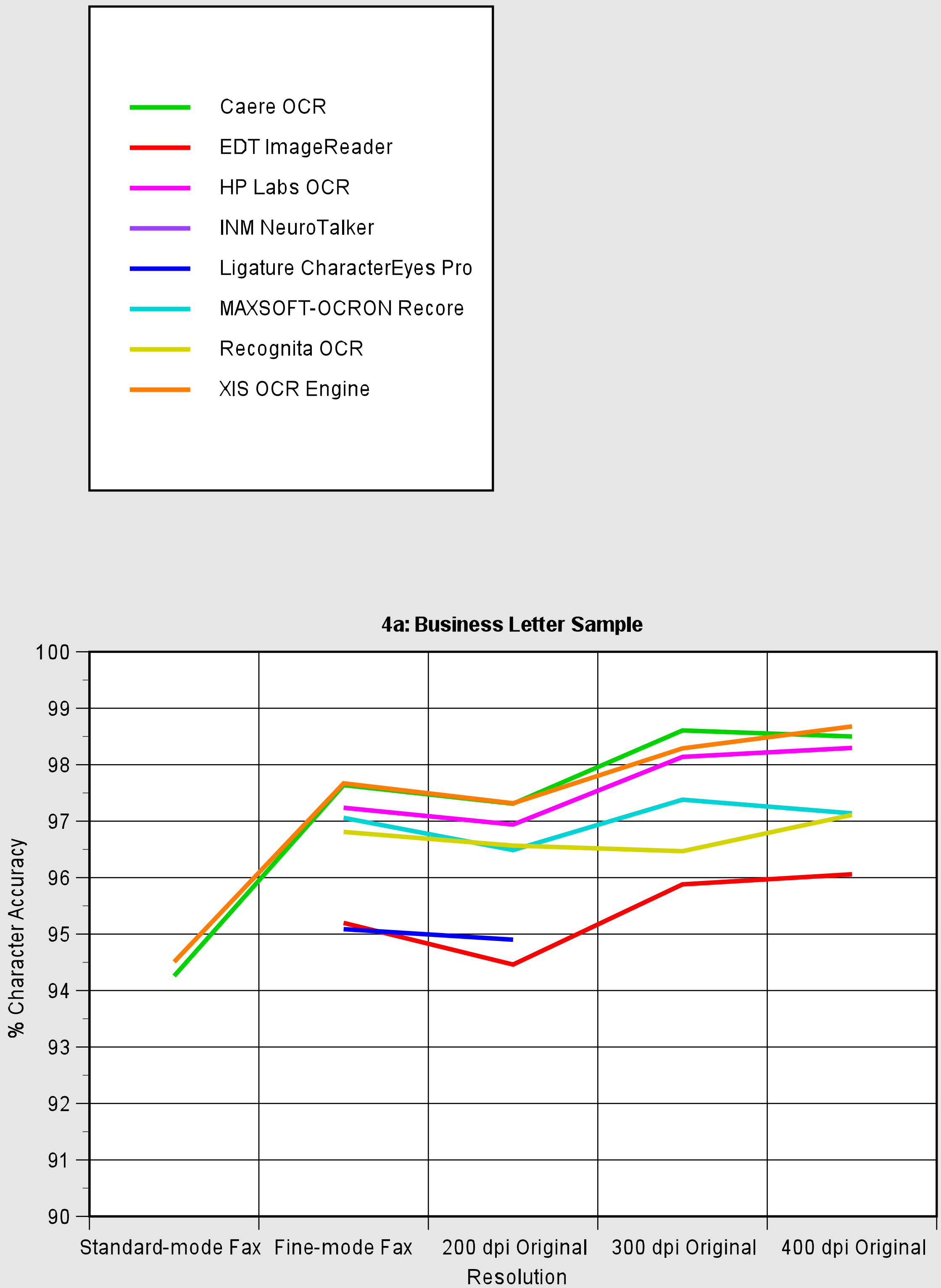
2d: DOE Sample
2e: Magazine Sample
2f: English Newspaper Sample
2g: Spanish Newspaper Sample
SPARC
PC
HP
(Characters - P x Errors) / Seconds
P
0
1
2
3
4
5
6
7
8
9
10
300
200
100
75
50
25
20

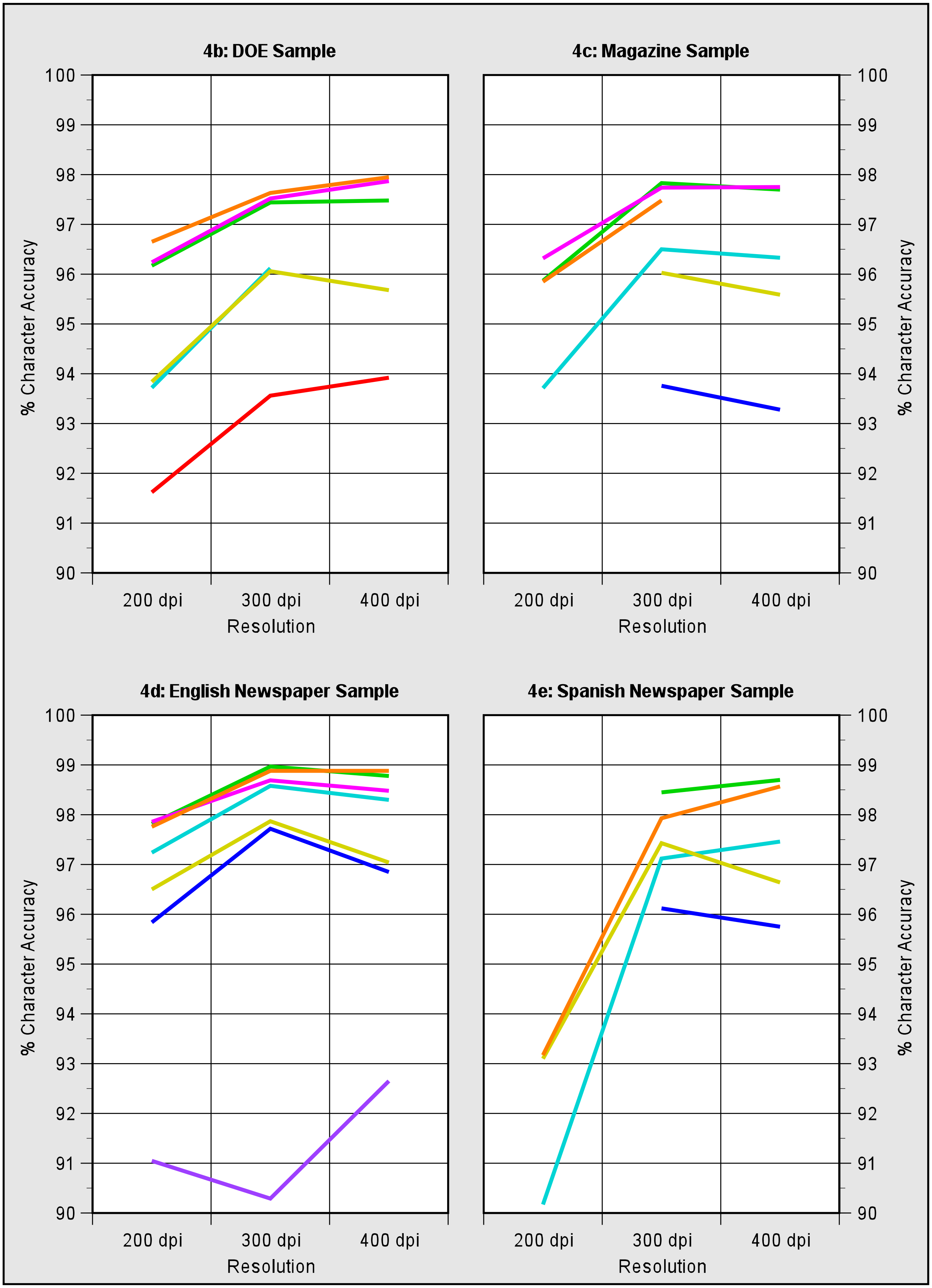
3 Accuracy by Character Class



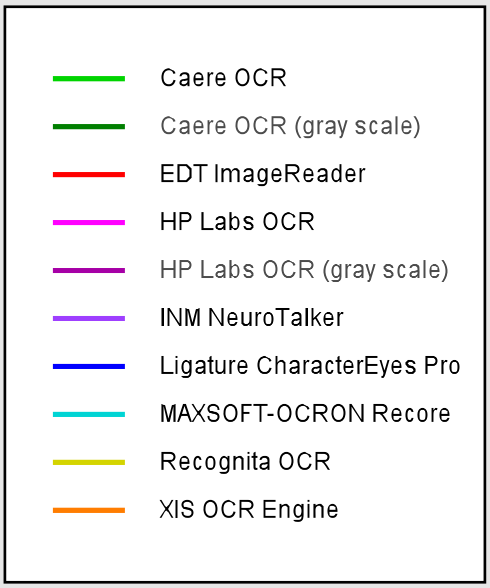

3d: DOE Sample
3e: Magazine Sample
3f: English Newspaper Sample
3g: Spanish Newspaper Sample
% Correct
100
98
96
94
92
90
88
86
84
82
80
Spacing
a - z
A - Z
0 - 9
Special
Spanish
Character Class

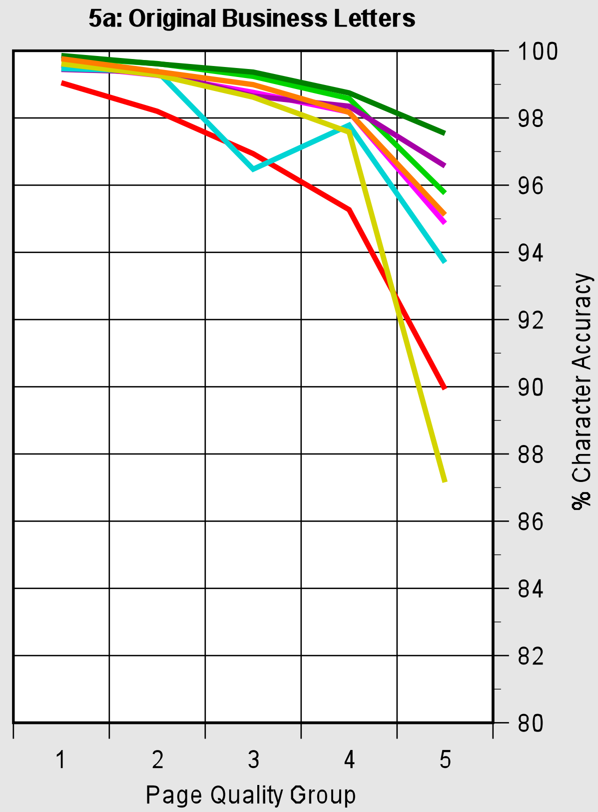
4 Effect of Resolution

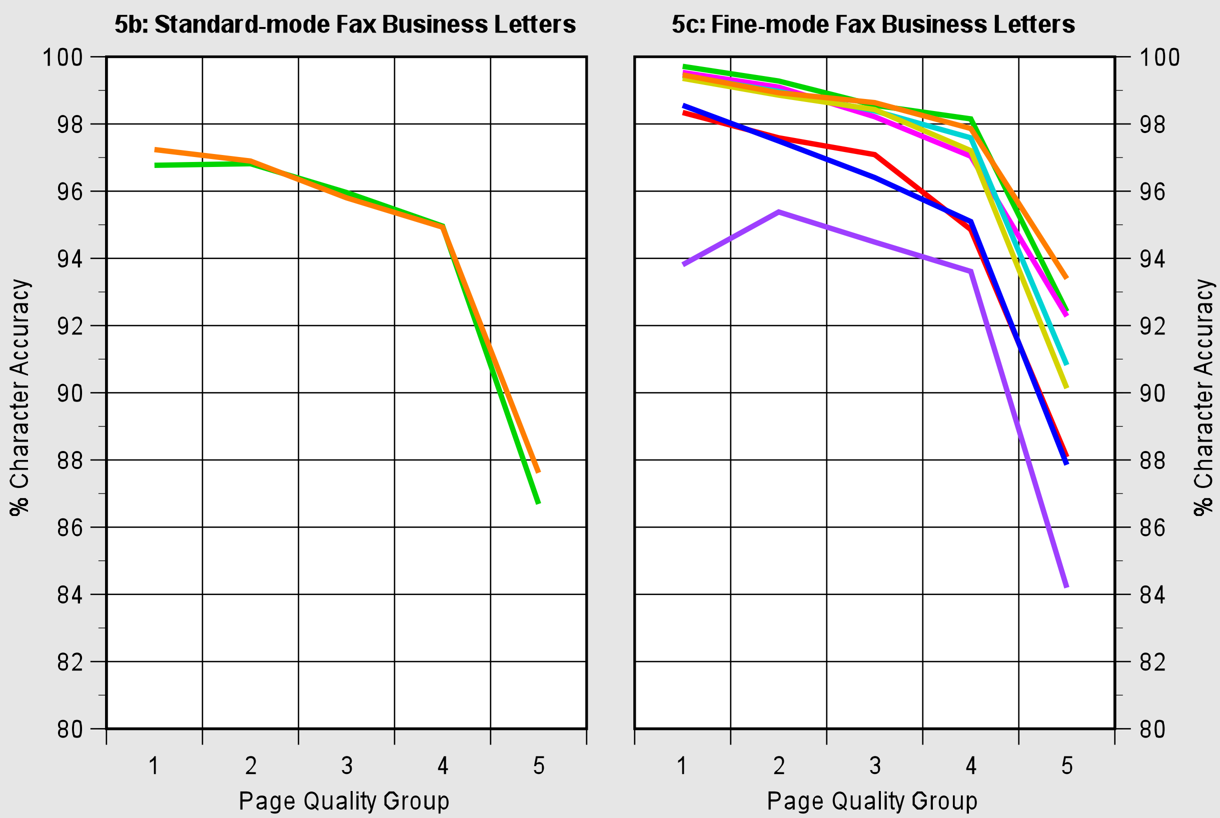


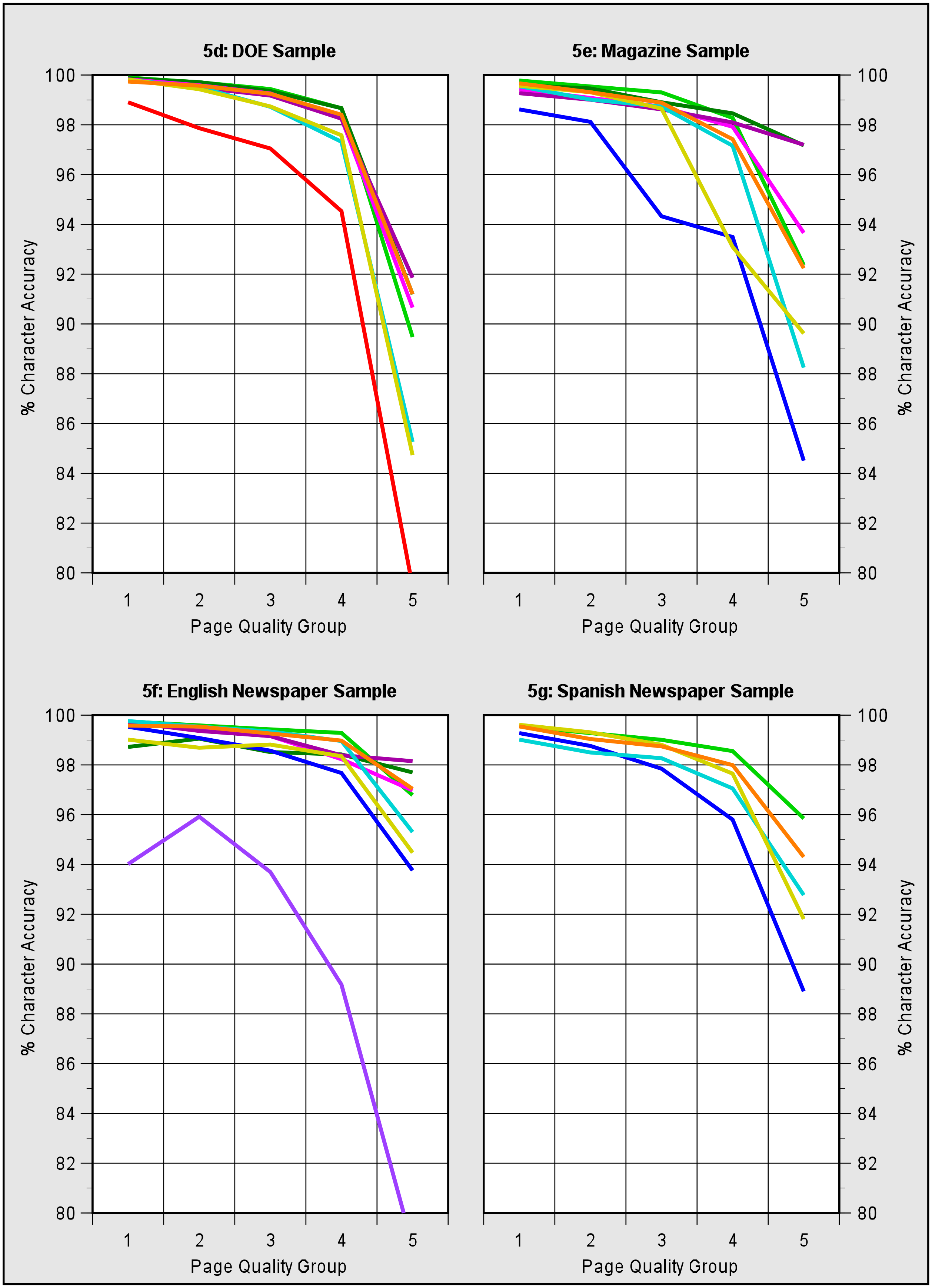

4b: DOE Sample
4c: Magazine Sample
4d: English Newspaper Sample
4e: Spanish Newspaper Sample
% Character Accuracy
100
99
98
97
96
95
94
93
92
91
90
200 dpi
300 dpi
400 dpi
Resolution

5 Character Accuracy vs. Page Quality

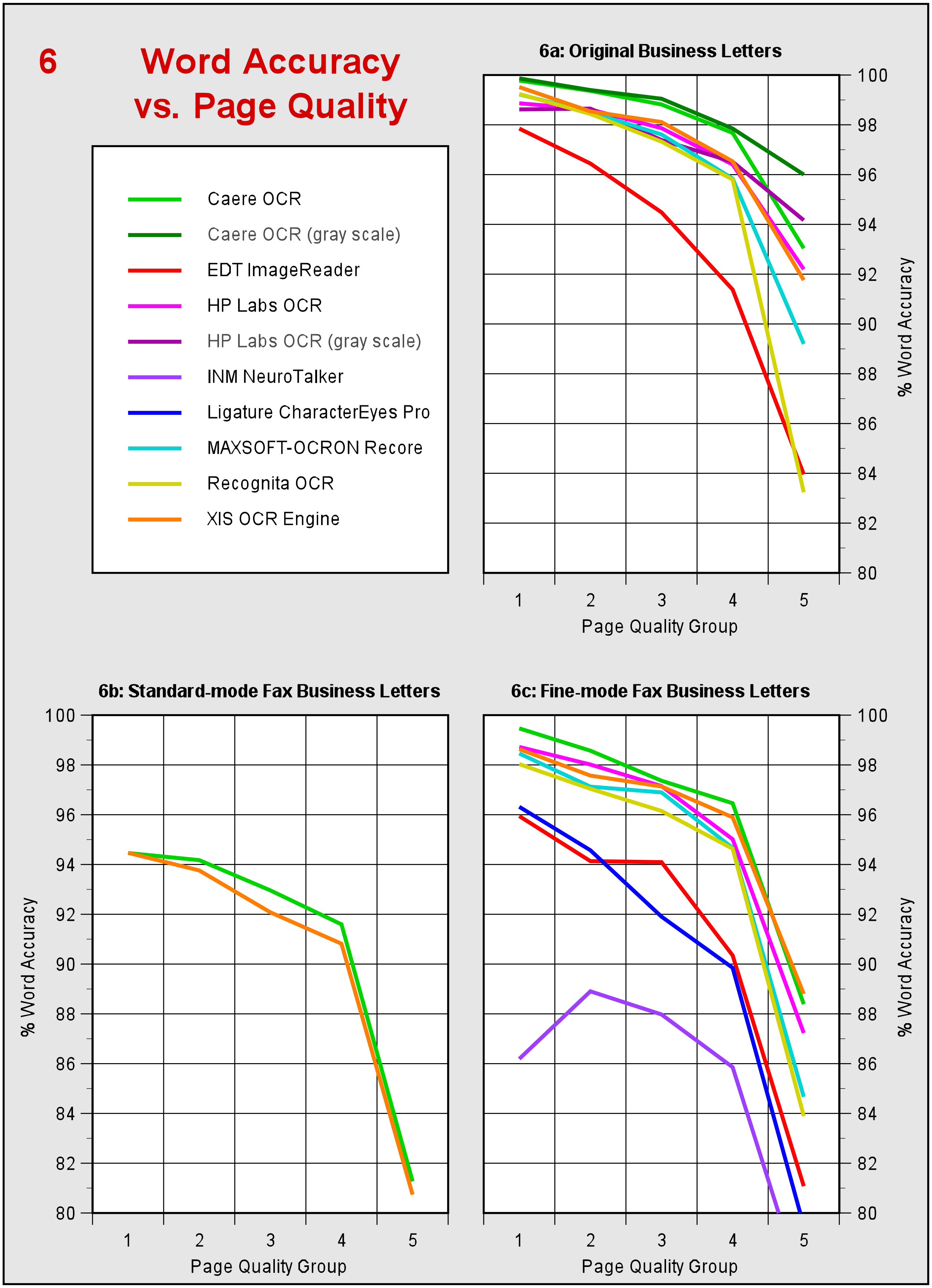


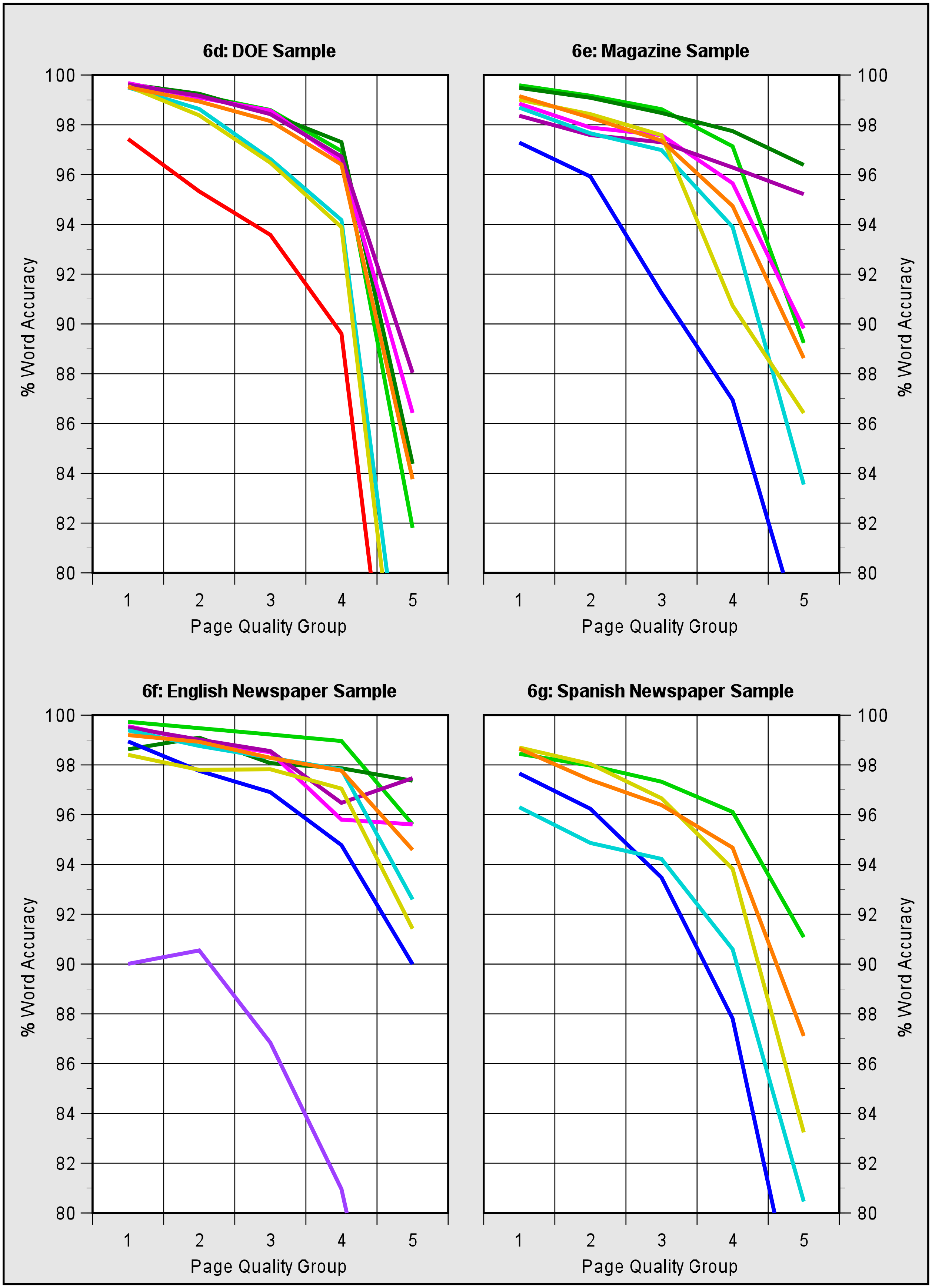




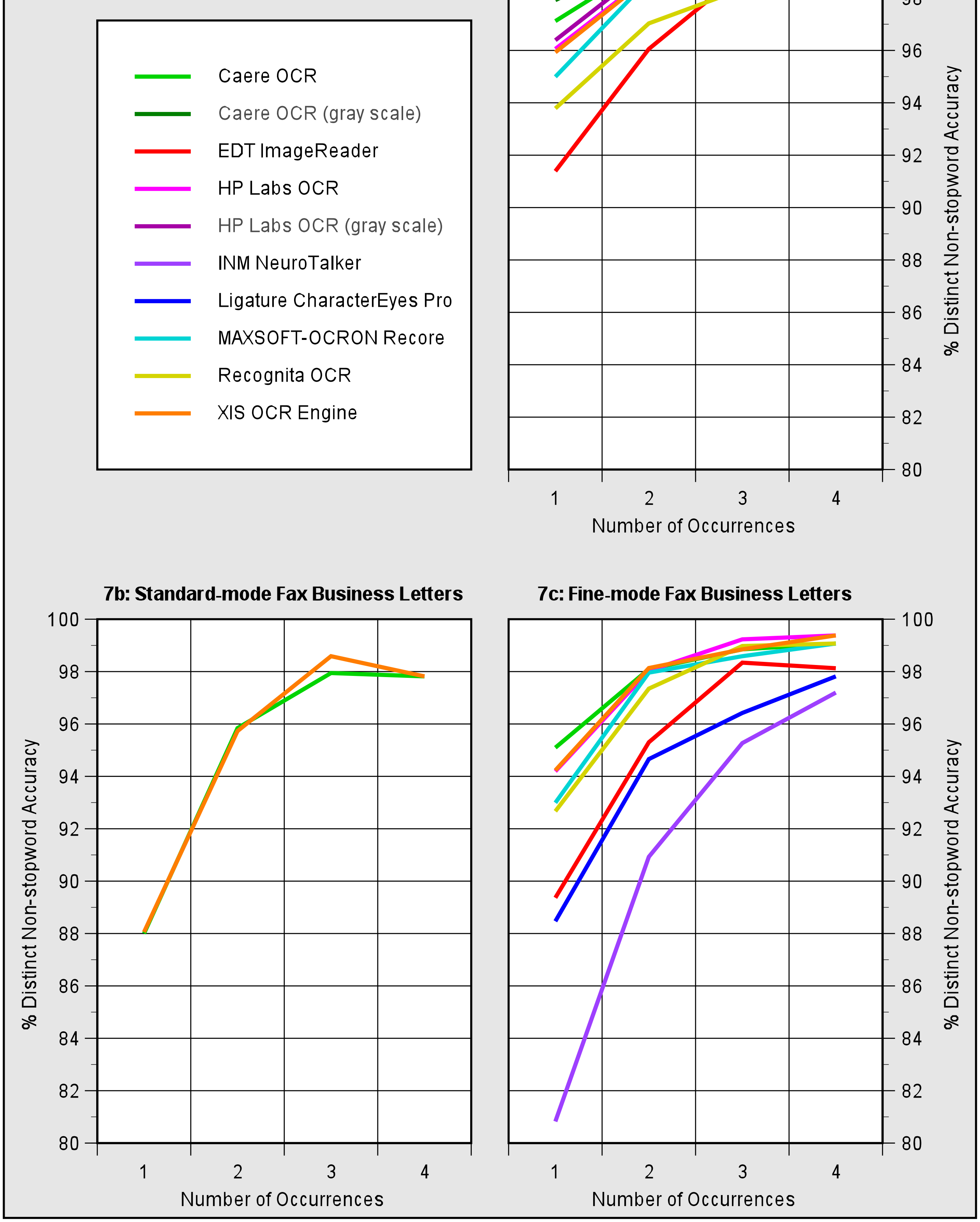

5d: DOE Sample
5e: Magazine Sample
5f: English Newspaper Sample
5g: Spanish Newspaper Sample
% Character Accuracy
Page Quality Group
100
98
96
94
92
90
88
86
84
82
80
1
2
3
4
5

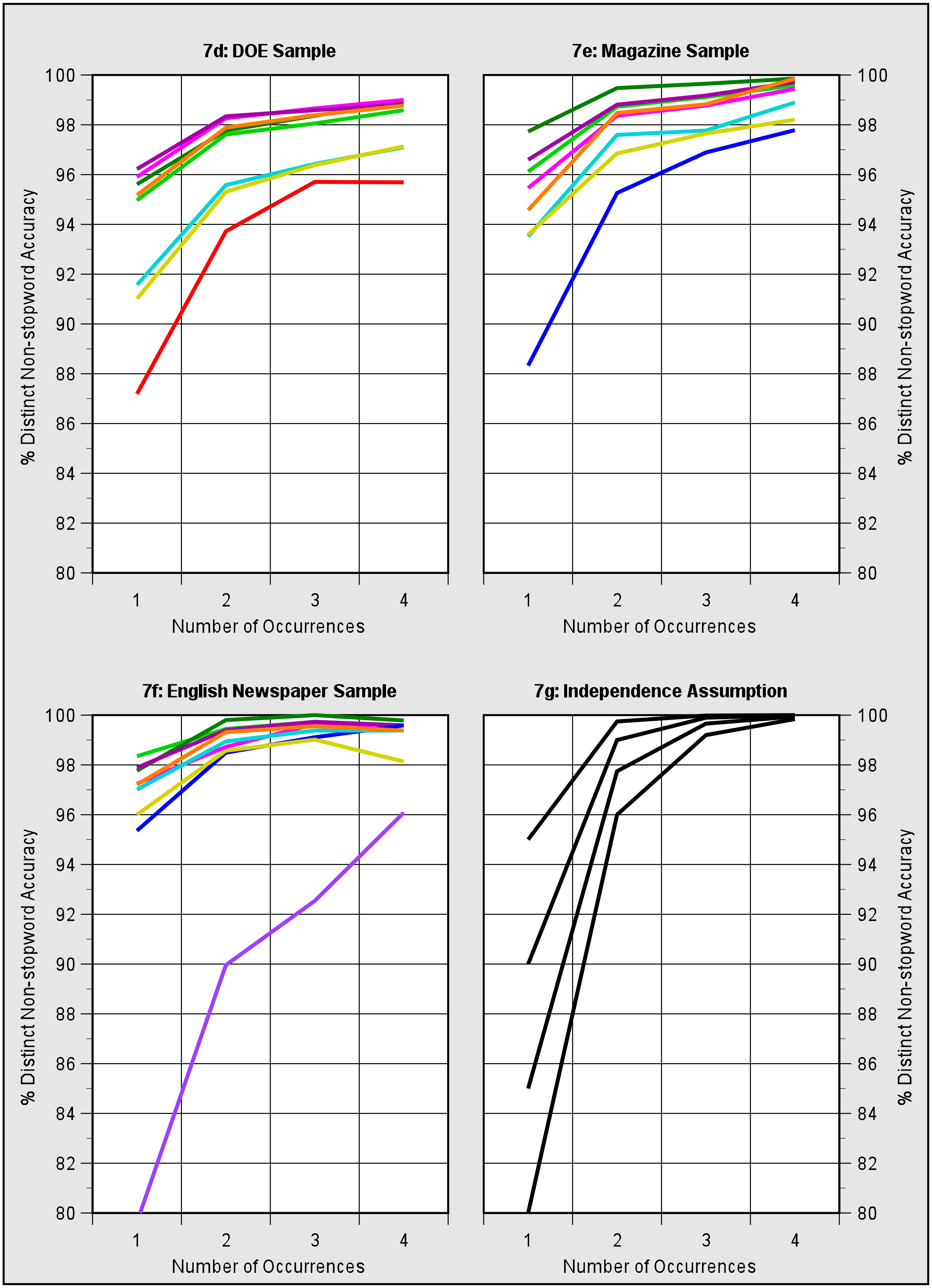
6 Word Accuracy vs. Page Quality



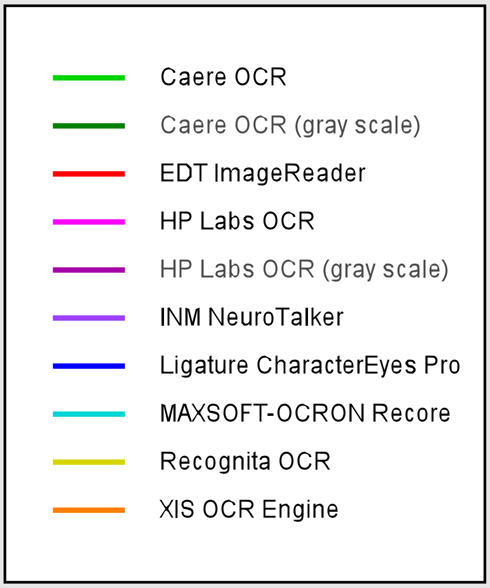

6d: DOE Sample
6e: Magazine Sample
6f: English Newspaper Sample
6g: Spanish Newspaper Sample
% Word Accuracy
Page Quality Group
100
98
96
94
92
90
88
86
84
82
80
1
2
3
4
5

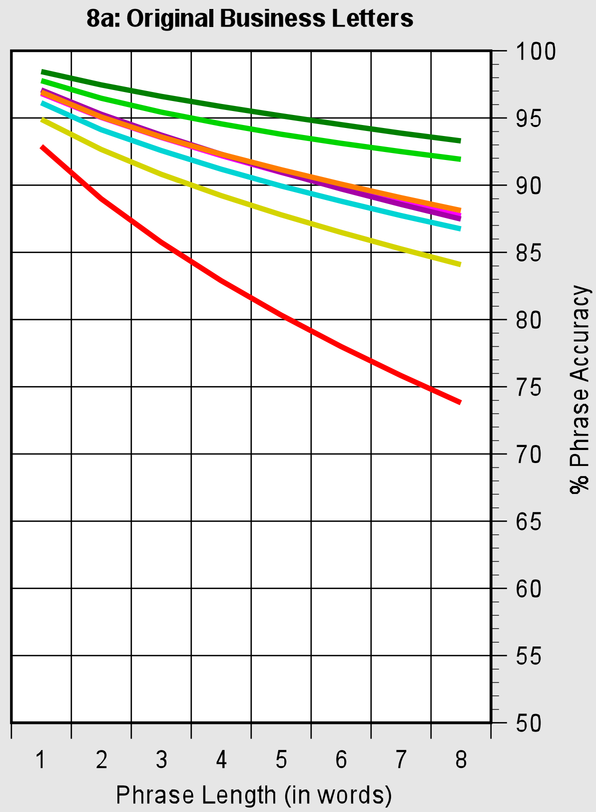
7 Distinct Non-stopword Accuracy

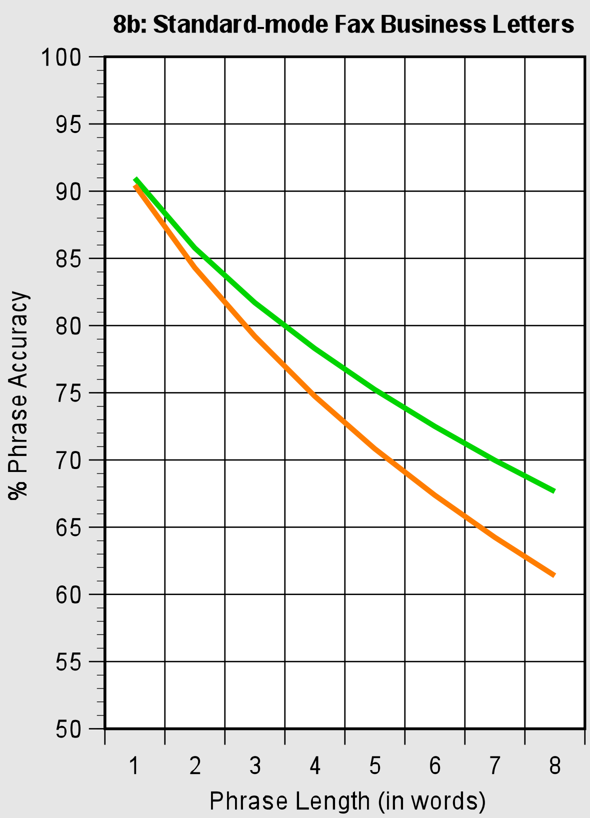


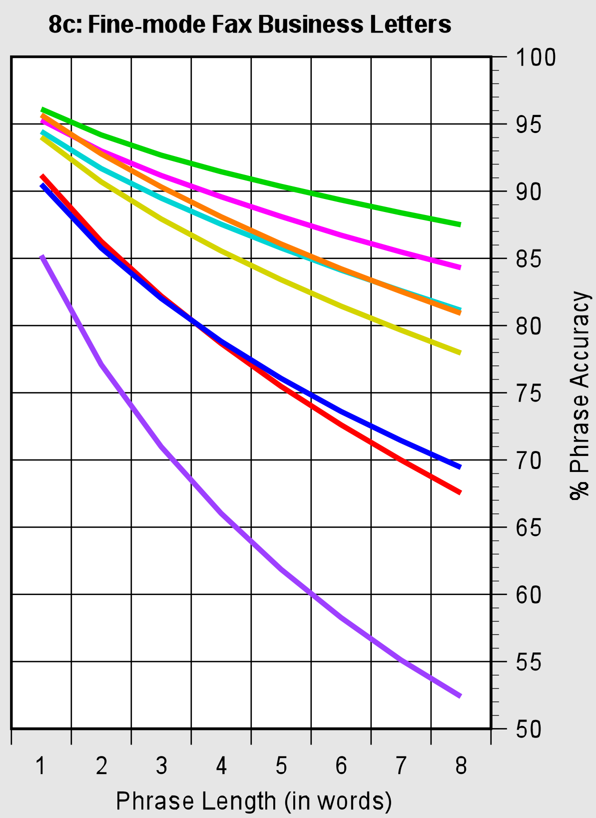

7d: DOE Sample
7e: Magazine Sample
7f: English Newspaper Sample
7g: Independence Assumption
% Distinct Non-stopword Accuracy
Number of Occurrences
100
98
96
94
92
90
88
86
84
82
80
1
2
3
4

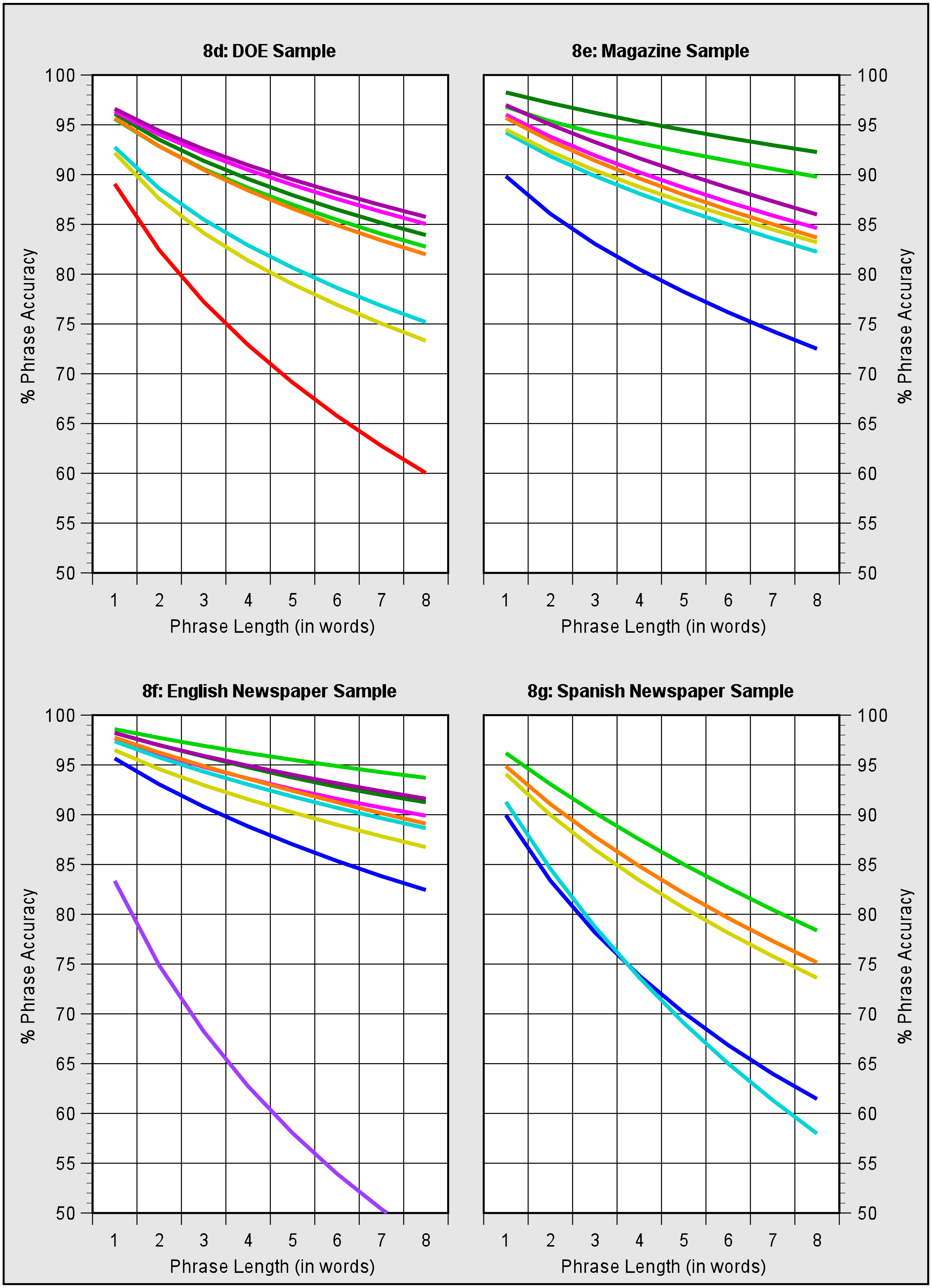
8 Phrase Accuracy



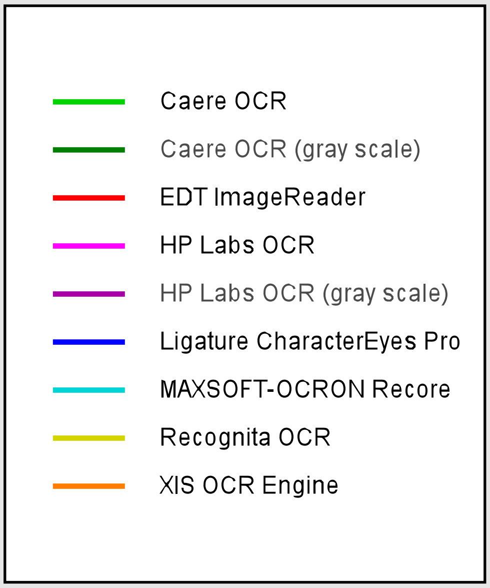


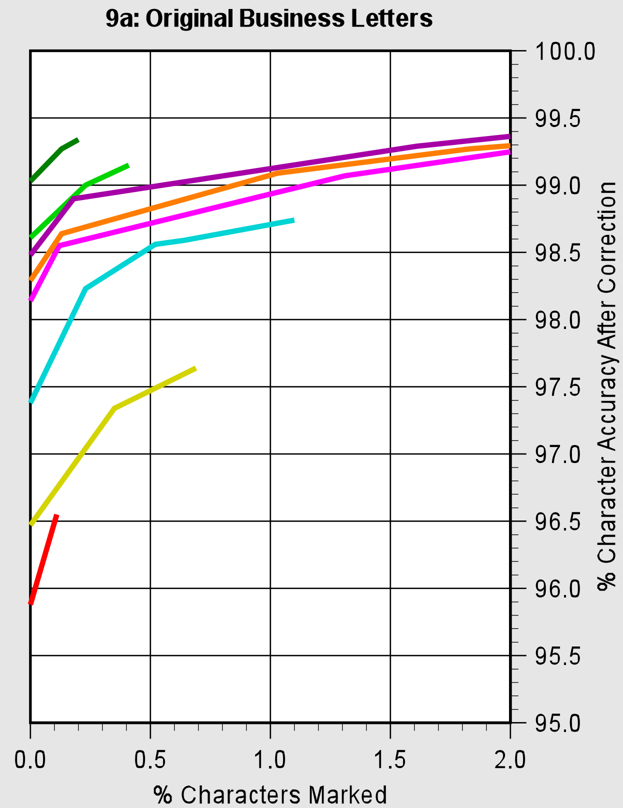


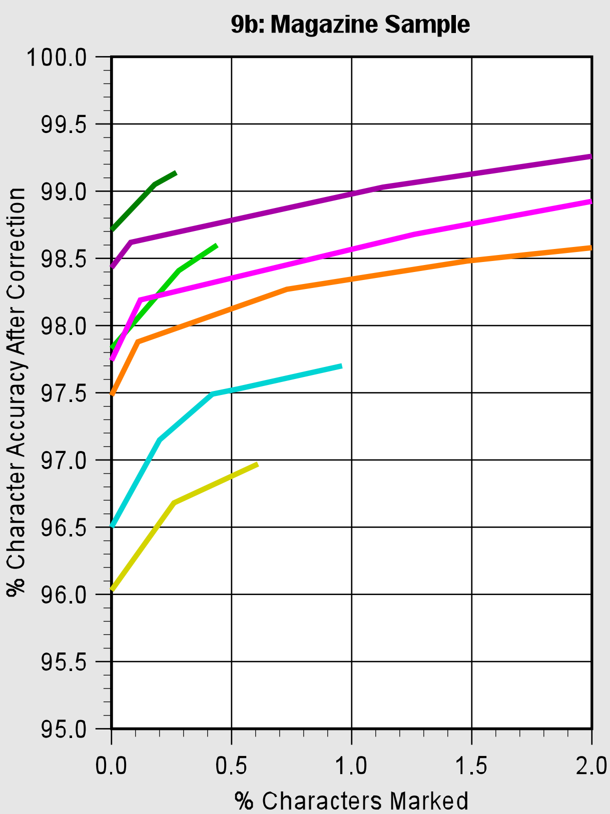


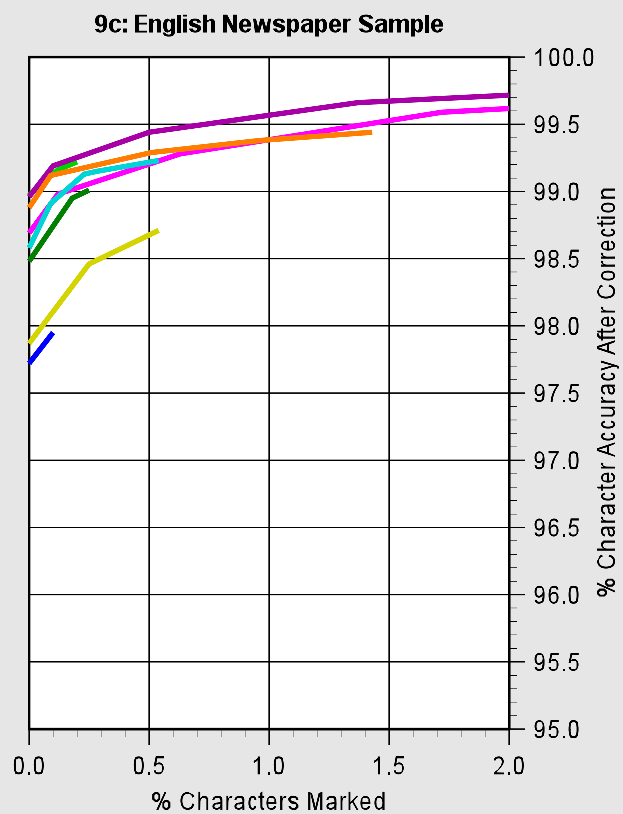

8d: DOE Sample
8e: Magazine Sample
8f: English Newspaper Sample
8g: Spanish Newspaper Sample
% Phrase Accuracy
Phrase Length (in words)
100
95
90
85
80
75
70
65
60
55
50
1
2
3
4
5
6
7
8

9 Marked Character Efficiency









10 Automatic Zoning