The Fifth Annual Test of OCR Accuracy

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

Information Science Research Institute

TR-96-01

April 1996

A Technical Report

published by the
Information Science Research Institute
University of Nevada, Las Vegas
4505 Maryland Parkway • Box 454021
Las Vegas, NV 89154-4021

Telephone: +1 (702) 895–3338
Fax: +1 (702) 895–1183
Email: isri-info@isri.unlv.edu
URL: http://www.isri.unlv.edu/

Copyright © 1996 by the Information Science Research Institute.

All rights reserved. This document may not be reproduced in whole or in part by any means without permission. For information write: Information Science Research Institute, University of Nevada, Las Vegas, 4505 Maryland Parkway, Box 454021, Las Vegas, Nevada 89154-4021. USA. Email: *isri-info@isri.unlv.edu*

The Fifth Annual Test of OCR Accuracy

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

1 Introduction

The Information Science Research Institute (ISRI) at the University of Nevada, Las Vegas, conducts an annual test of page-reading systems. A page-reading system, or "page reader," accepts as input a bitmapped image of any document page. This image is a digitized representation of the page and is produced by a scanner. Through a process known as optical character recognition (OCR), the page reader analyzes the image and attempts to locate and identify the machine-printed characters on the page. The output is a text file containing a coded representation of the characters which may be edited or searched.

This process is far from foolproof and the output may contain errors. Characters that have not been located on the page will be missing from the output; speckles in the image may be mistaken for characters, causing extraneous symbols to be produced; or most commonly, characters may simply be misidentified.

In comparison to other types of software, such as spreadsheets or word processor programs, page-reading systems are very unpredictable. The universe of pages containing machine-printed text is enormous, yet all of these pages are potential inputs to such systems. How well they perform when confronted with a broad range of typefaces, type sizes, and print qualities can be estimated only through rigorous testing.

Since its formation in 1990, ISRI has devoted much effort to developing a comprehensive methodology for the evaluation of page readers. With considerable input from the research, vendor and user communities, a methodology has evolved for conducting large-scale, automated tests of this technology. Using this methodology, ISRI performs an annual test of the "latest and greatest" page-reading systems provided by participating organizations. The purpose of the test is to make visible the capabilities of these systems and to identify problems at the state-of-the-art.

For the fifth annual test, more than 2,000 pages were selected from seven different types of documents. These pages contain nearly 5 million characters in total. Five leading page readers processed images of these pages. Their performance was measured and is presented in this report.

2 Test Description

Any organization may participate in the annual test by submitting a page-reading system by the established deadline, which was December 15, 1995 for the fifth annual test. The system must be able to run unattended on a PC or Sun SPARCstation. Participation in the test is voluntary and free, but only one entry is allowed per organization, and participants must sign an agreement regarding the use of the test results in advertising. Table 1 lists the participants in this year's test.

Organization	Version Name	Version No.	Platform	Version Type
International Neural Machines Inc. Waterloo, Ontario	INM NeuroTalker	3.3	PC DOS	pre-release
Maxsoft-Ocron, Inc. Fremont, California	Maxsoft-Ocron Recore	4.1	PC Windows	pre-release
Nankai University Tianjin, China	Nankai Reader	4.0	PC Windows	research prototype
Recognita Corp. Budapest, Hungary	Recognita OCR	3.02	PC Windows	commercial release
Xerox Corp. Peabody, Massachusetts	Xerox OCR Engine	11.0	Sun SPARCstation	pre-release

Table 1: Test Participants

The pages chosen for the test belong to seven distinct "samples":

1. The *Corporate Annual Report Sample* consists of pages from the annual financial reports to stockholders from 75 "Fortune 500" corporations.
2. The *DOE Sample* contains pages from a large group of scientific and technical documents collected by the U.S. Department of Energy and its contractors for the Yucca Mountain Project.
3. The *Magazine Sample* is composed of articles from 75 popular magazines.
4. The *Legal Document Sample* contains pages from a variety of legal documents that were obtained from a local law firm and a local bankruptcy court.
5. The *English Business Letter Sample* is a collection of business letters gathered by ISRI.
6. The *Spanish Newspaper Sample* contains Spanish-language articles clipped from 12 popular newspapers from Argentina, Mexico, and Spain.
7. The *German Business Letter Sample* is a group of German-language business letters collected by our academic affiliate, DFKI in Kaiserslautern, Germany.

With the exception of the two business letter samples, pages were selected from documents at random. All samples contain English text except the *Spanish Newspaper Sample* and the *German Business Letter Sample*. Three of the samples were used in last year's test: the *DOE Sample*, the *English Business Letter Sample*, and the *Spanish Newspaper Sample*.

Each page was digitized four times using a Fujitsu M3096G scanner to produce three binary images and one 8-bit gray scale image. The binary images were created using a global threshold at resolutions of 200, 300, and 400 dots per inch (dpi). The gray scale image was scanned at 300 dpi. In addition, for the *Legal Document Sample* and the two business letter samples, each page was twice transmitted locally from a Xerox 7024 fax machine to a fax modem to obtain both a standard-mode fax image (204 x 98 dpi) and a fine-mode fax image (204 x 196 dpi). Although four or six images were created for each page, all results given in this report are for the 300 dpi binary and gray scale images unless otherwise indicated.

An important step in preparing a page for testing is to define its "zones." We manually delineated the text regions of each page and determined their reading order. With the exception of the automatic zoning test, the page-reading systems processed only the zoned portions of each image. Furthermore, each system processed the same zoned portions of the same images; that is, the pages were not re-zoned or re-scanned for each system. Some text regions were not zoned and were thereby excluded from the test. These include advertisements, logos, mathematical equations, and graph and map labels.

To minimize errors in the "ground truth," the text for each zone was entered by four typists working independently. Differences in their entries were reconciled with the help of a string-matching program. Table 2 gives the number of pages, zones, words, and characters in each sample.

	Pages	Zones	Words	Characters
Corporate Annual Report Sample	300	1,704	122,007	892,266
DOE Sample	785	2,280	213,552	1,463,512
Magazine Sample	300	2,332	206,648	1,244,171
Legal Document Sample	300	847	58,703	372,098
English Business Letter Sample	200	1,419	51,460	319,756
Spanish Newspaper Sample	144	558	57,670	348,091
German Business Letter Sample	200	1,814	36,884	298,590
Total	2,229	10,954	746,924	4,938,484

Table 2: Test Data

The PCs used in the test were identically-configured 486DX/33 machines with 8 megabytes of memory running MS-DOS 5.0 and MS Windows 3.1. The Xerox OCR Engine was operated under SunOS 4.1.3 on a single-processor Sun SPARCstation 10 with 64 megabytes of memory.

The test was conducted using Version 5.1 of the OCR Experimental Environment. This is a suite of software tools developed at ISRI that includes programs for operating page readers, for matching their outputs with correct text, and for computing all of the performance measures described in this report.

3 Character Accuracy

The text generated by a page-reading system is matched with the correct text to determine the minimum number of edit operations (character insertions, deletions, and substitutions) needed to correct the generated text. This quantity is termed the number of *errors*. If there are *n* characters in the correct text, then the character accuracy is given by $\frac{n - \#\text{errors}}{n}$.

Tables 3a to 3g show the number of errors for each system and the corresponding character accuracy. Missing entries in these tables are due to a variety of reasons:

1. The INM NeuroTalker and Maxsoft-Ocron Recore do not accept gray scale input.
2. The gray scale test for the *Magazine Sample* was not run due to time limitations.
3. Only Recognita OCR and the Xerox OCR Engine support Spanish or German.
4. Results are omitted when the character accuracy is less than 90%.
5. Results are omitted when the number of characters on failed pages exceeds one percent of the sample.

A failure is detected when a page reader crashes, hangs, or returns an error status when processing a page. If the Failures column indicates "5 / 1.78" for example, then failures were detected on five pages containing a total of 1.78% of the characters in the sample. Since this exceeds one percent, no further results are given. But if there were only two failed pages containing 0.40% of the sample (i.e., "2 / 0.40"), then the character accuracy is reported, but it is reduced by 0.40% because one error is charged for each character on a failed page.

The ISRI Voting Machine was applied to the outputs (when available) of the following page-reading systems: the Nankai Reader, Recognita OCR, and the Xerox OCR Engine. A single output was produced by finding disagreements among these three systems and resolving them by a majority vote. Shaded table entries give the accuracy of this voting output. Additional shaded entries in Tables 3b, 3e and 3f show the highest accuracy achieved among participants in last year's test.

Graphs 1a to 1g display approximate 95% confidence intervals for character accuracy. A narrow interval indicates consistent performance within a sample whereas a wide interval reflects considerable variability. Two intervals with no overlap imply a statistically significant difference in character accuracy.

	Binary Input			Gray Scale Input		
	Errors	Accuracy	Failures	Errors	Accuracy	Failures
INM NeuroTalker	—	< 90.00	none	—	—	—
Maxsoft-Ocron Recore	42,766	95.21	2 / 0.40	—	—	—
Nankai Reader	54,841	93.85	none	—	—	5 / 1.78
Recognita OCR	33,884	96.20	none	20,652	97.69	none
Xerox OCR Engine	26,579	97.02	none	13,396	98.50	none
ISRI Voting Machine	23,621	97.35	none	—	—	—

Table 3a: Character Accuracy, *Corporate Annual Report Sample*

	Binary Input			Gray Scale Input		
	Errors	Accuracy	Failures	Errors	Accuracy	Failures
INM NeuroTalker	—	< 90.00	none	—	—	—
Maxsoft-Ocron Recore	51,258	96.50	none	—	—	—
Nankai Reader	37,147	97.46	none	48,260	96.70	6 / 0.87
Recognita OCR	45,460	96.89	none	44,635	96.95	none
Xerox OCR Engine	30,823	97.89	none	27,775	98.10	none
ISRI Voting Machine	23,274	98.41	none	24,335	98.34	none
Best Last Year (Xerox)	34,644	97.63	none	—	—	—
Best Last Year (Caere)	—	—	—	32,791	97.76	1 / 0.33

Table 3b: Character Accuracy, *DOE Sample*

	Binary Input			Gray Scale Input		
	Errors	Accuracy	Failures	Errors	Accuracy	Failures
INM NeuroTalker	—	< 90.00	none	—	—	—
Maxsoft-Ocron Recore	84,978	93.17	1 / 0.17	—	—	—
Nankai Reader	69,414	94.42	none	—	—	—
Recognita OCR	37,564	96.98	none	—	—	—
Xerox OCR Engine	44,949	96.39	none	—	—	—
ISRI Voting Machine	42,888	96.55	none	—	—	—

Table 3c: Character Accuracy, *Magazine Sample*

	Binary Input			Gray Scale Input		
	Errors	Accuracy	Failures	Errors	Accuracy	Failures
INM NeuroTalker	8,115	97.82	none	—	—	—
Maxsoft-Ocron Recore	3,082	99.17	none	—	—	—
Nankai Reader	1,891	99.49	none	4,995	98.66	4 / 0.79
Recognita OCR	2,929	99.21	none	3,118	99.16	none
Xerox OCR Engine	4,603	98.76	none	4,575	98.77	none
ISRI Voting Machine	1,267	99.66	none	1,182	99.68	none

Table 3d: Character Accuracy, *Legal Document Sample*

	Binary Input			Gray Scale Input		
	Errors	Accuracy	Failures	Errors	Accuracy	Failures
INM NeuroTalker	31,645	90.10	none	—	—	—
Maxsoft-Ocron Recore	7,796	97.56	none	—	—	—
Nankai Reader	3,746	98.83	none	2,800	99.12	none
Recognita OCR	6,756	97.89	none	5,923	98.15	none
Xerox OCR Engine	4,952	98.45	none	3,270	98.98	none
ISRI Voting Machine	2,715	99.15	none	1,841	99.42	none
Best Last Year (Caere)	4,459	98.61	none	3,102	99.03	none

Table 3e: Character Accuracy, *English Business Letter Sample*

	Binary Input			Gray Scale Input		
	Errors	Accuracy	Failures	Errors	Accuracy	Failures
Recognita OCR	7,704	97.79	none	7,551	97.83	none
Xerox OCR Engine	5,804	98.33	none	3,043	99.13	none
Best Last Year (Caere)	5,394	98.45	none	—	—	—

Table 3f: Character Accuracy, *Spanish Newspaper Sample*

	Binary Input			Gray Scale Input		
	Errors	Accuracy	Failures	Errors	Accuracy	Failures
Recognita OCR	14,302	95.21	none	9,438	96.84	none
Xerox OCR Engine	14,921	95.00	none	9,218	96.91	none

Table 3g: Character Accuracy, *German Business Letter Sample*

4 Throughput

Throughput reflects both the speed and the accuracy of a page reader. It is defined as $\frac{n - P \times \#\text{errors}}{\#\text{seconds}}$ where n is the number of characters in the correct text. P is the penalty charged for each error. When there is no penalty for errors, i.e., $P = 0$, throughput is equal to the speed of the page reader in terms of characters per second. Accuracy becomes more important as P increases.

Graphs 2a to 2g display throughput separately for the SPARC-based Xerox OCR Engine and the PC-based page readers. Throughput is plotted ranging from no penalty for errors ($P = 0$) to a large penalty ($P = 100$). Notice that throughput becomes negative for large P. It is interesting to observe that the throughput lines intersect the y-axis in order of speed and the x-axis in order of accuracy.

Throughput indicates the trade-off between speed and accuracy. For example, the Xerox OCR Engine processed binary images of the *Corporate Annual Report Sample* at a rate of 220 characters per second. However, when given the gray scale images, it produced one-half the errors but took about three times longer to do so (75 characters per second). When is it desirable to use gray scale instead of binary images? Since the throughput lines intersect at $P = 27$, we conclude that it is advantageous to use gray scale images when the penalty for errors exceeds 27.

In practice, the actual value of P may be difficult to determine and depends on the application. For an application with a high accuracy requirement, errors must be corrected manually; hence, the penalty for errors is large. However, when errors have little consequence to an application, P is small.

5 Accuracy by Character Class

We divided the “ground truth” characters into five classes and determined the percentage of characters in each class that were correctly recognized. Graphs 3a to 3g display the results. The following classes were used:

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

For each non-English sample, a sixth class was added to contain the non-ASCII symbols of the language. This class includes, for example, ¿, á, and ñ for Spanish, and ß, ä, and ö for German.

Depending on the sample, between 66 and 75% of the characters are lowercase letters. Spacing symbols account for 14 to 17% of all characters. Uppercase letters range from 3% of the *Spanish Newspaper Sample* to 11% of the *Legal Document Sample*. Digits account for only

1% of the *Spanish Newspaper Sample*, but as much as 6% of both the *Corporate Annual Report Sample* and the *DOE Sample*, which contain many numeric tables. Punctuation and special symbols make up 3 to 5% of all characters. Approximately 1 out of 50 characters of the *Spanish Newspaper Sample* is a non-ASCII Spanish symbol, whereas non-ASCII German symbols account for about 1 out of 75 characters of the *German Business Letter Sample*.

6 Accuracy vs. Frequency

It is not surprising that in general, the page-reading systems recognize the most frequently occurring symbols, i.e., the lowercase letters, with greater accuracy than less common characters. Tables 4a to 4g show in detail the relationship between accuracy and frequency. For each character and page reader, we determined the percentage of occurrences of this character that were correctly identified by the page reader when processing binary page images. We then computed the median percentage for the set of page readers to obtain an overall accuracy with which the character was recognized. This accuracy determines the row in which the character is placed. The column is based on the frequency of the character within the sample. Very rare symbols occurring less than 1 in 8,192 characters are not shown.

	1/8192 to 1/4096	1/4096 to 1/2048	1/2048 to 1/1024	1/1024 to 1/512	1/512 to 1/256	1/256 to 1/128	1/128 to 1/64	1/64 to 1/32	1/32 to 1/16	1/16 to 1/8
99-100%										
98- 99%			z							
97- 98%				q	x		f g v	c d h l m p u	a i n o r s t	e
96- 97%			:		6 k	2 3 4 5 w	9 b y			
95- 96%			j		7 8 P	C	, .			
94- 95%				B G	A D S T	0	1			
93- 94%		K	J	' F						
92- 93%	X	;	W		) I N	$				
91- 92%		Y	"	M R U	(E O	-				
90- 91%	&		% H	L						
< 90%	* /		V							

Table 4a: Accuracy vs. Frequency, *Corporate Annual Report Sample*

	1/8192 to 1/4096	1/4096 to 1/2048	1/2048 to 1/1024	1/1024 to 1/512	1/512 to 1/256	1/256 to 1/128	1/128 to 1/64	1/64 to 1/32	1/32 to 1/16	1/16 to 1/8
99-100%								d	r t	
98- 99%				q	k	v w	b y	c f h l p u	a i n o s	e
97- 98%			:		C	T	g	m		
96- 97%				x z	F L R	A S		.		
95- 96%		j	" Y		) N	9				
94- 95%	Z		;	W	(8 P					
93- 94%		'	J	H U	7 D I	2 3 5	,			
92- 93%				G	6					
91- 92%			V	B	O	E	1			
90- 91%					M	4				
< 90%	% * Q X	+ =	/ K				- 0			

Table 4b: Accuracy vs. Frequency, *DOE Sample*

	1/8192 to 1/4096	1/4096 to 1/2048	1/2048 to 1/1024	1/1024 to 1/512	1/512 to 1/256	1/256 to 1/128	1/128 to 1/64	1/64 to 1/32	1/32 to 1/16	1/16 to 1/8
99-100%										
98- 99%										
97- 98%										
96- 97%			j q			k v	f g p w y	c d u	a h i n r s	e t
95- 96%							b	m	l o	
94- 95%		?		x			, -			
93- 94%			z							
92- 93%				3	" '	A	.			
91- 92%	X		J	4 7	9 C M	T				
90- 91%			K	F W	0 5 P					
< 90%	& Q	%	$ / ; V	( ) 6 8 : D G H L U Y	1 2 B E I N O R	S				

Table 4c: Accuracy vs. Frequency, *Magazine Sample*

	1/8192 to 1/4096	1/4096 to 1/2048	1/2048 to 1/1024	1/1024 to 1/512	1/512 to 1/256	1/256 to 1/128	1/128 to 1/64	1/64 to 1/32	1/32 to 1/16	1/16 to 1/8
99-100%		z	j	4 k x	5 9	v w	b g m p y	c d f h l u	a i n o r s	e t
98- 99%		&	W	' ( ) 6 7 V Y q	- 2 3 : B F	1 C D I L N O P	, A E R T			
97- 98%		$;		G	8 M U	0 S	.			
96- 97%			"	K	H					
95- 96%			J							
94- 95%										
93- 94%										
92- 93%										
91- 92%			/							
90- 91%										
< 90%	#	Q X								

Table 4d: Accuracy vs. Frequency, *Legal Document Sample*

	1/8192 to 1/4096	1/4096 to 1/2048	1/2048 to 1/1024	1/1024 to 1/512	1/512 to 1/256	1/256 to 1/128	1/128 to 1/64	1/64 to 1/32	1/32 to 1/16	1/16 to 1/8
99-100%	+		z				w	c d h u	i n r s t	e
98- 99%	;						b f g v	m p	a l o	
97- 98%	?		j q	: x	3 4 9	- k		y		
96- 97%			"		' 5 P	0 A C S	.			
95- 96%			)	8	2 D L R W	E				
94- 95%			!	6 7 V Y	F M	1 I T	,			
93- 94%		K X	(	B						
92- 93%					O					
91- 92%			$ J	H	N					
90- 91%		/		U						
< 90%	% &	*		G						

Table 4e: Accuracy vs. Frequency, *English Business Letter Sample*

	1/8192 to 1/4096	1/4096 to 1/2048	1/2048 to 1/1024	1/1024 to 1/512	1/512 to 1/256	1/256 to 1/128	1/128 to 1/64	1/64 to 1/32	1/32 to 1/16	1/16 to 1/8
99-100%		4 V k	9		z		b	p	d t	
98- 99%	7 K	3 8	2 B J	x	j	h q v	- g	u	c n r s	a e o
97- 98%		5 :	F	D M N S é	C E P	f		m	i l	
96- 97%	Z	6 Y	0	L	A á	. ó				
95- 96%			T	1 O R ú		y				
94- 95%			G U	ñ						
93- 94%	?		(H	I						
92- 93%										
91- 92%							,			
90- 91%	;									
< 90%	¿	« »	)		"	í				

Table 4f: Accuracy vs. Frequency, *Spanish Newspaper Sample*

	1/8192 to 1/4096	1/4096 to 1/2048	1/2048 to 1/1024	1/1024 to 1/512	1/512 to 1/256	1/256 to 1/128	1/128 to 1/64	1/64 to 1/32	1/32 to 1/16	1/16 to 1/8
99-100%										
98- 99%			j							
97- 98%						P p v w z	k	d g u	i s t	e n r
96- 97%		X					S b	h m o	a	
95- 96%					V W ä ö	3 A D F K	- f	c l		
94- 95%						, 2 7 E H M				
93- 94%		!		x y	4 L U	1 5 6 9 G T ü	.			
92- 93%	Q				8 C R	I				
91- 92%			"	Z	N O	B	0			
90- 91%		;	J							
< 90%	' Y Ü	+		( ) /	: ß					

Table 4g: Accuracy vs. Frequency, *German Business Letter Sample*

7 Effect of Resolution

Graphs 4a to 4g display the character accuracy for binary images of various resolutions. Data points that are missing from these graphs are due to excessive failures.

This data shows that in general, a substantial increase in errors can be expected when decreasing the resolution from 300 to 200 dpi. However, little or no benefit may be obtained when increasing the resolution from 300 to 400 dpi.

The fine-mode fax images have essentially the same resolution as the 200 dpi images, but the standard-mode fax images have only half the resolution in the vertical dimension. This additional loss of fidelity is the source of many more errors.

8 Effect of Page Quality

For each page, we computed the median character accuracy achieved by the set of page readers, which gives us a measure of the page's "quality" or "OCR difficulty." We used this measure to partition each sample into five Page Quality Groups of approximately equal size. Page Quality Group 1 contains the pages with the highest median accuracy (best page quality), while Page Quality Group 5 contains the pages with the lowest median accuracy (worst page quality). Graphs 5a to 5g plot the character accuracy within each group to show the effect of page quality.

In general, 50 to 75% of all errors are made on the worst 20% of each sample, i.e., Group 5. By examining pages within this group, we can gain insight into what makes OCR difficult. Figures 1 to 7 show snippets selected from the 300 dpi binary images of pages in Group 5. Each snippet has been scaled to twice its original size to make it easier to observe its properties. Each sample provides a unique combination of challenges:

1. *Corporate Annual Report Sample* – creative typefaces, reverse video, shaded backgrounds, numeric tables, broken characters.
2. *DOE Sample* – broken and touching characters from photocopies, skewed and curved baselines, numeric tables.
3. *Magazine Sample* – creative typefaces, reverse video, shaded backgrounds, shadowing, drop-caps.
4. *Legal Document Sample* – broken and touching characters from photocopies, signatures intersecting text, underlined fields that are filled in, or left blank as in _____.
5. *English Business Letter Sample* – creative letterheads, broken characters (some caused by creases in the hard copy), signatures intersecting text.
6. *Spanish Newspaper Sample* – broken and touching characters, speckling.
7. *German Business Letter Sample* – creative letterheads with much small print, broken and touching characters, signatures intersecting text.

9 Word Accuracy

A popular use of a page-reading system is to create a text database from a collection of hard-copy documents. Information retrieval techniques can then be applied to locate documents of interest. For this application, the correct recognition of words is paramount.

We define a word to be any sequence of one or more letters. In word accuracy, we determine the percentage of words that are correctly recognized. Each letter of the word must be correctly identified, although we permit errors in case (e.g., "uniVerSity" generated for "University") because full-text searching is normally insensitive to case. Errors in recognizing digits or punctuation have no effect on word accuracy.

Graphs 6a to 6g display approximate 95% confidence intervals for word accuracy.

10 Non-stopword Accuracy

Stopwords are common words such as *the*, *of*, *and*, *to*, and *a* in English; *de*, *la*, *el*, *y*, and *en* in Spanish; and *der*, *die*, *in*, *und*, and *von* in German. These words are normally not indexed by a text retrieval system because they are not useful for retrieval. Users search for documents by specifying non-stopwords in queries.

With this in mind, we wish to determine the percentage of non-stopwords that are correctly recognized, i.e., the non-stopword accuracy. To do this, we need a list of stopwords for each language. In past years, we computed non-stopword accuracy using the default set of 110 English stopwords provided by the BASISPLUS text retrieval system. For this year's test, we obtained a list of 200 stopwords for each language. Furthermore, the stopwords in each list are ordered by their frequency of occurrence, which was determined from a large corpus. For English, we chose the 200 most common stopwords from a well-known stopword list [Frakes 92] using the frequency data of the Brown Corpus [Kucera 67]. We obtained a list of 200 Spanish stopwords from Cornell University, and a list of 200 German stopwords from ETH in Switzerland; both of these lists were ordered by frequency. The 200 English stopwords account for 37 to 45% of the words in the English samples; the 200 Spanish stopwords make up 49% of the words in the *Spanish Newspaper Sample*; and the 200 German stopwords account for 30% of the words in the *German Business Letter Sample*.

These ordered lists of stopwords enabled us to compute non-stopword accuracy as a function of the number of stopwords, and the results are presented in Graphs 7a to 7g. For N stopwords, non-stopword accuracy was computed by excluding the N most common stopwords. The graphs show non-stopword accuracy for $N = 0$ to 200. When $N = 0$, no words are excluded and non-stopword accuracy is equal to the word accuracy. When $N = 1$, we see the effect of excluding the most common stopword (*the* in English), and as N increases, we observe the effect of excluding more and more stopwords. Eventually, the curve levels off to indicate the page reader's ability to recognize uncommon words.

11 Phrase Accuracy

Users search for documents containing specific phrases. We define a phrase of length *L* to be any sequence of *L* words. For example, the phrases of length 3 in "University of Nevada, Las Vegas" are "University of Nevada," "of Nevada, Las," and "Nevada, Las Vegas." For a phrase to be correctly recognized, all of its words must be correctly identified. Phrase accuracy is the percentage of phrases that are correctly recognized, and we have computed it for $L = 1$ through 8. Graphs 8a to 8g display the results. The phrase accuracy for length 1 is equal to the word accuracy.

Phrase accuracy reflects the extent to which errors are bunched or scattered within generated text. Suppose two page readers, *A* and *B*, have the same word accuracy but *A* has a higher phrase accuracy than *B*. Then *A*'s errors are more closely bunched, and hence, easier to correct, than *B*'s errors.

12 Marked Character Efficiency

Page-reading systems place flags in their output to help users locate errors. If a system has no clue as to the identity of a symbol, it will generate a reject character, as in "N~vada." If the system believes that the symbol is most likely an "a" but is not confident with its choice, then it may precede the "a" with a suspect marker to call attention to it, as in "N^avada." These examples show how a flag in the generated text can direct a user to an error. But if a correctly-generated character is marked as suspect, as in "N^evada," then it is a false mark which slows the user who must take time to verify the correctness of this character. The goal of a page reader is to mark as many of its errors as possible while minimizing the number of false marks.

In marked character efficiency, we measure the utility of marked characters (reject characters and suspect markers). Graphs 9a to 9g show how the accuracy of generated text increases as more and more marked characters are examined and errors are corrected. The starting point for each curve is the base character accuracy of the text, that is, the accuracy before any corrections are made. The curve then rises sharply to show the effect of correcting the reject characters. It then gradually flattens as more and more suspect markers are examined. This flattening reflects an increasing percentage of false marks and a corresponding decrease in the efficiency of the correction process. (The INM NeuroTalker and Recognita OCR do not support suspect markers so their curves show only the effect of correcting reject characters.) Since not all errors are flagged, the accuracy of the generated text falls short of 100% even after all marked characters have been examined. The remaining errors can be located by proofreading the text, but this is tedious and costly.

13 Automatic Zoning

In the test of automatic zoning, the page readers were tasked with locating the text regions on each page and determining their correct reading order. The method used to evaluate their performance is described in [Kanai 95]. Essentially, the generated text is matched with the correct text to identify missing blocks of text and blocks that are out of order. The cost of correcting the generated text is then estimated based on the number of insertions needed to enter missing blocks, and the number and length of move operations needed to re-order blocks. By converting moves into an equivalent number of insertions, the cost of correction is given solely in terms of insertions and is normalized by dividing it by the number of characters in the sample.

The automatic zoning test was performed only for three of the seven samples, and the results are displayed in Graphs 10a to 10c. Each curve shows the normalized cost of correcting the automatic zoning errors for a range of conversion factors, 0 to 100. The INM NeuroTalker does not support automatic zoning and is therefore missing from these graphs. The Xerox OCR Engine does not support automatic zoning of a gray scale image. Gray scale results for the Nankai Reader are missing due to excessive failures.

14 Conclusion

There is no clear "winner" in this year's test. Depending on the sample, a case can be made for either the Nankai Reader, Recognita OCR, or the Xerox OCR Engine as the top system. Maxsoft-Ocron Recore is not far behind these three systems; however, the INM NeuroTalker is substantially behind.

Gray scale input demonstrated its usefulness for some, but not all of the samples. While it offered little or no advantage over binary input for the *DOE Sample* and the *Legal Document Sample*, it enabled one page reader, the Xerox OCR Engine, to reduce its errors by nearly 50% in the *Corporate Annual Report Sample* and the *Spanish Newspaper Sample*. However, processing gray scale input comes at a price. Gray scale images require considerably more storage than binary images, and they take much longer to process. The Nankai Reader was especially slow, needing at least eight times longer to process a gray scale image than a binary image. Recognita OCR and the Xerox OCR Engine were less affected, taking only two to three times longer.

By re-using some samples from the previous year's test, we are able to chart the progress made by participants from one year to the next. Table 5 shows how the accuracy of three page-reading systems has improved in the past year. However, for all three systems, the increased accuracy was achieved at the expense of decreased speed. Maxsoft-Ocron Recore slowed down the most: Version 4.1 took about 2.5 times longer to process a page than last year's entry, Version 3.2. Recognita OCR Version 3.02 consumed 30% more time than its predecessor (Version 3.0), and the Xerox OCR Engine Version 11.0 needed about 40% more time than Version 10.5.

One final note: ISRI is a strong advocate of page-reading technology, but does not endorse any particular page-reading system or systems.

	DOE Sample	*English Business Letter Sample*	*Spanish Newspaper Sample*
Maxsoft-Ocron Recore	10	7	—
Recognita OCR	21	40	14
Xerox OCR Engine	11	10	20

Table 5: % Error Reduction in the Past Year

Acknowledgments

We thank Phil Cheatle for suggesting the non-stopword accuracy curves, and Donna Harman, Chris Buckley, Philipp Haefliger, and Peter Schauble for their assistance with the stopword lists. We also thank Sheila Rice, Diego Vinas, and Julia Cataldo for their help with the selection and layout of the snippets appearing in Figures 1 to 7.

Bibliography

[Frakes 92] William B. Frakes and Ricardo Baeza-Yates, editors. *Information Retrieval: Data Structures & Algorithms*, pages 114-115. Prentice Hall, 1992.

[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.

[Kucera 67] Henry Kucera and W. Nelson Francis. *Computational Analysis of Present-Day American English*. Brown University Press, 1967.

[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.

[Nartker 95] Thomas A. Nartker, Stephen V. Rice, and Frank R. Jenkins. OCR accuracy: UNLV's fourth annual test. *Inform*, Association for Information and Image Management, 9(7):38+, July 1995.

[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.

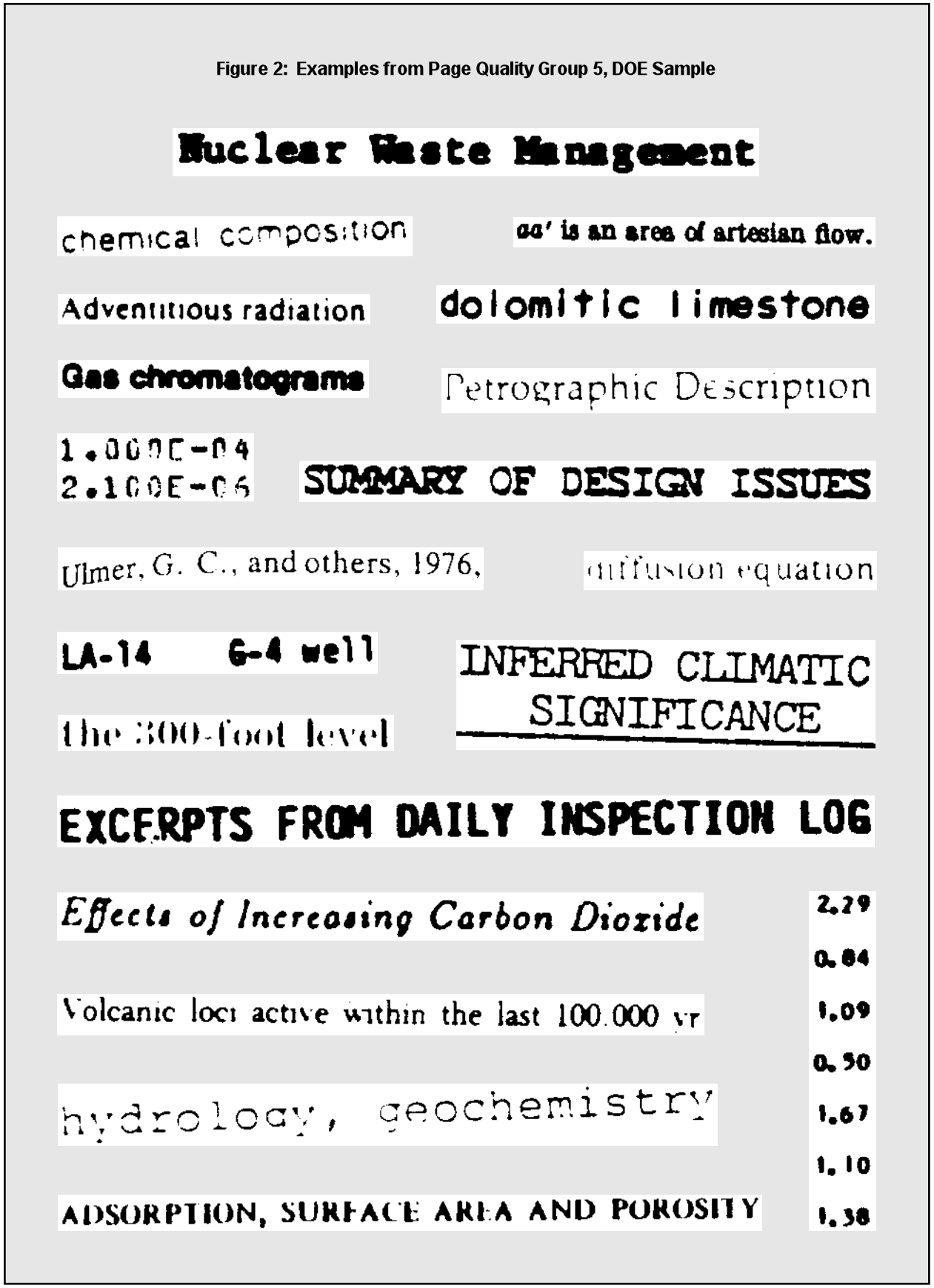
[Rice 93] Stephen V. Rice, Junichi Kanai, and Thomas A. Nartker. An evaluation of OCR accuracy. Technical Report 93-01, Information Science Research Institute, University of Nevada, Las Vegas, April 1993.

[Rice 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.

[Rice 95] Stephen V. Rice, Frank R. Jenkins, and Thomas A. Nartker. The fourth annual test of OCR accuracy. Technical Report 95-04, Information Science Research Institute, University of Nevada, Las Vegas, April 1995.

Figure 1: Examples from Page Quality Group 5, Corporate Annual Report Sample

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



Nuclear Waste Management

chemical composition

aa' is an area of artesian flow.

Adventitious radiation

dolomitic limestone

Gas chromatograms

Petrographic Description

1.000E-04
2.100E-06

SUMMARY OF DESIGN ISSUES

Ulmer, G. C., and others, 1976,

diffusion equation

LA-14 G-4 well

INFERRED CLIMATIC SIGNIFICANCE

the 300-foot level

EXCERPTS FROM DAILY INSPECTION LOG

Effects of Increasing Carbon Dioxide

2.29
0.84
1.09
0.50
1.67
1.10
1.38

Volcanic loci active within the last 100,000 yr

hydrology, geochemistry

ADSORPTION, SURFACE AREA AND POROSITY

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

Figure 4: Examples from Page Quality Group 5, Legal Document Sample

THE DEKALB CIRCUIT COURT

IT IS HEREBY ORDERED,

disbursement costs **EX PARTE MOTION**

assumes all of such obligations in writing.

EDWARD R.J. KANE, Esq.
Nevada Bar No. 001438

David L. Tanner, Esq., P.C. Gibson, Dunn & Crutcher

IT IS FURTHER ORDERED

$ 100 per month

the Bankruptcy Code

David M. Crosby, Esq.
Attorney for Debtor(s)

EXHIBIT A

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

Figure 6: Examples from Page Quality Group 5, Spanish Newspaper Sample

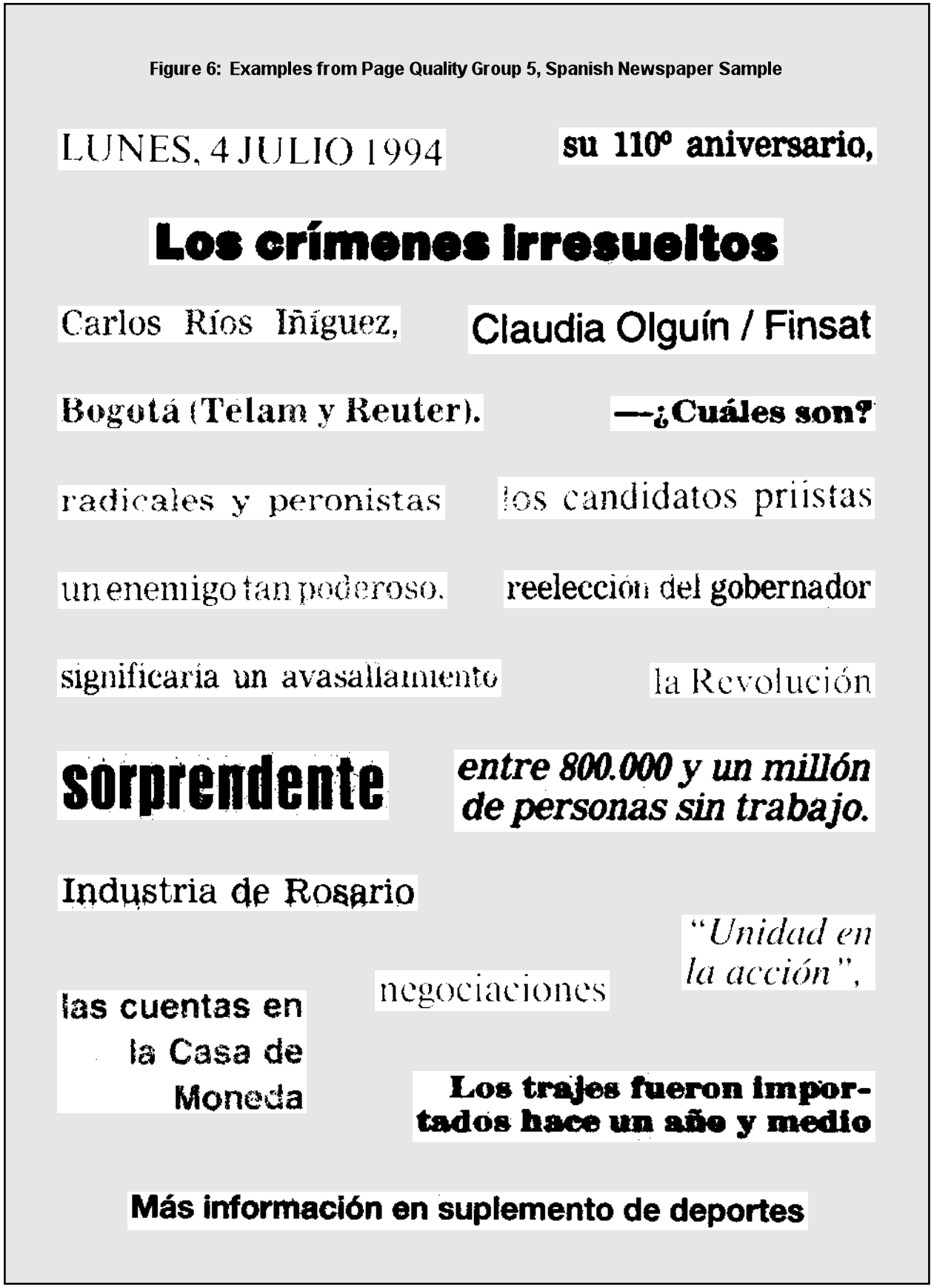
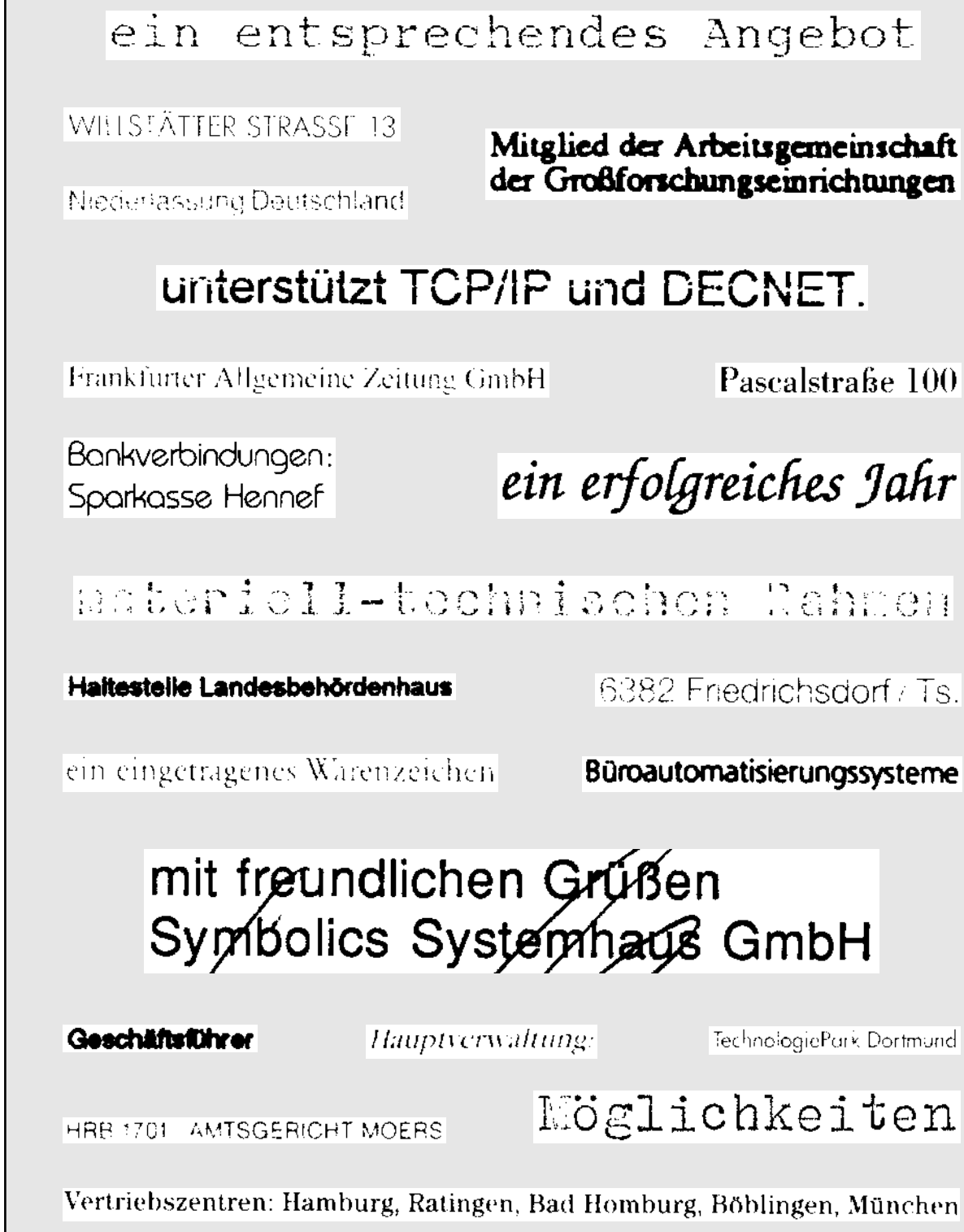
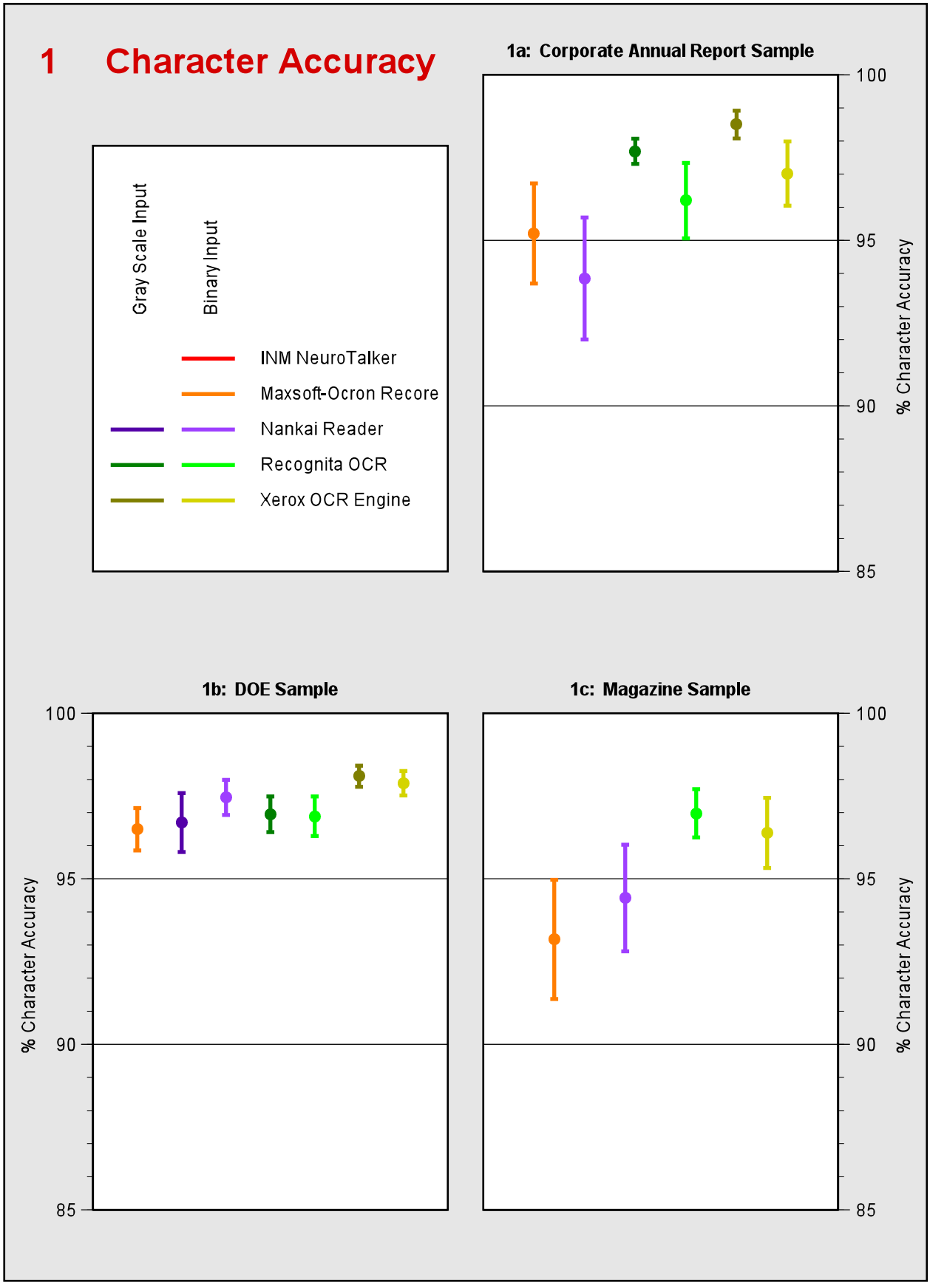
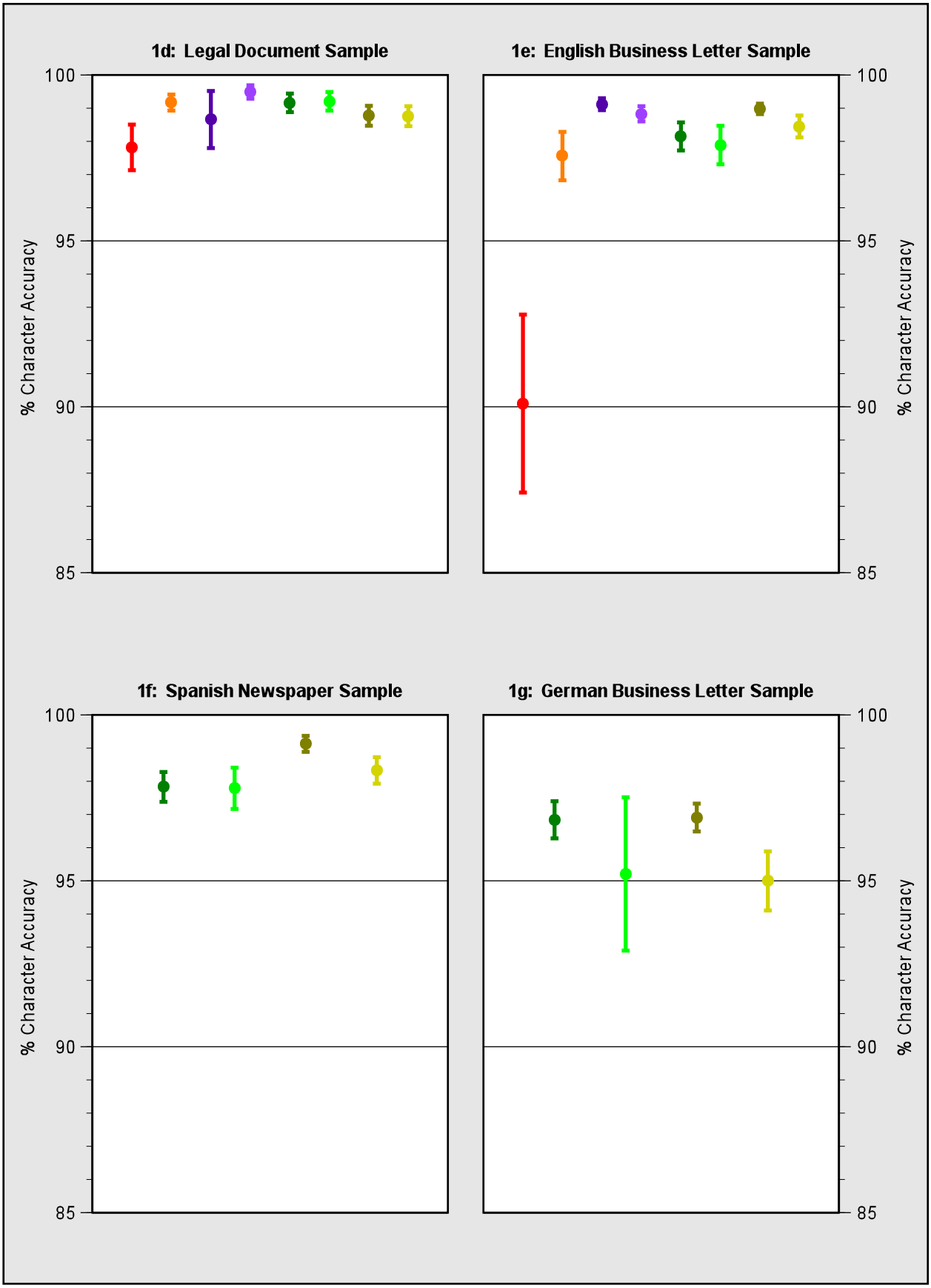
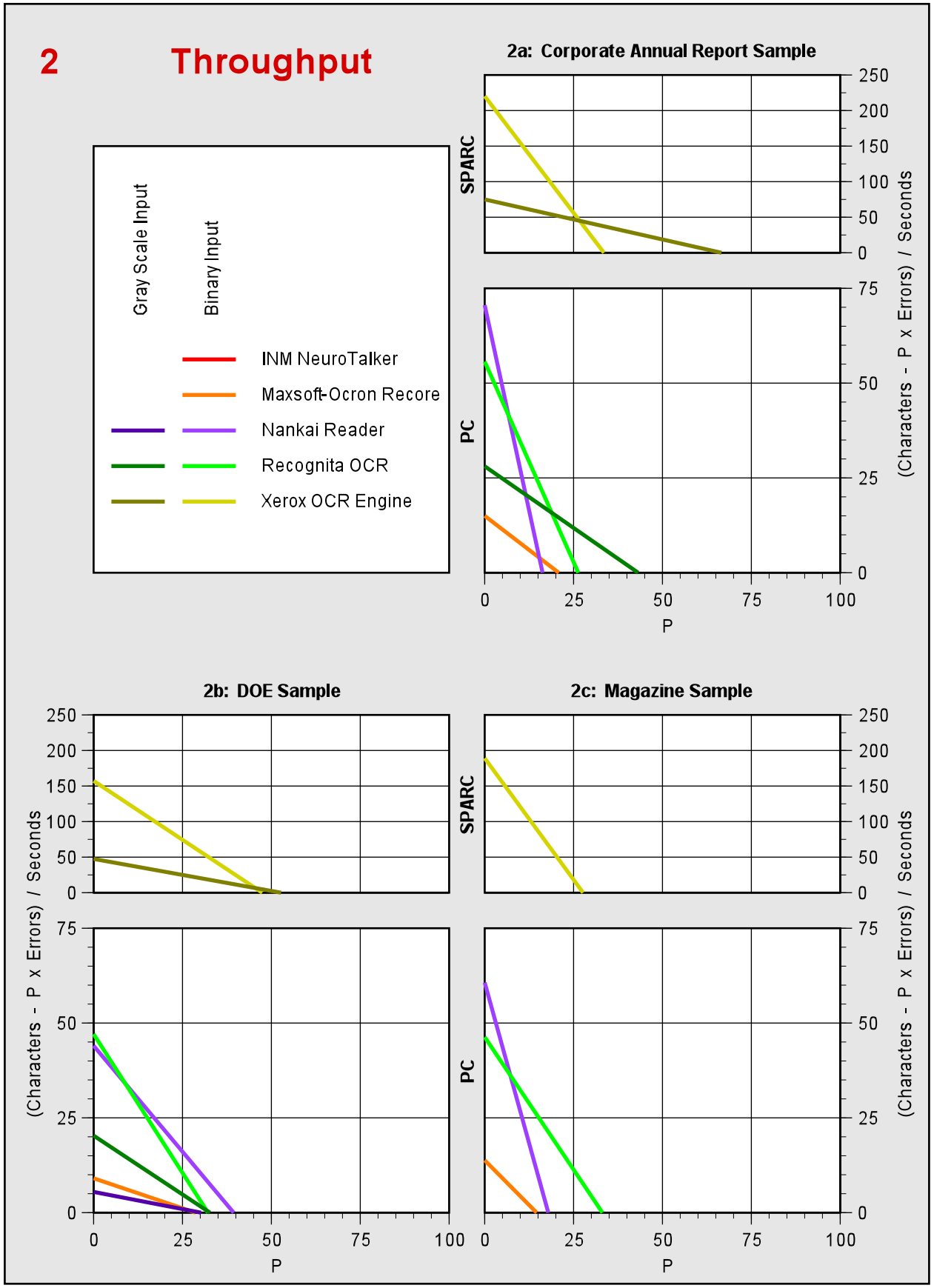


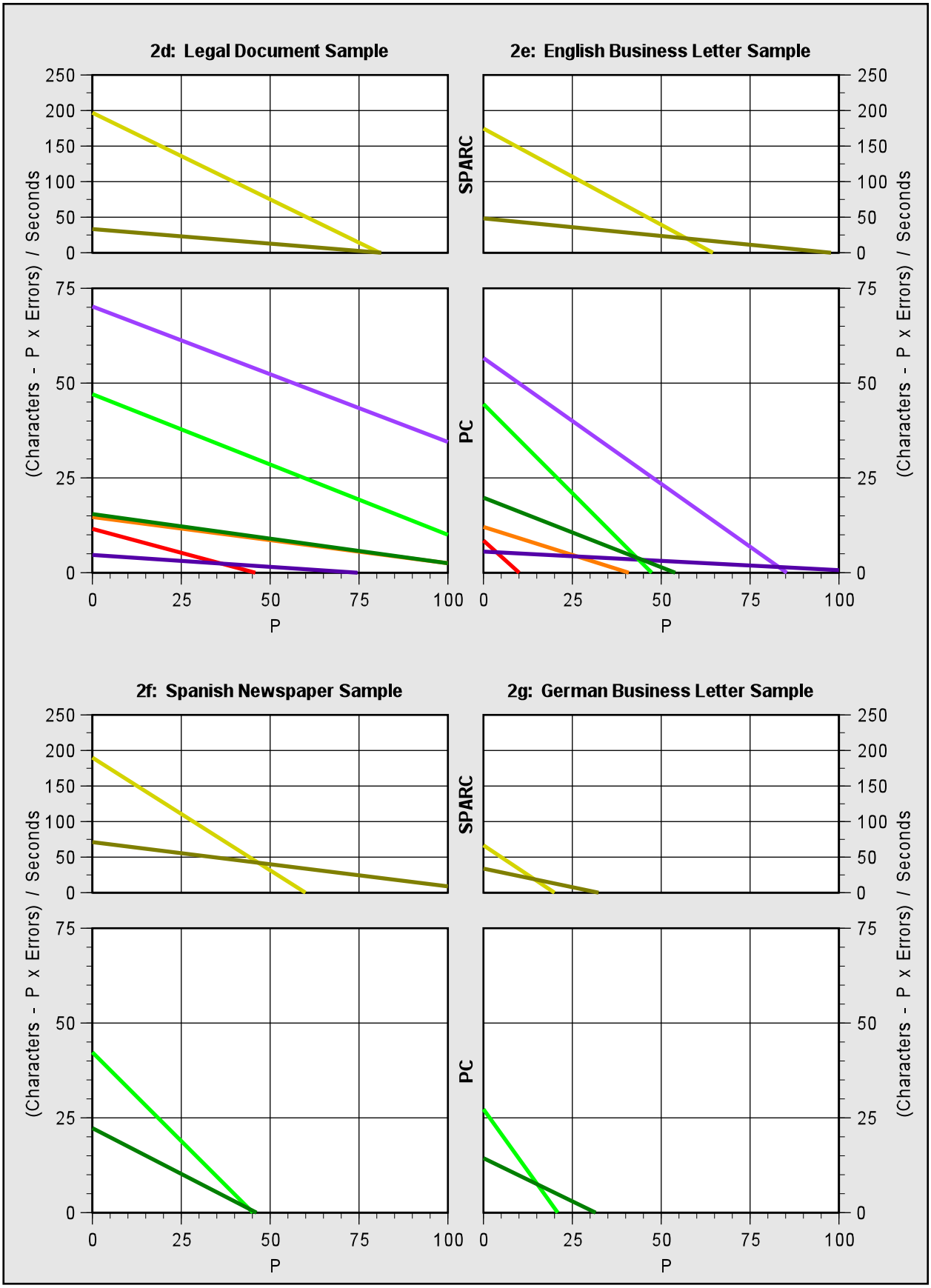
Figure 7: Examples from Page Quality Group 5, German Business Letter Sample



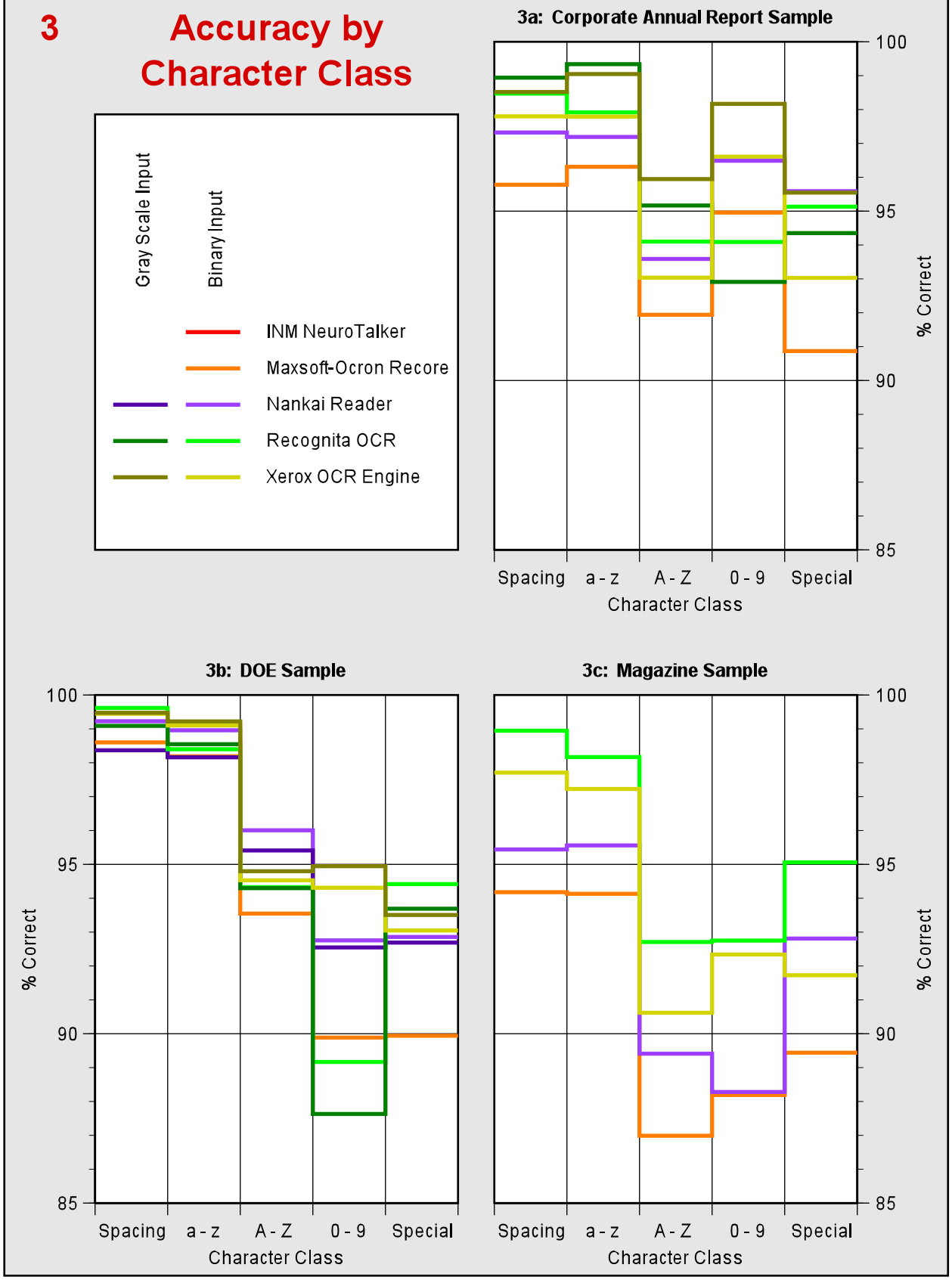

1 Character Accuracy
1a: Corporate Annual Report Sample
Gray Scale Input
Binary Input
INM NeuroTalker
Maxsoft-Ocron Recore
Nankai Reader
Recognita OCR
Xerox OCR Engine
100
95
90
85
% Character Accuracy
1b: DOE Sample
1c: Magazine Sample

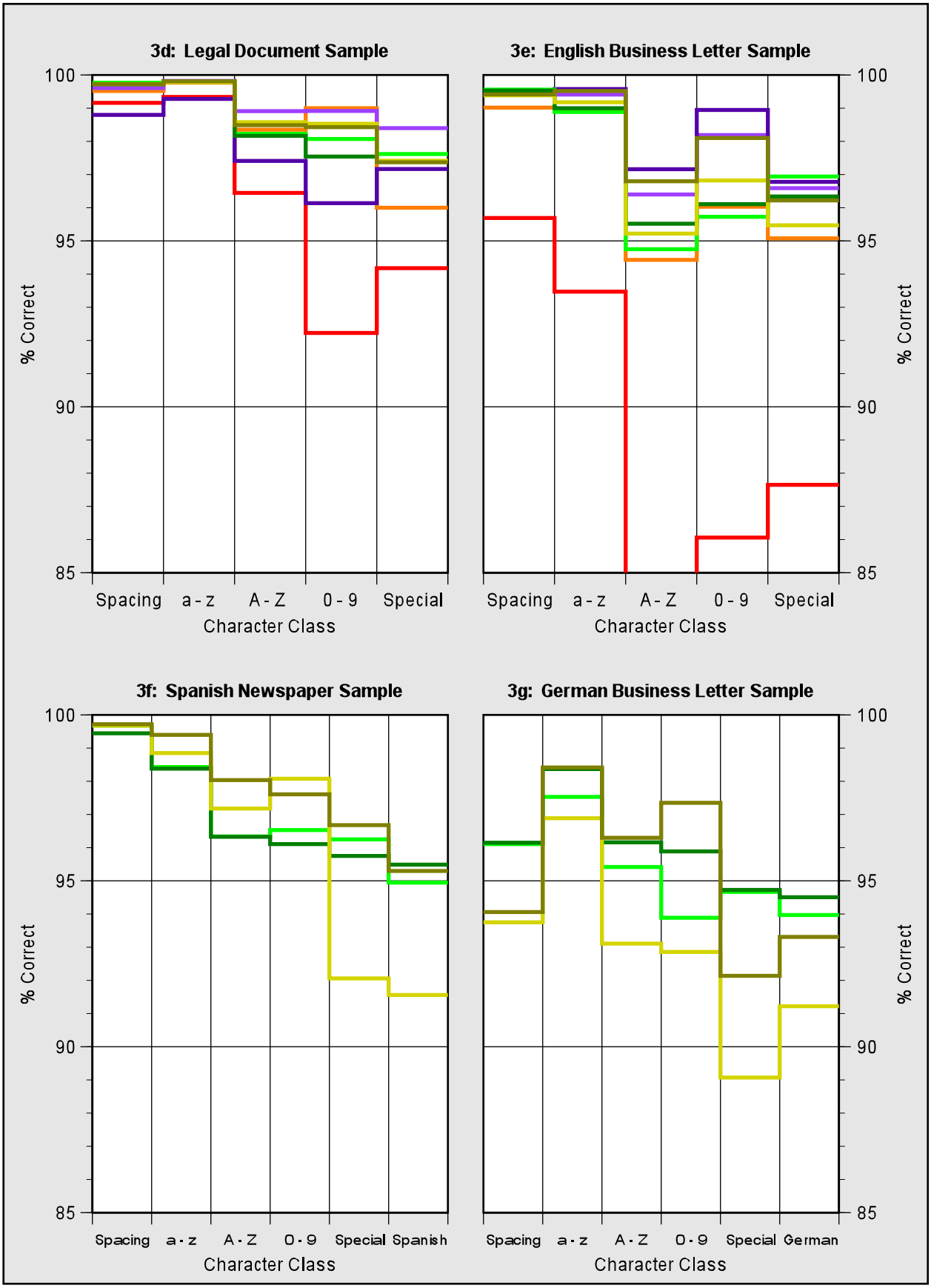

1d: Legal Document Sample
1e: English Business Letter Sample
1f: Spanish Newspaper Sample
1g: German Business Letter Sample
% Character Accuracy
100
95
90
85


2 Throughput
2a: Corporate Annual Report Sample
Gray Scale Input
Binary Input
INM NeuroTalker
Maxsoft-Ocron Recore
Nankai Reader
Recognita OCR
Xerox OCR Engine
SPARC
PC
(Characters - P x Errors) / Seconds
P
2b: DOE Sample
2c: Magazine Sample

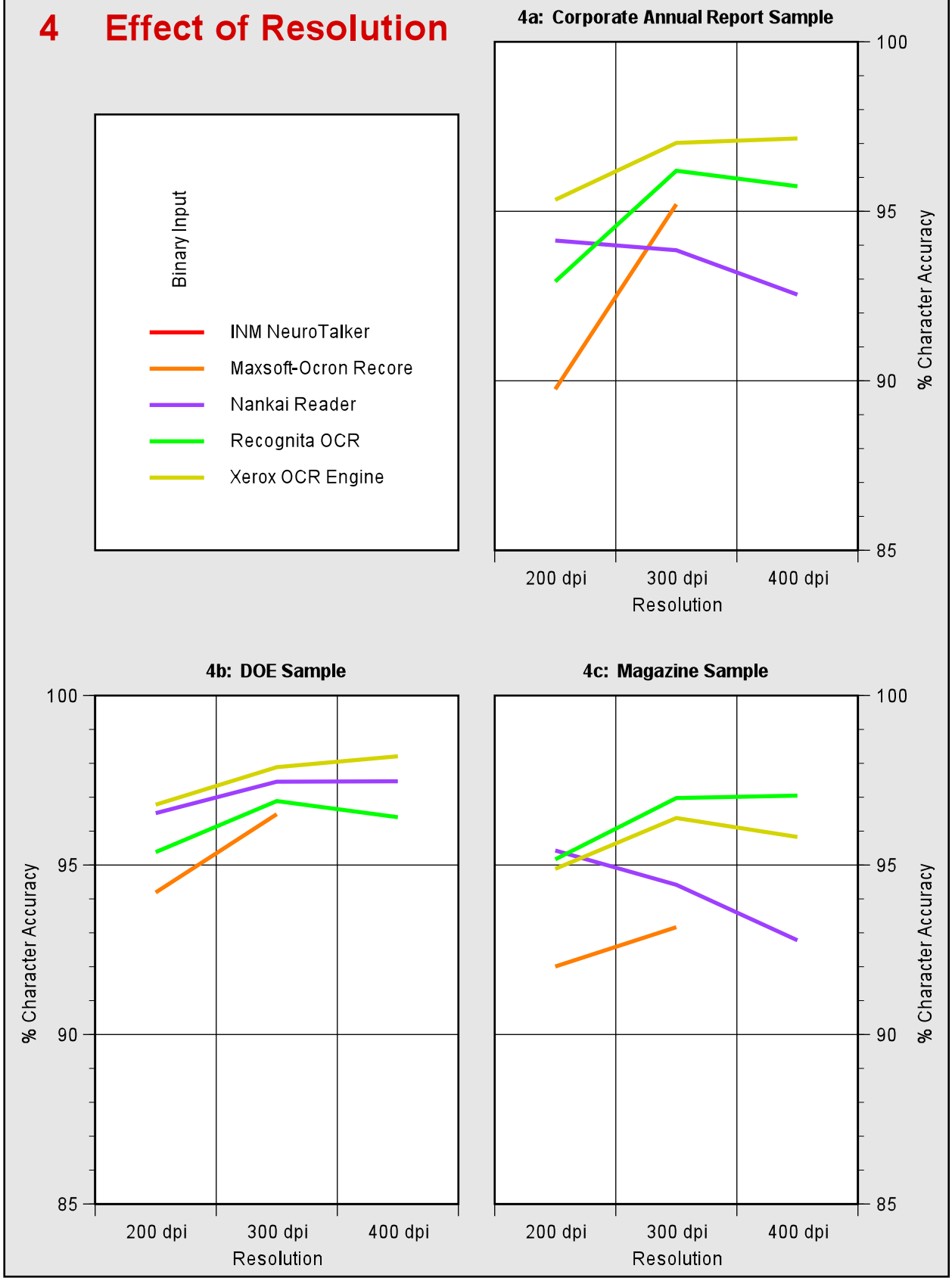

2d: Legal Document Sample
2e: English Business Letter Sample
2f: Spanish Newspaper Sample
2g: German Business Letter Sample
SPARC
PC
(Characters - P x Errors) / Seconds
P
0
25
50
75
100
150
200
250

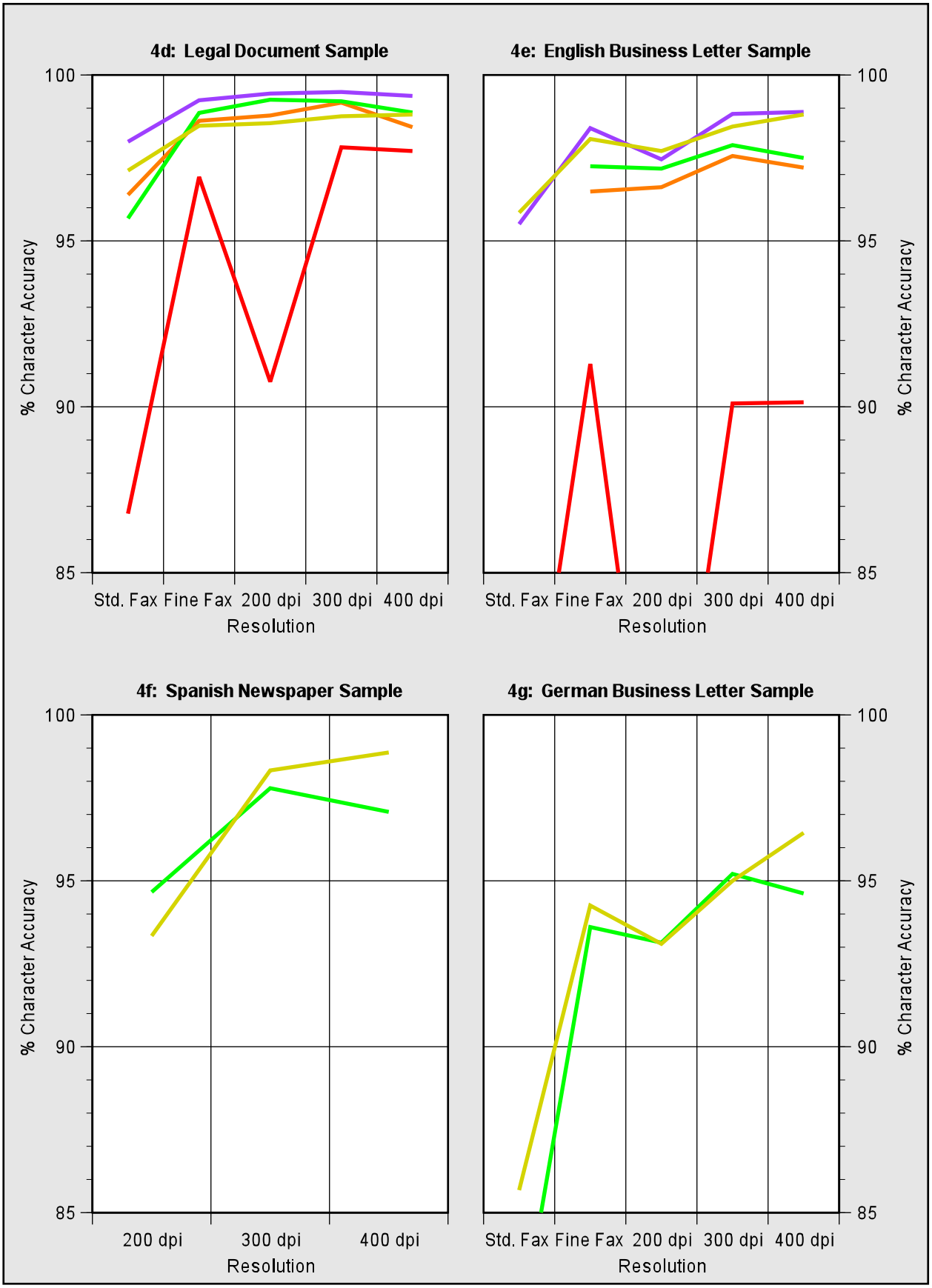
3 Accuracy by Character Class

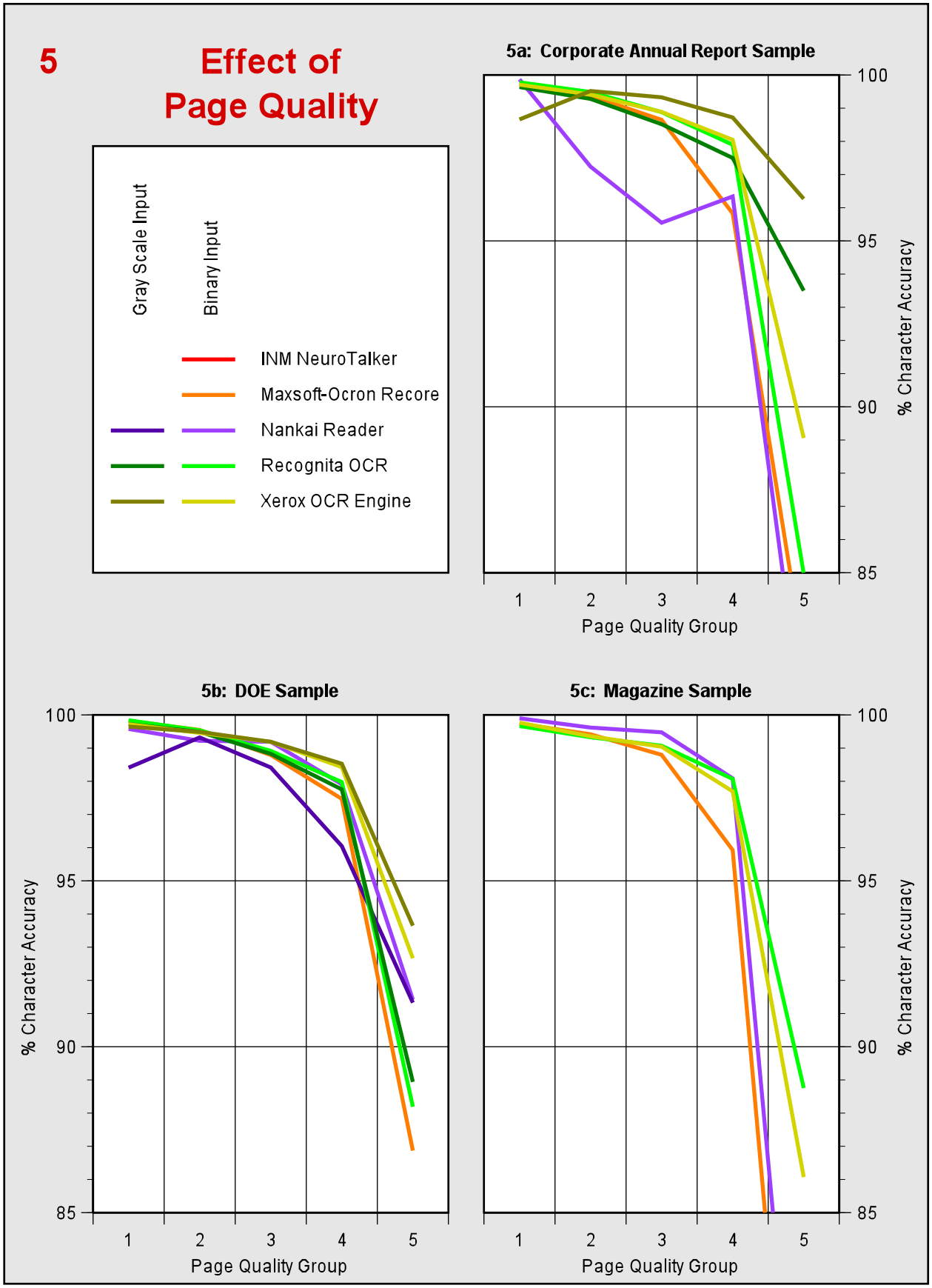



3d: Legal Document Sample
3e: English Business Letter Sample
3f: Spanish Newspaper Sample
3g: German Business Letter Sample
100
95
90
85
% Correct
Spacing
a - z
A - Z
0 - 9
Special
Special Spanish
Special German
Character Class

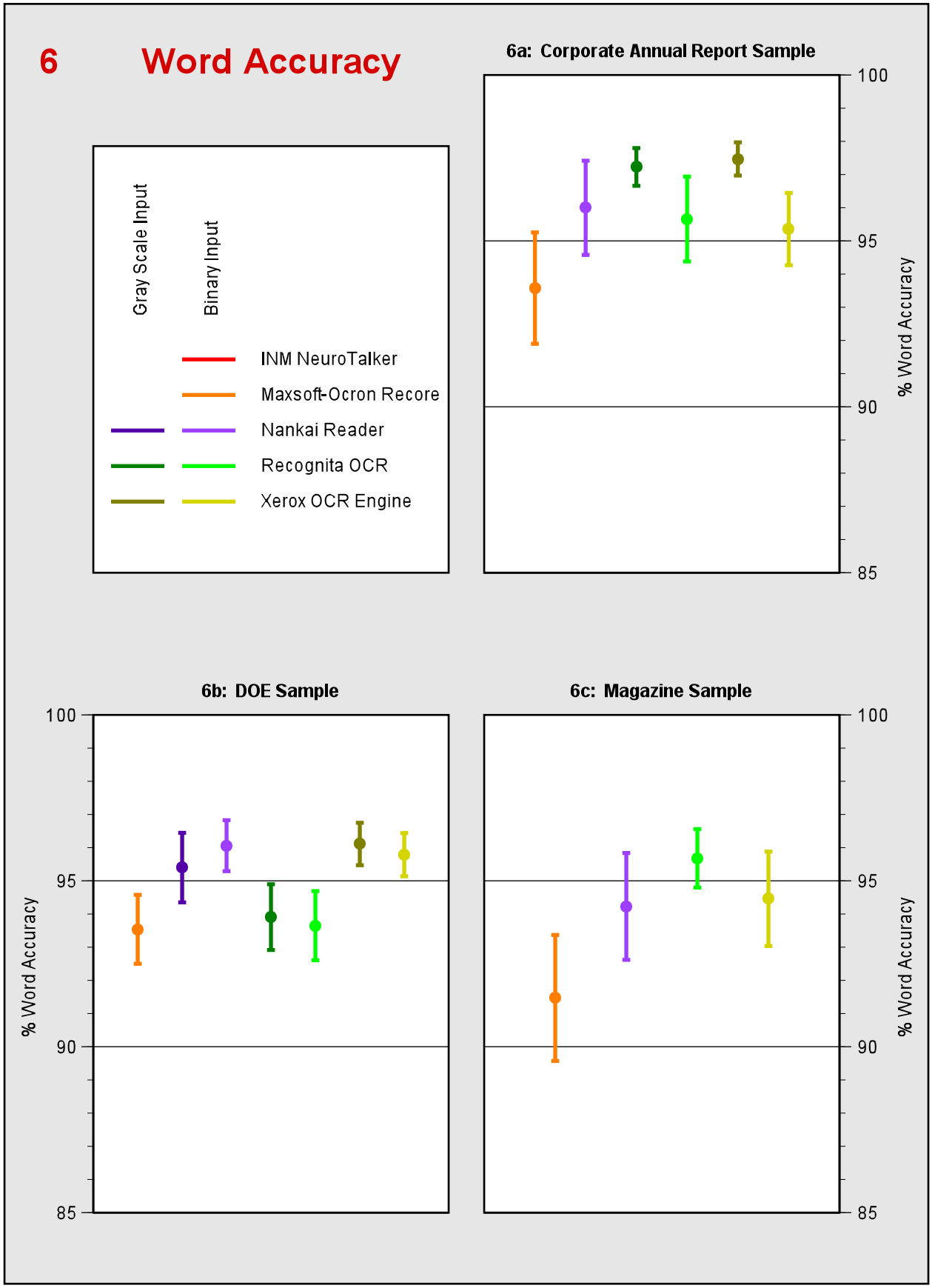
4 Effect of Resolution

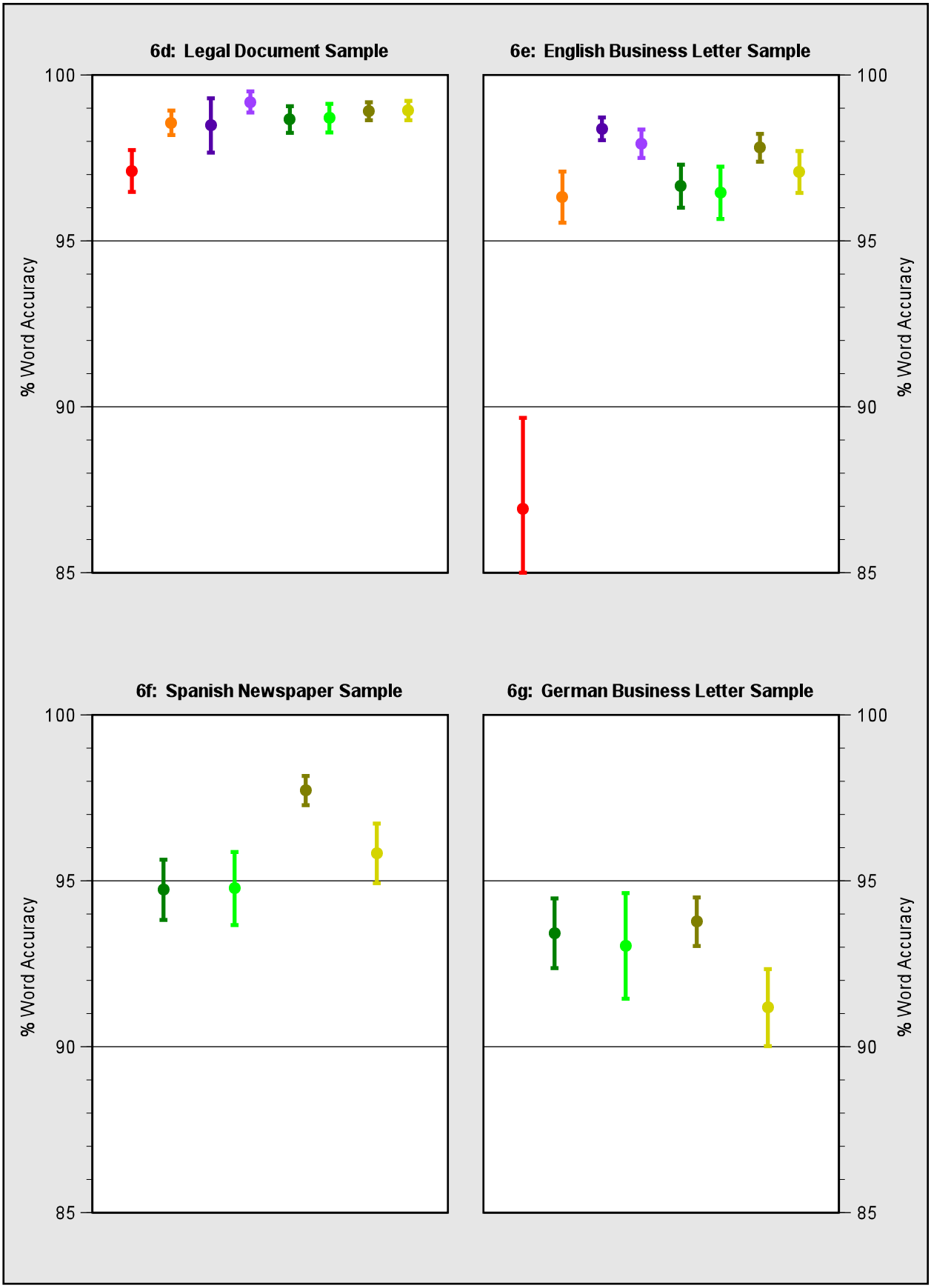



4d: Legal Document Sample
4e: English Business Letter Sample
4f: Spanish Newspaper Sample
4g: German Business Letter Sample
% Character Accuracy
100
95
90
85
Std. Fax
Fine Fax
200 dpi
300 dpi
400 dpi
Resolution

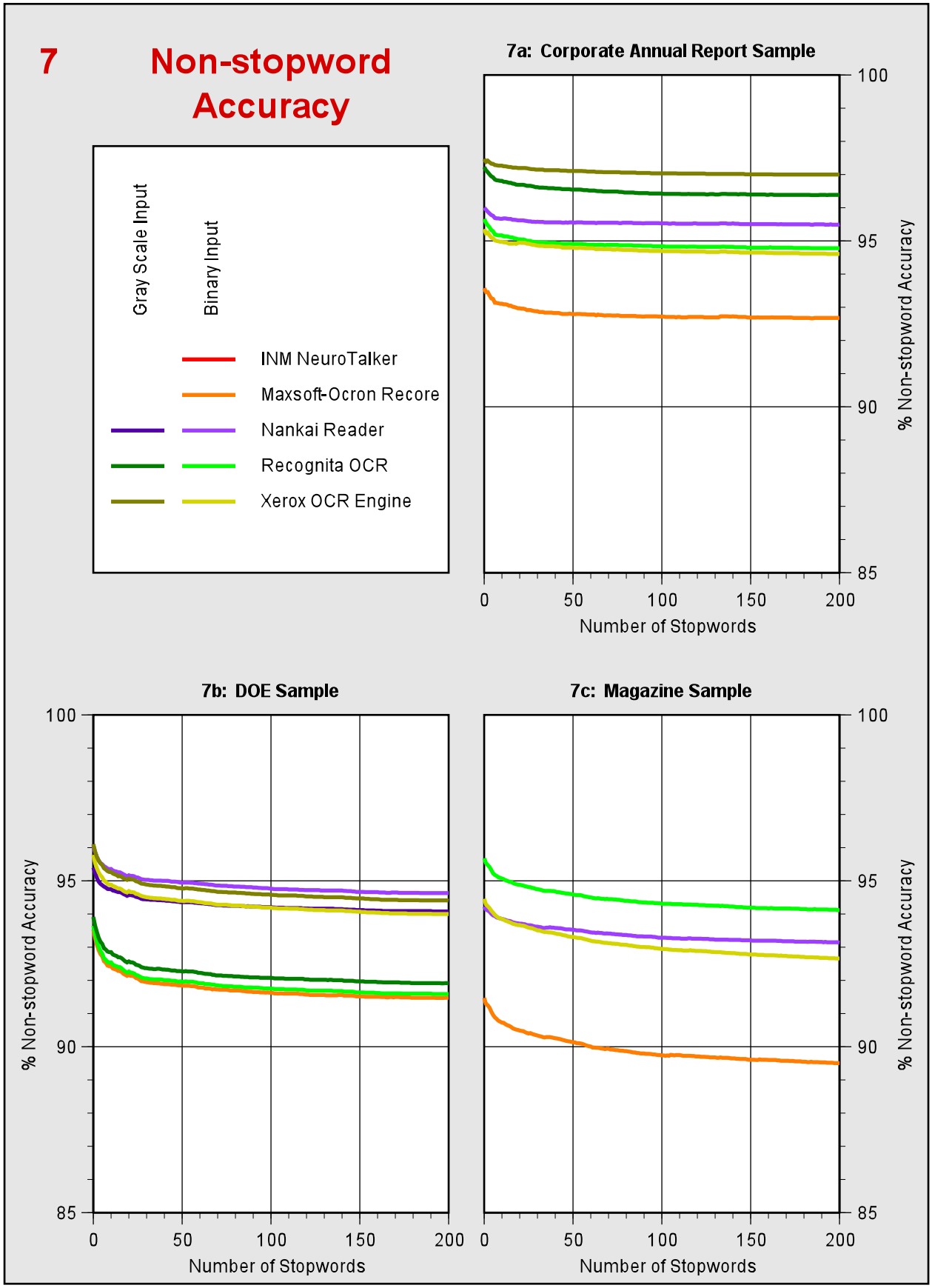

5
Effect of Page Quality
Gray Scale Input
Binary Input
INM NeuroTalker
Maxsoft-Ocron Recore
Nankai Reader
Recognita OCR
Xerox OCR Engine
5a: Corporate Annual Report Sample
5b: DOE Sample
5c: Magazine Sample
% Character Accuracy
Page Quality Group
100
95
90
85
1
2
3
4
5

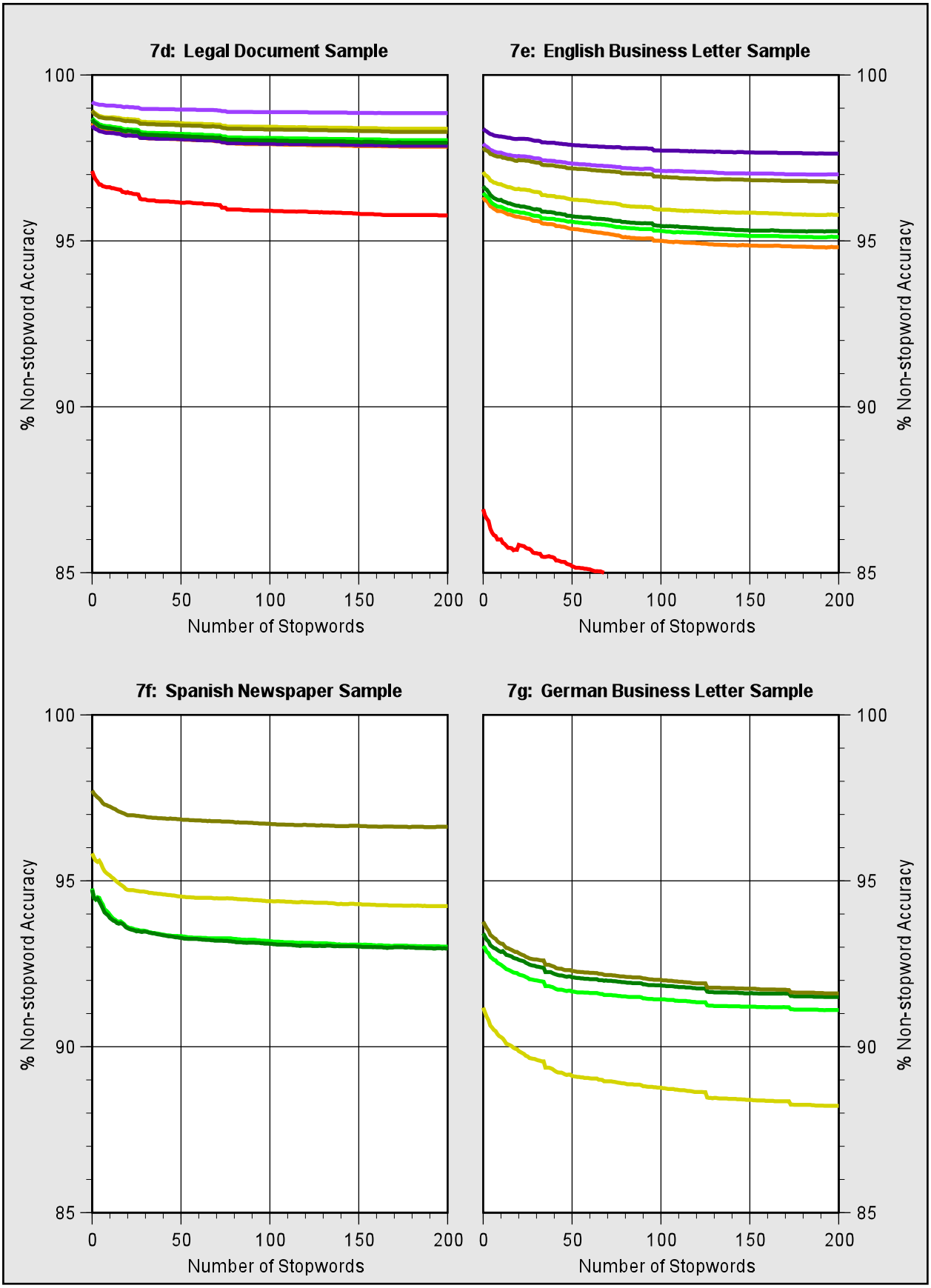

5d: Legal Document Sample
5e: English Business Letter Sample
5f: Spanish Newspaper Sample
5g: German Business Letter Sample
% Character Accuracy
100
95
90
85
1
2
3
4
5
Page Quality Group


6 Word Accuracy
Gray Scale Input
Binary Input
INM NeuroTalker
Maxsoft-Ocron Recore
Nankai Reader
Recognita OCR
Xerox OCR Engine
6a: Corporate Annual Report Sample
6b: DOE Sample
6c: Magazine Sample
% Word Accuracy
100
95
90
85

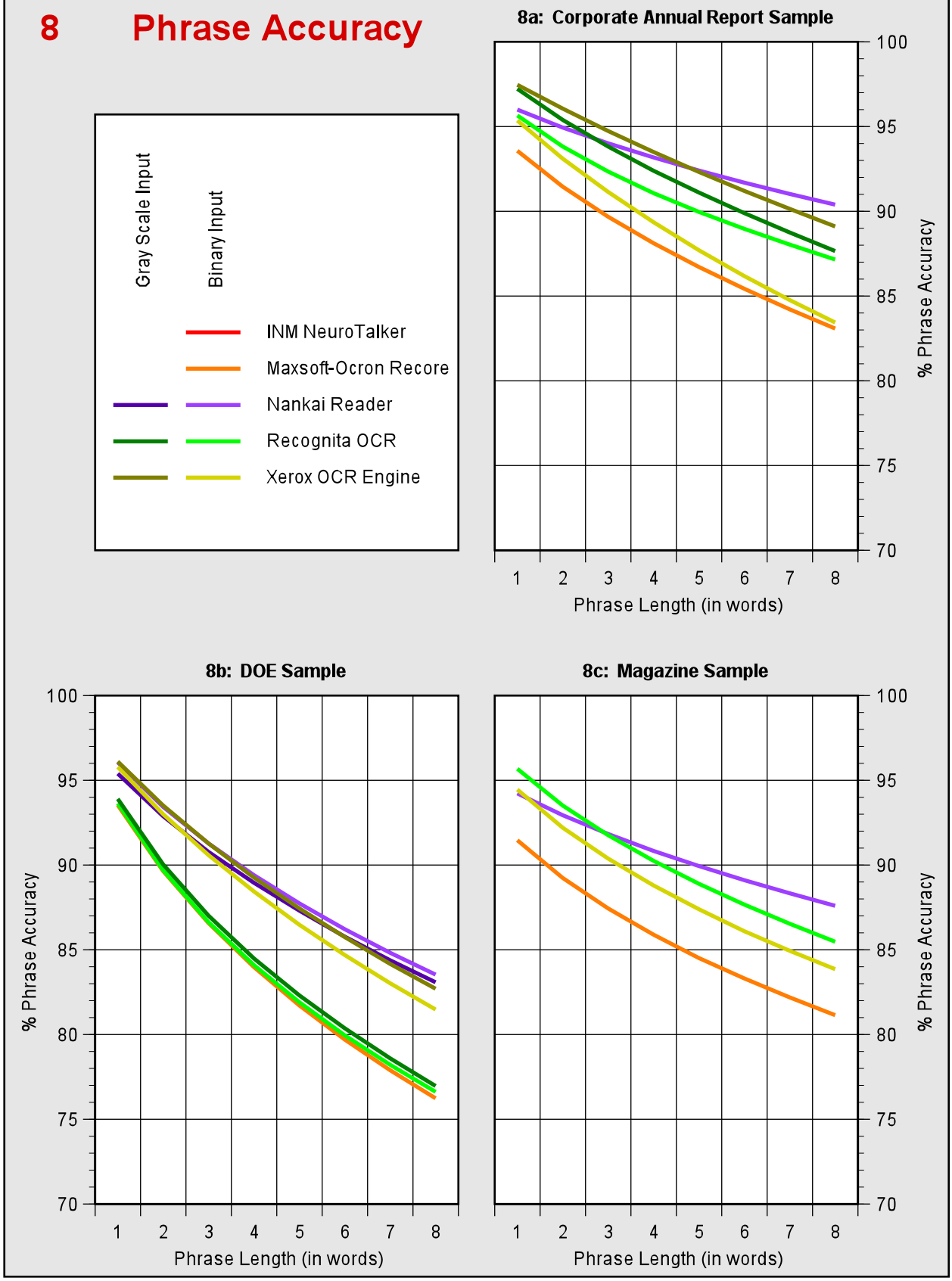

6d: Legal Document Sample
6e: English Business Letter Sample
6f: Spanish Newspaper Sample
6g: German Business Letter Sample
% Word Accuracy
100
95
90
85

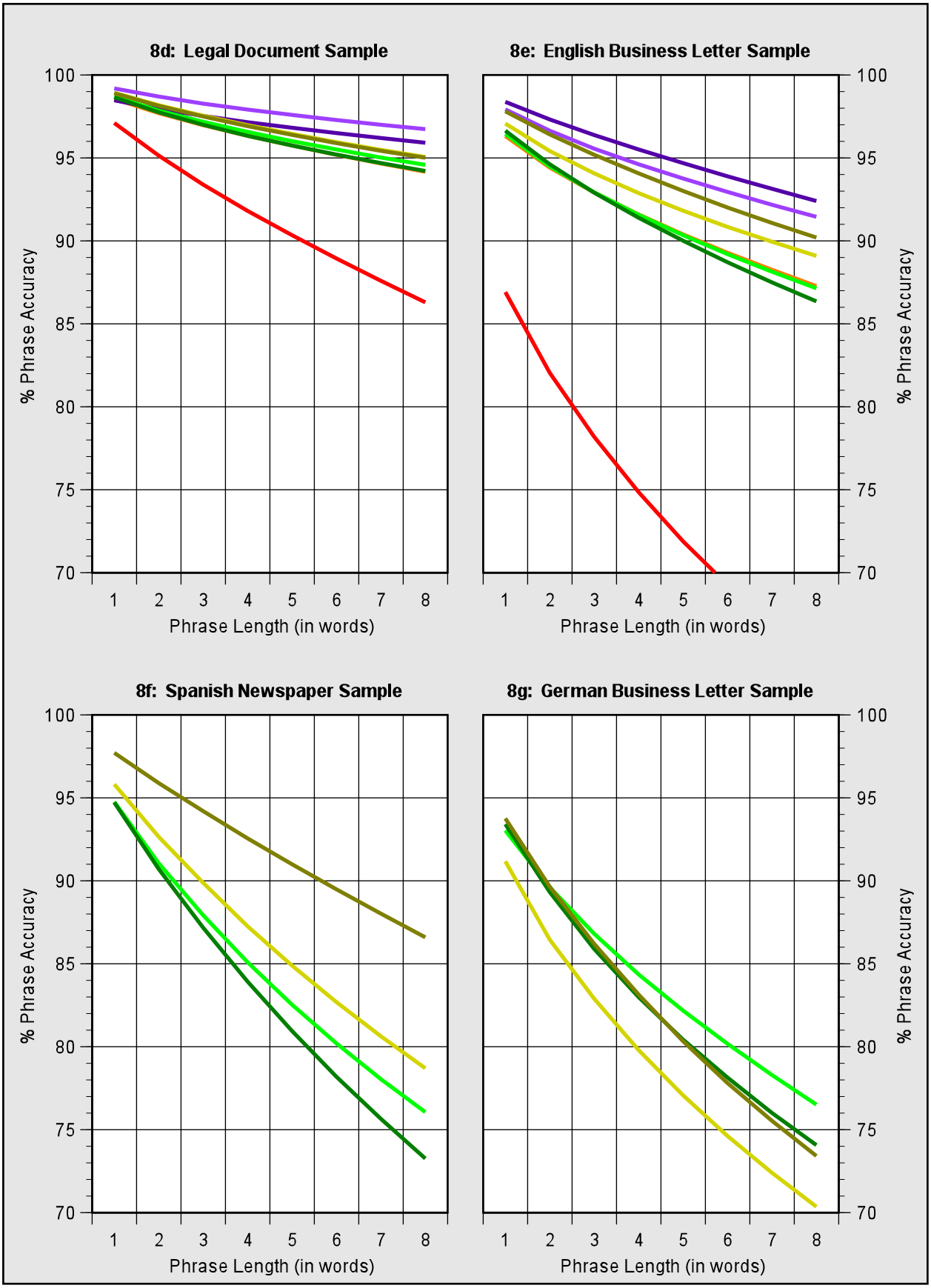
7 Non-stopword Accuracy



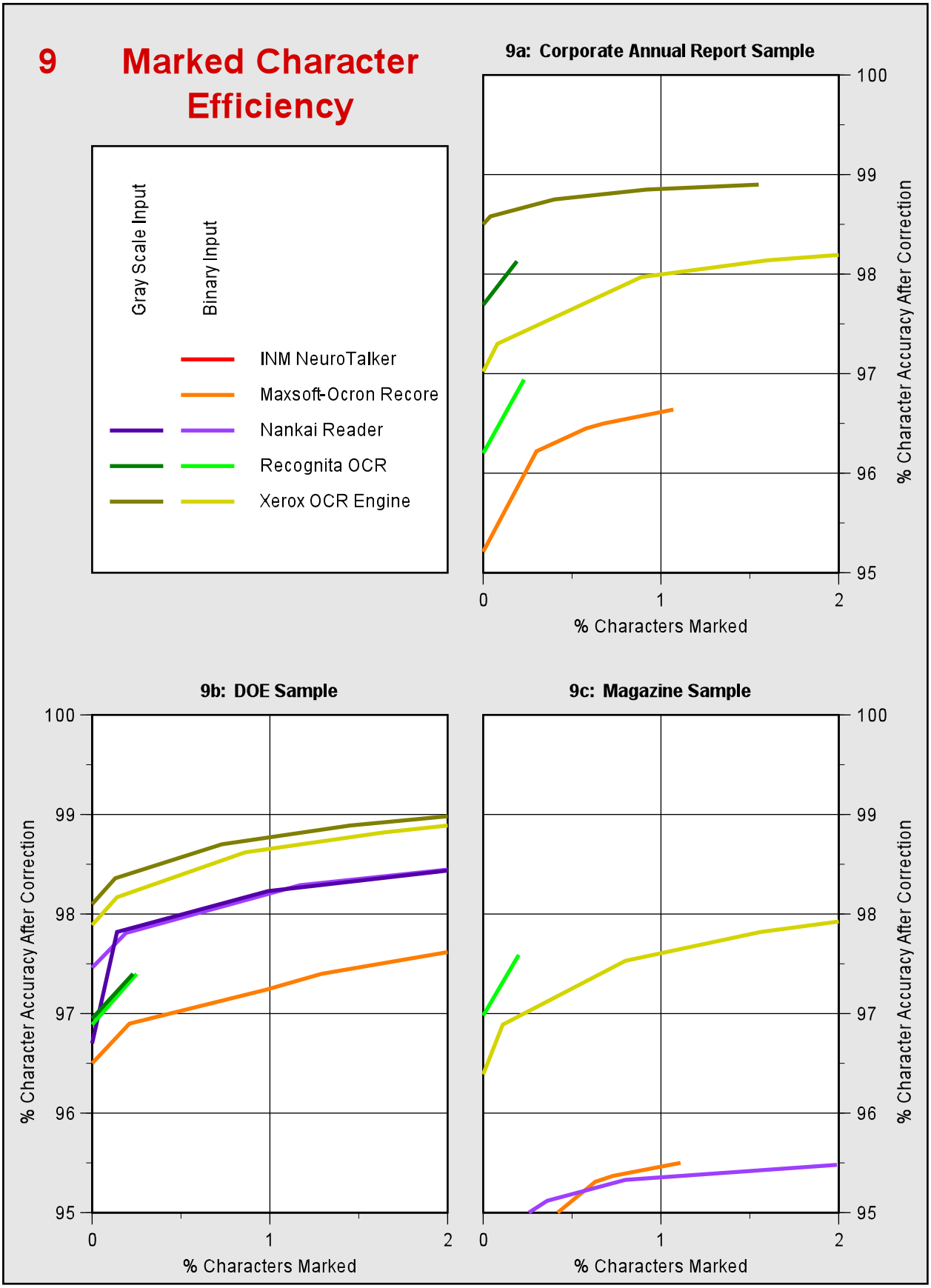

7d: Legal Document Sample
7e: English Business Letter Sample
7f: Spanish Newspaper Sample
7g: German Business Letter Sample
% Non-stopword Accuracy
Number of Stopwords
100
95
90
85
0
50
100
150
200

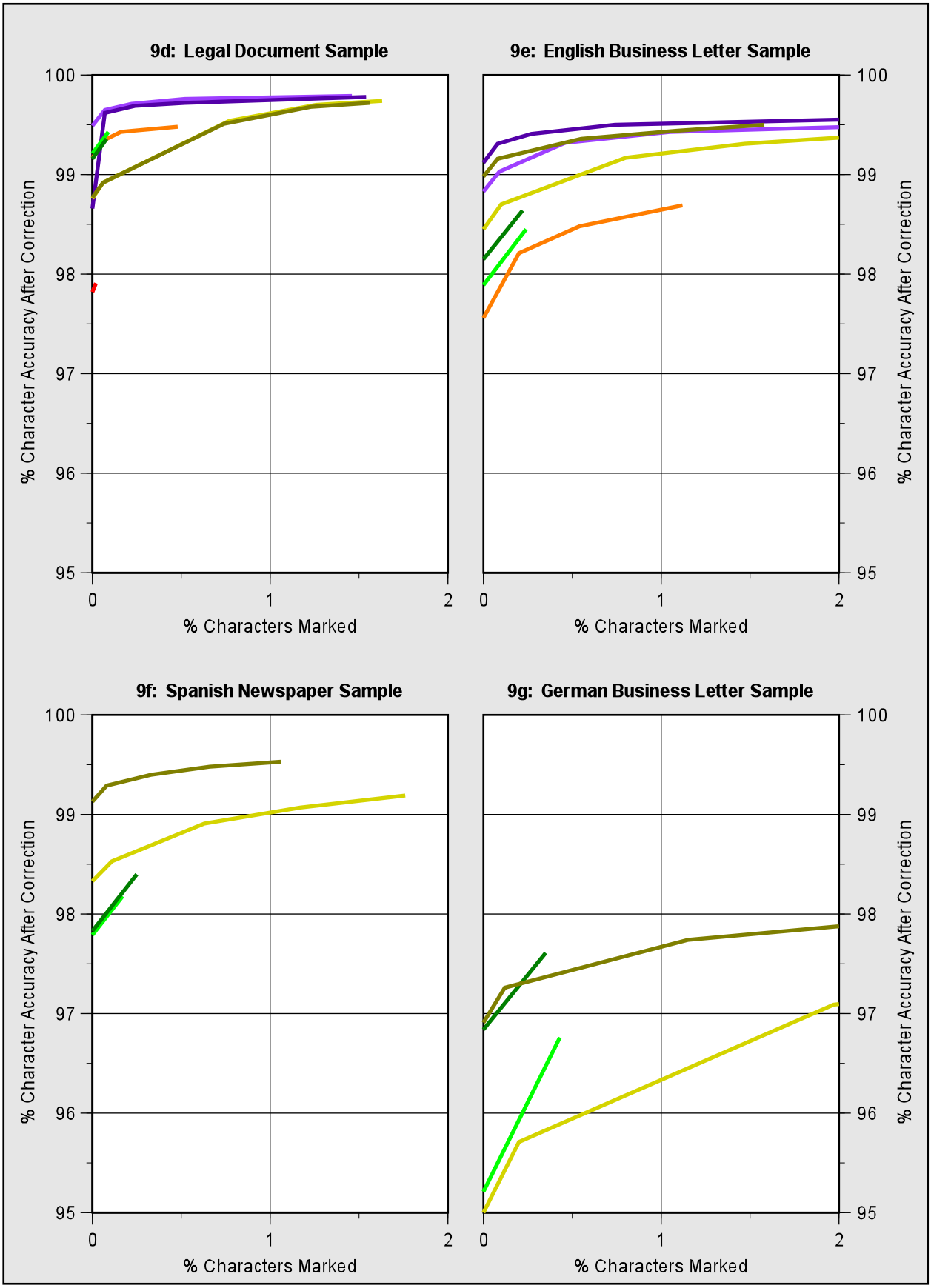
8 Phrase Accuracy



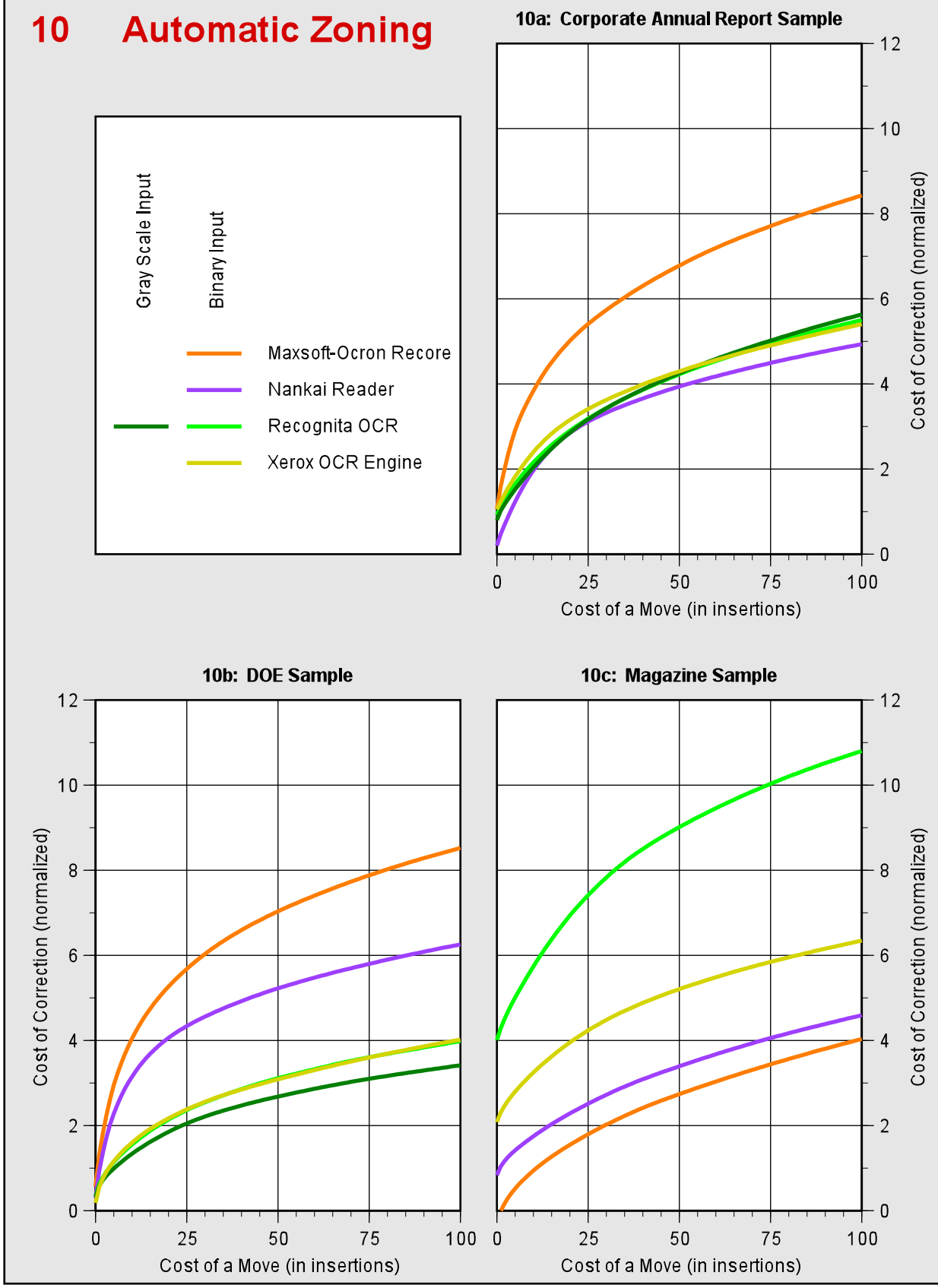

8d: Legal Document Sample
8e: English Business Letter Sample
8f: Spanish Newspaper Sample
8g: German Business Letter Sample
% Phrase Accuracy
Phrase Length (in words)
100
95
90
85
80
75
70
1
2
3
4
5
6
7
8

9 Marked Character Efficiency




9d: Legal Document Sample
9e: English Business Letter Sample
9f: Spanish Newspaper Sample
9g: German Business Letter Sample
% Character Accuracy After Correction
% Characters Marked
100
99
98
97
96
95
0
1
2

10 Automatic Zoning



About ISRI

ISRI was established in 1990 with funding from the U.S. Department of Energy. Its mission is to foster the improvement of automated technologies for understanding machine-printed documents. To pursue this goal, four specific programs have been established:

1. ISRI conducts a program of applied research in recognition of information from machine-printed documents. Our research focuses on developing new metrics of recognition performance, on measures of print quality, on document image enhancement, and on characterization of document analysis techniques.

2. ISRI conducts a program of applied research in Information Retrieval. This research is focused on issues related to the combined use of recognition and retrieval technologies. For example, we are focused on evaluating the effectiveness of different retrieval models in the presence of OCR errors. We are interested in improvements that can be made in the retrieval environment to reduce the effects that recognition errors have on retrieval. Further we are developing systems to automatically tag the physical and logical structure of documents to establish a mapping between the text and the image. This mapping can be exploited in various ways to improve both retrieval and display of documents.

3. Each year, ISRI sponsors a "Symposium on Document Analysis and Information Retrieval" (SDAIR). This symposium provides a forum for presenting the results of research into improved technologies for document understanding with emphasis on both recognition and retrieval from machine-printed documents.

4. ISRI conducts an annual "OCR Technology Assessment" program. Each year, using its automated test facilities, ISRI prepares an in-depth, independent comparison of the performance characteristics of all available technologies for character recognition from machine-printed documents. The results of this test are first made public at the SDAIR symposium.

These programs interact very strongly. We expect that continued development of new measures of OCR system performance will contribute to a better understanding of recognition problems. Our Technology Assessment program provides an opportunity each year to apply new metrics. Metrics, such as non-stopword and phrase accuracy, reflect on our ability to retrieve information. Our view is that new measures of recognition technologies are needed and that goal-directed measures may be the most important. Finally, SDAIR is a natural forum not only for presenting and discussing detailed test results but also for stimulating interaction between recognition and retrieval researchers. Our goals are to promote improved understanding of the current state-of-the-art in both recognition and retrieval and to promote the exchange of information among the user, vendor, and academic communities.