

第四次OCR准确性年度测试
The Fourth Annual T

斯蒂芬·赖斯，弗兰克·詹金斯和托马斯·纳特克

1引言

四年來，ISRI进行了光学字符识别（OCR）系统的年度测试，该系统被称为“页面阅读器”。这些系统接受任何文档页面的位图图像作为输入，并尝试识别页面上的机器打印字符。在年度测试中，我们通过比较输出的文本和正确的文本来测量此过程的准确性。测试的目标包括：

1. 提供最新，独立的系统性能评估，
2. 逐年衡量技术的进步，
3. 深入了解OCR的复杂性，以及
4. 识别最新技术中的问题。

在过去四年中，测试范围已大大增加。在第一个测试[Rice 92]中，六个OCR系统处理了132页的二进制图像，总共包含278,000个字符。这些页面是从美国能源部（DOE）的科学和技术文档数据库中随机选择的。在第二年，采用了新的性能指标来评估使用较大的DOE样本（460页和817,000个字符）的8个OCR系统[Rice 93a, Kanai 93, Nartker 94a]。第三次年度测试重用了该DOE样本，并精选了200页的美国流行杂志文章的样本，这些文章以三种不同的分辨率进行了扫描。因此，在包含将近150万个字符的页面上测试了六个OCR系统[Rice 94, Nartker 94b]。在此报告中，我们介绍了第四次年度测试的结果，这是迄今为止规模最大，最全面的测试。测试样本包含来自商务信函，DOE文档以及杂志和报纸文章的三百万个字符。每页已扫描四次，以生成三种不同分辨率的二进制图像，再加上一张灰度图像。此外，已经为每个商务信函页面获得了两种不同分辨率的传真图像。我们介绍了我们的第一个非英语样本，该样本是西班牙语报纸文章的集合，并且首次报告了OCR系统的速度

1.1参加者

任何组织均可参加提供的年度测试：

- 1.在规定的截止日期（1994年12月15日，第四次年度测试）之前提交OCR系统的版本，
- 2.该版本在PC或Sun SPARCstation上运行，并且
- 3.该版本可以以全自动（非交互）方式处理TIFF图像的特定区域。

此外，每个组织只允许输入一个条目。此测试中评估了OCR系统的许多功能。提交的版本不需要支持所有这些功能。例如，如果某个版本不支持自动分区或西班牙语OCR，则将其直接排除在测试的那部分之外。表1列出了参加本年度测试的八个组织及其提交的版本。惠普实验室提交了仅可在HP工作站上运行的研究原型。之所以可以这样做，是因为HP在截止日期之前就提供了硬件并简化了界面。

1.2测试数据

今年的测试中使用了五个测试样本。

- 1.商业信函样本包含企业和个人收到并捐赠给ISRI的各种信函。
2. DOE样本是我们通过从DOE的科学和技术文档集合中随机选择页面而准备的第三大样本。
- 3.在第三次年度测试中使用的杂志样本，由从发行量最大的100本美国杂志中随机选择的页面组成。
- 4.英文报纸样本包含从发行量最大的50份美国报纸中随机选择的文章。
- 5.西班牙报纸样本包含从阿根廷，墨西哥和西班牙的12种流行新闻中随机选择的文章。

对于报纸样本，仅选择报纸第一部分的文章，然后从报纸上剪下每篇文章。将每个测试页手动放置在Fujits u M3096G扫描仪的压板上，然后进行四次数字化处理，以生成二进制图像。每英寸200、300和400点（dpi），以及300 dpi的8位灰度图像。全局阈值127（255个阈值）用于创建商务信函，DOE和杂志样本的二进制图像。为报纸样本选择了不同的阈值：英文文章为75，西班牙文文章为95

Table 1: Participating Organizations

<u>Organization</u>	<u>Version Name</u>	<u>Version No.</u>	<u>Platform</u>	<u>Version Type</u>
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

	<u>Pages</u>	<u>Zones</u>	<u>Words</u>	<u>Characters</u>
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

我们通过使用Xerox 7024传真机到标准传真机和本地模式在本地传输每一页，来创建商务信函样本的传真图像。标准模式传真图像的X方向分辨率为204 dpi，Y方向的分辨率为98 dpi；精细模式图像的X方向的分辨率为204 dpi，但Y-方向的分辨率为196 dpi。我们手动对每个页面进行“分区”，即，我们划定了页面的文本区域并对其进行了排序。OCR系统仅处理这些“区域”。一些文本被认为不适合该测试，因此被排除在外；示例包括等式，广告，作为表象的一部分的文本（例如，图形或地图的标签）以及被人类认为不可读的文本。我们精心准备了与之对应的正确文本或“真实性”每个区域。为了确保最高的准确性，每个区域的文本由独立工作的打字员输入了四次。借助差异算法对这四个版本进行了协调。表2列出了每个测试样本中的页面，区域，单词和字符的数量。

1.3 测试操作

OCR实验环境的5.0版用于进行第四次年度测试。这是ISRI开发的一套软件工具，用于OCR的大规模，自动化测试和实验研究。[Rice 93b]中描述了该软件的早期版本。该软件在Sun SPARCstations上运行，并提供对PC的远程控制。每个OCR系统都以全自动方式操作，即无需人工干预。OCR生成的文本与正确文本的比较，以及准确性统计信息的列表，都是在计算机的控制下进行的。没有为每个OCR系统重新扫描页面，也没有对页面图像进行重新分区；所有OCR系统都处理相同页面图像的相同分区部分。除非另有说明，否则使用300 dpi二进制和灰度图像进行测试。例外情况是涉及传真商务信函的测试和分辨率效果的测试（还使用了200和400 dpi二进制图像）。CaereOCR和XIS OCR引擎在单处理器Sun SPARCstation 10上以SunOS 4.1.3进行操作。具有64 MB的内存。五个基于PC的OCR系统在具有8 MB内存的相同配置的486DX / 33计算机上执行，在MS-DOS 5.0下运行，对于其中三个，运行Windows Windows 3.1。HP Labs OCR在HP 9000 Model 735上以32 MB的内存运行在HP-UX A.09.01下运行，记录定时数据时每台机器都没有负担，即只有OCR系统在机器上运行。每个OCR系统每次都处理一页图像。因此，时序图包括为每个图像初始化OCR系统的适度开销。

2个字符精度

尽管有许多方法可以量化OCR生成的文本与正确文本之间的偏差，但在我们最基本的衡量标准中，我们反映了人工编辑人员需要进行的工作以校正OCR生成的文本。具体来说，我们计算编辑操作的最小数量（字符

插入，删除和替换）以完全更正文本。我们将此数量称为OCR系统产生的错误数量。将其表示为字符总数的百分比，可以得到字符精度：

$$\# \text{errors} \times 100\% = \frac{\# \text{errors}}{\# \text{characters}} \times 100\%$$

过去，我们使用的算法往往将最低的编辑操作数量高估5至10%。为了精确报告，今年我们已切换到一种可以精确计算最小数字的算法。它是Ukkonen [Ukkonen 85]的一种优化算法。表3a-3f给出了每个测试样本的字符精度结果。由于不支持的功能，有些条目丢失了。仅Caere OCR和HP Labs OCR接受灰度输入。EDT ImageReader，HP Labs OCR和INM NeuroTalker不支持西班牙语OCR。

2.1失败

当OCR系统在处理页面图像时“崩溃”或“挂起”时，或者在终止时返回错误状态时，将检测到故障。在“失败”列中输入none表示未检测到任何失败。否则，将指定失败的页面数，然后指定这些页面上的字符数，以样本中字符总数的百分比表示。如果后者超过百分之一，则认为失败是过分的，并且不报告准确性结果；否则，将按照与失败页面上的字符数相等的数量收取错误费用。错误可能导致输出多余的字符，或者可能阻止生成正确的字符。无法自动可靠地将此类故障与识别错误区分开。它们不会被检测到，并且错误会与纠正损坏所需的编辑工作成比例地收取。如果OCR系统的字符精度对于特定样本而言小于90%，则我们仅注意到精度低于此阈值。我们不会在此样本上进一步报告该系统的性能。

2.2置信区间

图1a-1g显示了大约95%的置信区间，以提高字符准确性。这些间隔是使用称为折刀估计器[Dudewicz 88]的统计技术计算出来的。在应用此技术时，我们假设样本中的页面是独立的，但我们并未假设页面中的字符一个独立的OCR系统在一个样本中始终如一地表现为一个狭窄的时间间隔，而一个宽的时间间隔则表明了相当大的可变性。当比较两个系统的性能时，非重叠间隔意味着两个系统之间存在统计学上的显着差异。

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速度和吞吐量

对于大多数应用而言，速度远不如准确性重要。的确，快速OCR系统几乎没有用，它产生的输出大部分都是乱码。但是，鉴于OCR系统具有可比的精度，因此速度成为一个重要因素。我们通常反对在不考虑精度的情况下报告原始速度数据，因此，我们引入了以下吞吐量功能，该功能在报告速度的同时对错误进行了惩罚：

$$\frac{\# \text{个字符} P \# \text{个错误}}{\# \text{秒}}$$

P代表分配给每个错误的代价。当P =

0时，该函数给出每秒字符的原始速度项。一些作者将吞吐量定义为表示“正确的字符”/秒，相当于P = 1。因此，在图2a-2g中，我们给出了P = 0到10的吞吐量。

2.4按字符分类的准确性

在按字符分类的准确性方面，我们将基本字符分类，并确定每个分类中正确识别的字符的百分比。使用了以下类。

1. 间距：空白和行尾字符，
2. a-z：小写字母，
3. A-Z：大写字母，
4. 0-9：十进制数字，以及
5. 特殊：标点符号和其他特殊符号。

对于西班牙报纸样本，添加了一个西班牙语类，其中包含西班牙语的重音符号和标点符号。图3a-3g显示了每个测试样品的结果。这些类别中最大的类别是a-z类别；根据样本的不同，有68%到75%的地面真实字符属于此类。第二大类是Spacing类，占字符的16%至17%。A-Z, 0-9和Special类包含3-7%，1-6%和3-5%的字符，分别。西班牙语类包含“西班牙报纸样本”中2%的字符。在这些较小的类别上，OCR系统的准确性较差。

2.5分辨率的影响

图4a-4e显示了如何通过将二进制图像的分辨率从300 dpi降低到200 dpi，以及将其提高到400 dpi来影响字符精度。图4a还包括传真图像。如果缺少特定分辨率的数据点，则说明OCR系统难以处理以该分辨率扫描的图像：出现过大的故障或者其准确性低于90%。将分辨率从300 dpi降低到200 dpi会导致分辨率的大幅提高。的数量

错误：商务信函和DOE样本大约增加了50%，杂志和英文报纸样本大约增加了75%。西班牙报纸样本的错误数量跃升了200%，人们希望通过将分辨率从300 dpi增加到400 dpi来减少错误数量。在某些情况下，它的数量很少，但错误的数量却经常增加。较高的分辨率几乎没有优势，甚至没有优势。精细模式传真图像的分辨率与200 dpi图像基本相同。但是，在处理前者时，OCR系统所犯的错误要比后者少5%至15%。通过检查，我们发现，与使用Fujitsu扫描仪生成的图像相比，传真机创建的图像更“暗”，并且包含更少的破字符。标准模式传真图像对于OCR系统而言是非常困难的测试。在这些图像上所犯错误的数量是在精细模式图像上所犯错误数量的两倍以上，只有Caere OCR和XIS OCR Engine这两个系统的字符精度达到了90%以上。Ligature和Recognita这两个组织选择不参加此测试。

2.6 页面质量组

如果我们使用多个OCR系统处理给定页面，并在此页面上确定每个系统的字符精度，那么我们可以计算这些精度的中位数以获得该页面的质量或“OCR难度”的度量值。我们使用这种方法将每个样本的页面划分为五个大小大致相等的“页面质量组”。第1组包含中位数准确度最高（最佳页面质量）的页面，第5组包含中位数准确度最低（最差页面质量）的页面。在图5a-5g中，绘制了每个组中的字符准确度以显示页面质量的影响，其中大部分错误（约50%至60%）是在每个样本中最差的20%（即第5组）上产生的。对于DOE样本，此百分比甚至更高，大约为70%到80%。通过检查属于第5组的页面的图像，我们可以了解导致OCR困难的原因。在图1-5中，我们提供了从这些图像中摘录的摘录。每个字符都是从300 dpi的二进制图像中复制出来的，并放大了50%，以便于查看。断线和触摸字符（也称为“拆分”和“joins”）是非常常见的错误来源，在每次测试中都会发生样品。在处理商务信函时，OCR系统在读取信头时会遇到一些困难，这些信头通常以样式打印。同样，印刷版中的折痕也会影响整行的识别。DOE样本包含许多具有挑战性的表格，有些页面带有倾斜的文本和/或弯曲的基线。在阴影背景上打印的文本在杂志文章中很常见，并且是造成错误的重要原因。新闻纸的渗漏和其他不规则现象引起了报纸图像的斑点。

3个字的准确度

OCR的一种流行应用是从一系列硬拷贝文档中构建文本数据库。然后可以将信息检索技术应用于查找感兴趣的文档。在这种环境中，正确识别的单词的百分比或

Word Accuracy

Table 4a: Original Business Letters

	Misrec. Words	300 dpi Binary			300 dpi 8-bit Gray Scale			Non- stopword Accuracy
		Word Accuracy	Stopword Accuracy	Non- stopword Accuracy	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

	Misrec. Words	Standard-mode Fax			Fine-mode Fax			Non- stopword Accuracy
		Word Accuracy	Stopword Accuracy	Non- stopword Accuracy	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

	Misrec. Words	300 dpi Binary			300 dpi 8-bit Gray Scale			Non- stopword Accuracy
		Word Accuracy	Stopword Accuracy	Non- stopword Accuracy	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
Ligature CharacterEyes Pro	5,785
MAXSOFT-OCRON Recore	5,015
Recognita OCR	3,400
XIS OCR Engine	2,966

OCR生成的文本是一项重要措施。我们将单词定义为一个或多个字母的任意序列。如果单词的所有字母均已正确识别，则认为该单词被正确识别。由于全文搜索通常是在不区分大小写的基础上执行的，因此在错误的情况下生成的字母（例如，C代表c）仍然被认为是正确的。表4a-4f给出了错误识别的单词数和该单词每个测试样品的准确性。图表6a-6g显示了每个页面质量组中的单词准确性。

3.1停用词和非停用词

停用词是常见的词，例如the, of和in等。由于这些词的检索价值很小，因此通常不会被文本检索系统索引。由于用户仅搜索非停用词，因此特别需要关注的是正确识别的非停用词的百分比或非停用词的准确性。我们使用BASISPLUS文本检索系统默认的110个停用词集[IDI 90]。测试样本中大约40%的单词是停用词。表4a-4e显示了每个测试样本的停用词和非停用词准确性。（这些不是针对西班牙报纸样本计算的。）

3.2不同的非停用词准确性

假设用户希望在数据库中查找包含特定术语（非停用词）的每个文档。如果要查找包含该术语的文档，则OCR系统必须至少已正确识别该术语的一次出现。考虑到这一点，我们引入了一种称为独特的非停用词准确性的测量方法。对于给定的页面，我们将页面上出现的每个唯一术语都称为独特的非停用词，并且说如果至少有一个以下术语正确识别它的出现已被正确识别。不同的非停用词准确性是正确识别的不同非停用词的百分比。有人认为，由于文本固有的冗余性，全文本搜索可抵抗OCR错误。由于搜索词可能在文档中出现多次，因此OCR系统不太可能会误识别每次出现的情况。在独立性的假设下，如果OCR生成的文本的整体非停用词准确性为90%，则对于出现n次的项，每次出现错误的概率为 10^{-n} 。可以提出一个论点，即在每个事件中都可能出现导致OCR系统误识别一个事件的图像缺陷，从而很可能会丢失每个事件。在图7a-7f中，我们显示了至少一个频率发生次数从一增加到四时，可以正确识别发生次数。将这些图与图7g进行对比，图7g说明了在独立性假设下曲线的预期形状。

3.3词组准确度

用户还搜索包含特定短语的文档。我们将长度为n的短语定义为n个单词的任意序列。如果所有词都正确，则该词组将被正确识别。

确定的。短语准确度是正确识别的短语的百分比。图8a-8g显示了长度为1到8的短语的准确性。请注意，长度为1的词组精度等于词的精度。词组精度提供了“错误聚集”的有用度量。给定两个OCR系统，它们具有相同的单词精度，而短语精度较低的系统则产生了错误，这些错误在整个文本中的分布更为广泛。

4标记字符效率

对于用户而言，在OCR生成的文本中查找和纠正错误可能是一个繁琐而昂贵的过程。但是，OCR系统可以通过标记认为最有可能出错的生成字符来提供一些帮助。当OCR系统无法识别字符时，将在输出中放置一个拒绝字符（ \sim ）。另外，系统可能会在每个由低置信度生成的字符之前放置一个可疑标记（ \wedge ）。我们将拒绝字符和标记为可疑的字符称为标记字符。

这种感觉污染会拒绝特征和可疑标记。

在上面的句子中，有三个明显的错误：两个拒绝字符和“包含”中的“l”。 “标记”中的第二个“a”是未标记的错误。“and”中的“d”是错误标记，是正确生成的字符，被标记为可疑字符。为了提高标记字符的效率，我们测量了OCR系统生成的标记字符的效用。图9a-9c显示的曲线显示了随着人工编辑者检查越来越多的标记字符并纠正标记错误，OCR生成的文本的字符精度如何提高。最初，此过程非常有效，因为编辑者可以更正由拒绝字符和第一级可疑标记识别的错误。但是随着可疑标记数量的增加，由于伪标记百分比的增加，曲线变得平坦。标记的字符使编辑者可以只检查OCR生成文本的百分之一的一半，而可以纠正20%到45%的文本。文字错误。编辑者可能检查的字符更多，但是操作效率大大降低。

5自动分区

到目前为止，在讨论的每个测试中，都为OCR系统提供了要处理的文本区域的坐标。在自动分区测试中，未提供此信息，并且要求每个系统定位文本区域，并确定其正确的阅读顺序。为了衡量此任务的执行效果，我们应用了一种算法来估算字符插入的数量和纠正自动分区错误所需的块移动操作。如果OCR系统找不到文本区域，则需要插入以输入缺少的文本。如果文本块乱序，则需要移动操作来重新排序它们。使用转换因子以相等的插入次数来表示每个移动操作，最终仅以插入为单位给出校正成本。[Kanai 95]中介绍了此方法的详细信息。

图10a-10c显示了该测试的结果。针对一系列转换因子绘制了校正成本，并已使用每个样本中的字符数进行了标准化，以允许在各个样本之间进行比较。 HP Labs OCR和INM

NeuroTalker不支持自动分区，因此在图中没有显示。其他曲线丢失是由于过度的故障。 XIS OCR引擎在该测试中的整体表现最佳，而英文报纸样本的校正成本最低。由于每篇文章都是从报纸上剪下来的，因此OCR系统的争用主要是在一个或多个文字栏的上方有一个标题。但是杂志的文章没有被剪裁，而是页面布局的一部分，可能非常有创意（即很复杂）。 DOE页面提出了将表与多列文本区分开的挑战。

OCR系统应该“取消”列队，而不是“列队”。

6 OCR系统的比较：准确性和速度

6.1二进制输入

Caere OCR, HP Labs OCR和theXIS OCR Engine实现了二进制图像上最佳的整体精度。我们观察到这些系统之间的准确性没有显着差异，但有一个例外：在标准调制解调器传真图像上，HP Labs OCR的性能优于其他两个系统。但是速度上的明显差异是显而易见的。 XIS OCR引擎比Caere OCR快2.3到4.4倍，具体取决于样品。尽管Caere OCR和HP LabsOCR在不同的平台上运行，但显然前者比后者快得多。在提交HP Labs OCR进行此测试时，HP Labs代表表示该版本尚未优化，并且在速度方面没有竞争力.MAXSOFT-OCRON Recore和Recognita OCR构成了第二层系统。该系统在准确性方面相当，Recognita OCR的运行速度几乎是后者的两倍。第三层系统包括EDT ImageReader, INM NeuroTalker和LigatureCharacterEyes Pro。INM NeuroTalker的一个有趣功能是可以调整速度与精度之间的权衡。应INM的要求，我们开始使用最高精度（和最低速度）的设置对该系统进行测试。但是在遇到太多无法使系统保留在测试中的故障后，应INM的要求，我们将设置更改为降低精度，提高速度并避免故障的设置。

（INM将故障归因于系统附带的有故障的DOS扩展器。）由于在第三次年度测试中使用了Magazine Sample，因此它可以作为衡量过去一年进度的准绳。四个组织参加了第三和第四次年度测试：Caere, EDT, Recognita和XIS。在今年的测试中，来自Caere和XIS的系统在此样本上的错误比一年前的错误少27%至28%。TheRecognita版本今年的错误比去年增加了6%。由于故障，EDT版本无法进行比较

6.2灰阶输入

在处理灰度图像时，Caere OCR和HP Labs OCR使用根本不同的方法。HP Labs OCR首先将图像“二值化”，即从灰度图像创建二进制图像，然后识别二进制图像上的字符。另一方面，Caere OCR可以直接从灰度图像中识别字符，很明显，灰度输入在准确性方面优于二进制输入，根据样本，这些系统在给定灰度时产生的错误减少了10%至40%输入。但是有一个例外：Caere OCR在处理“英语报纸样本”的灰度图像时犯了更多的错误。这可能是由于渗漏引起的，渗漏通常在报纸的灰度图像中可见，但在二值化过程之后通常会消失。看来，灰度输入在识别阴影背景上打印的文本时具有最大价值，通常在杂志样本。此文本的二进制图像通常有斑点且有问题，但考虑到灰度图像，OCR系统有更好的机会将文本与背景分离。灰度图像比二进制图像需要更多的存储空间和更长的处理时间。平均而言，Caere OCR需要两倍的时间，而HP Labs OCR需要多20%的时间来处理灰度输入。

7小结

八个组织为第四次年度测试提交了OCR系统。这些系统处理了超过1500页商务信函，科学文件以及杂志和报纸文章的位图图像。使用OCR实验环境的分析工具将OCR生成的文本与正确的文本进行比较，并计算性能的几种度量。这些包括字符，单词，非停用词和短语的准确性，以及三个新的度量：吞吐量，按字符类别分类的准确性以及明显的非停用词准确性。此外，还观察到了页面质量的影响，评估了标记字符的效用，并估计了校正自动分区错误的成本。当图像的分辨率从300 dpi降低到200 dpi时，OCR系统的准确性急剧下降；但是，将分辨率提高到400 dpi几乎没有获得任何好处。传真图像，尤其是标准模式的传真图像提出了重大挑战。从灰度输入获得的精度提高表明这是一个重要的新方向。借助置信区间，我们根据二进制输入的精度结果将八个系统划分为三层。然后，我们注意到各层之间速度的差异。这种“排名”反映了这些系统在此测试中的性能。当处理其他类型的文档或在不同的测试条件下处理相似的文档时，它们的相对性能可能会有所不同。最后，我们要强调的是，ISRI不认可任何特定的OCR系统

致谢

我们感谢Ashok Singh博士和George Nagy博士在建立置信区间以提高字符准确性方面的帮助。此外，Junichi Kanai提出了“按字符分类的准确性”。

参考资料

[Dudewicz 88] Edward J. Dudewicz和Satya N. Mishra。现代数学统计，第743-748页。约翰·威利父子 (John Wiley& Sons) , 1988年。

[IDI 90] Information Dimensions, Inc., 俄亥俄州都柏林。1990年6月发行的L版《BASISPLUS数据库管理参考》。

[Kanai 93] Junichi Kanai, Stephen V. Rice和Thomas A.

Nartker。自动分区的初步评估。内华达大学信息科学研究所93-02技术报告，拉斯维加斯，1993年4月。

[Kanai 95] Junichi Kanai, Stephen V. Rice, Thomas A. Nartker和George Nagy。OCR分区的自动评估。IEEE Transactions on Pattern Analysis and Ma-Chine Intelligence 17 (1) : 86-90, 1995年1月。

[Nartker 94a] Thomas A. Nartker, Stephen V. Rice和Junichi Kanai。

OCR准确性：UNLV的第二次年度测试。通知，信息和图像管理协会。8 (1) : 40 +, 1994年1月。

[Nartker 94b] 托马斯·纳特克和斯蒂芬·赖斯。

OCR准确性：UNLV的第三次年度测试。通知，信息和图像管理协会。1994年9月9日，8 (8) : 30+。

[Rice 92] Stephen V. Rice, Junichi Kanai和Thomas A. Nartker。关于OCR设备准确性的报告。技术报告92-02，内华达大学信息科学研究院，拉斯维加斯，1992年3月。

[Rice 93a] Stephen V. Rice, Junichi Kanai和Thomas A.

Nartker。对OCR准确性的评估。内华达大学信息科学研究所93-01技术报告，拉斯维加斯，1993年4月。

[Rice 93b] Stephen V. 白饭。

OCR实验环境，版本3。技术报告93-04，内华达大学信息科学研究所，拉斯维加斯，1993年4月。

[Rice 94] Stephen V. Rice, Junichi Kanai和Thomas A. Nartker。

OCR准确性的第三次年度测试。内华达大学信息科学研究所94-03技术报告，拉斯维加斯，1994年4月。

[Ukkonen 85] Esko Ukkonen。近似字符串匹配的算法。信息与控制64: 100-118, 1985

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

**EMBASSY OF
COMMERCIAL
CHICAG**

quarter page adver
your support again
your convenience.

organization receiving i
expert information on t
research and maintains :

Please forward them to
returned to us together

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

Incentive Award - A mid
FAXed images with a sys

P.S. This enrollment
by June 15, 1993

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

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

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

UNITED STATES DEP.
National Institute of Sta
Gaithersburg, Maryland 20889

I would highly rec
organization. Sh

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

Service Representatives: sin
We have enclosed a C

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

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

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

Operation of pressure transducer tensiometers in an infil-

9. Precision

9.1 Criteria for judging the a
the maximum density and optim

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 to
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

would bring about dehydrati
acid sites (i.e. surface alomini
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 source"

consideration for the
heterogeneities and 2
thermally induced sa

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

ter. Still, much remains mysterious: Who? cent? The patient had no French ancestors. ventured from his hometown of Worcester,

enough to justify preventive treatments. Even if these medications made sense in my case, though, I'd b

achieved the same results, but how do they do it?

Answer: In years past, the s were fit with a considerable amount of

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

GALLONS USED DAILY BY A FAMILY

the cheapest plastic lenses with anti-reflective coating. The prices they were quoted were up to 50 percent less than by as much as 75 percent from

Summer's fresh barbecue sauces and marinades

PARKS

Children under 3 admitted free. * Prices do not include tax.

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-think session, I heard a story that I have never forgotten.

on scientific research, said Dr. Eric McDuffie, former senior vice president for medical affairs for the Leukemia Foundation.

Opry, Garth Brooks met backstage with cancer patient Libby Sharp of Gatlin-

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

WICKENBURY, ARIZONA

The Meadows

A leading treatment center for

Taste buds take note. *Tray Gourley's Your Own Chef in the College Classroom* (Lake Isle Press, \$10.95) is the book

cooked on the stovetop in your kitchen comfortably coo

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

TALK OF THE DAY:

"Names change for political, 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 he will be indicted on embezzlement charges, signaling a glitch in

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

them because of who he wants. The distinctive marks turn off prospective employers, threatening the life he plans away from the state.

Washington Post

THIMPU, Bhutan — Its citizens seldom write letters and there

and conversations with world leader:

neighbors and then leading them up the hill toward a new life.

Federal and state officials say him a lot.

cholera in the crammed Rwandan camps of eastern Zaire is going to United Nations said yesterday. But as the threat from cholera

heretofore unlabeled wines: Chateau Margaux; Chateau Lafite; Robert Mondavi cabernet sauvignon

Belize, a small Central American country, use plant medicines, losses of forests, and a lack of interest among young people in becoming

The village must rely on the McHenry County Sheriff's Department in Woodstock to enforce the limit,

ing a "large number of casual workers," spokesman Maj. Dacre Holloway said. In response, U.N. officials w

unemployment rate decreased to 7.3% of the work force in March from 8.3% in February, but rose from 7.1% in March 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 diente Juan Domingo Perón, que lo

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

PARANA. (Enviado especial). — Hombres con armas y largavistas apostados en las cúpulas de la

El extenso debate por Constituyente, ganad la posición conjunt

Pensiones sociales

La Caja de Pensiones Soles-Ley 5110, delegación Rosa

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

-¿Declararán cuando se llame a indagatorias?

-No. Existe el derecho a callar, eso no tiene nada que ver con la

sospechoso. Desde esferas sociales políticas se consideró que el acusado había sido Guillermo Luque, hijo entonces diputado nacional por

IXTAPALUCA (Notimex).— Enfrentamiento entre un grupo de muneros y agentes de la Dirección General de Seguridad Pública

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

De acuerdo con el plan de reorganización de la compañía, que aún no se da en su totalidad, la intención es diversificar los servicios con la construcción

Presidencia de la República Vicente Fox Zedillo, en materia económica, está siendo difundida por as

BARCELONA. (Agencias).— Los dirigentes de CiU y del PP catalanes cambiaron ayer duras críticas y

Y TIEMPO LIBRE

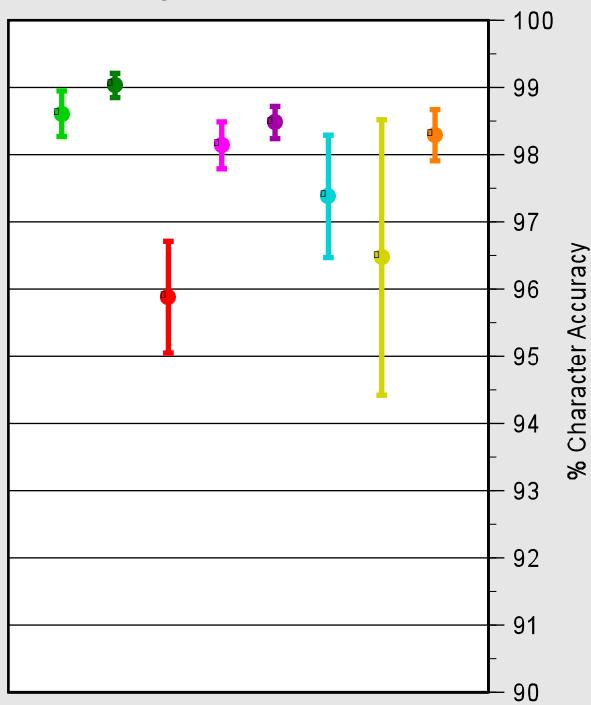
El histórico viaje de Joan Martí d'Empúries

El famoso paseo a pie entre Joan Martí d'Empúries que cantó el poeta Joan Maragall. ¿Por qué se ha quedado en la memoria? Porque Pla es lo que debía ser, un poeta

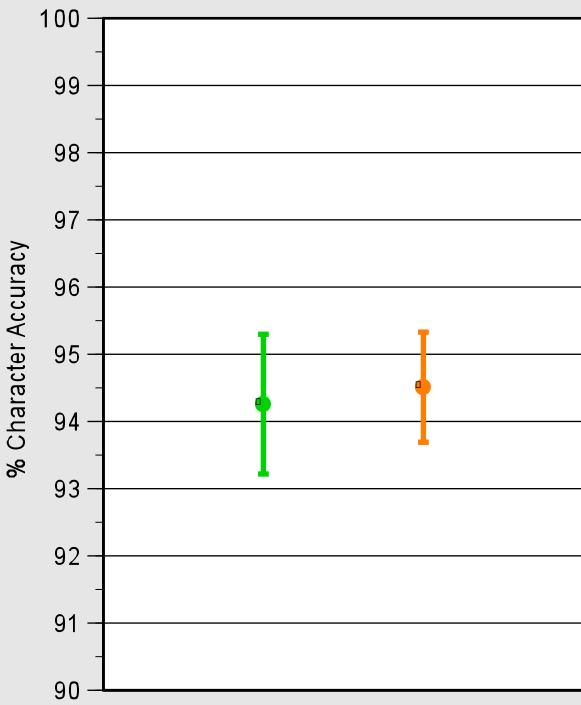
1 Character Accuracy

- 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

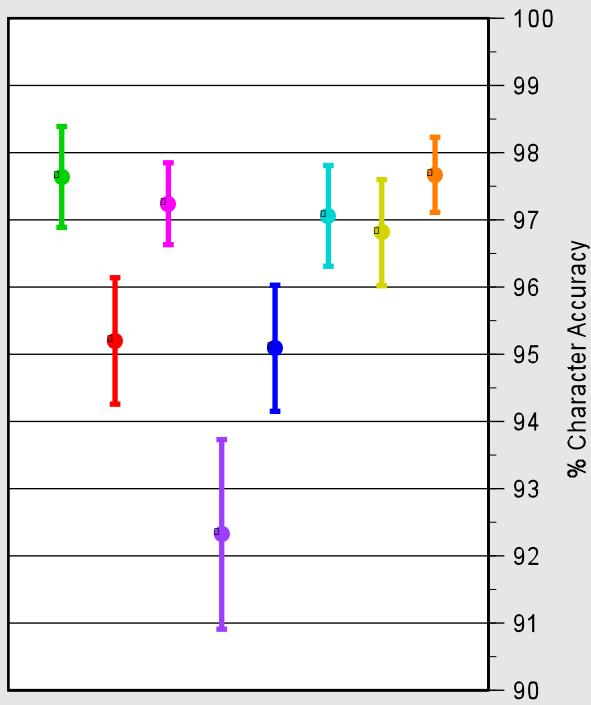
1a: Original Business Letters

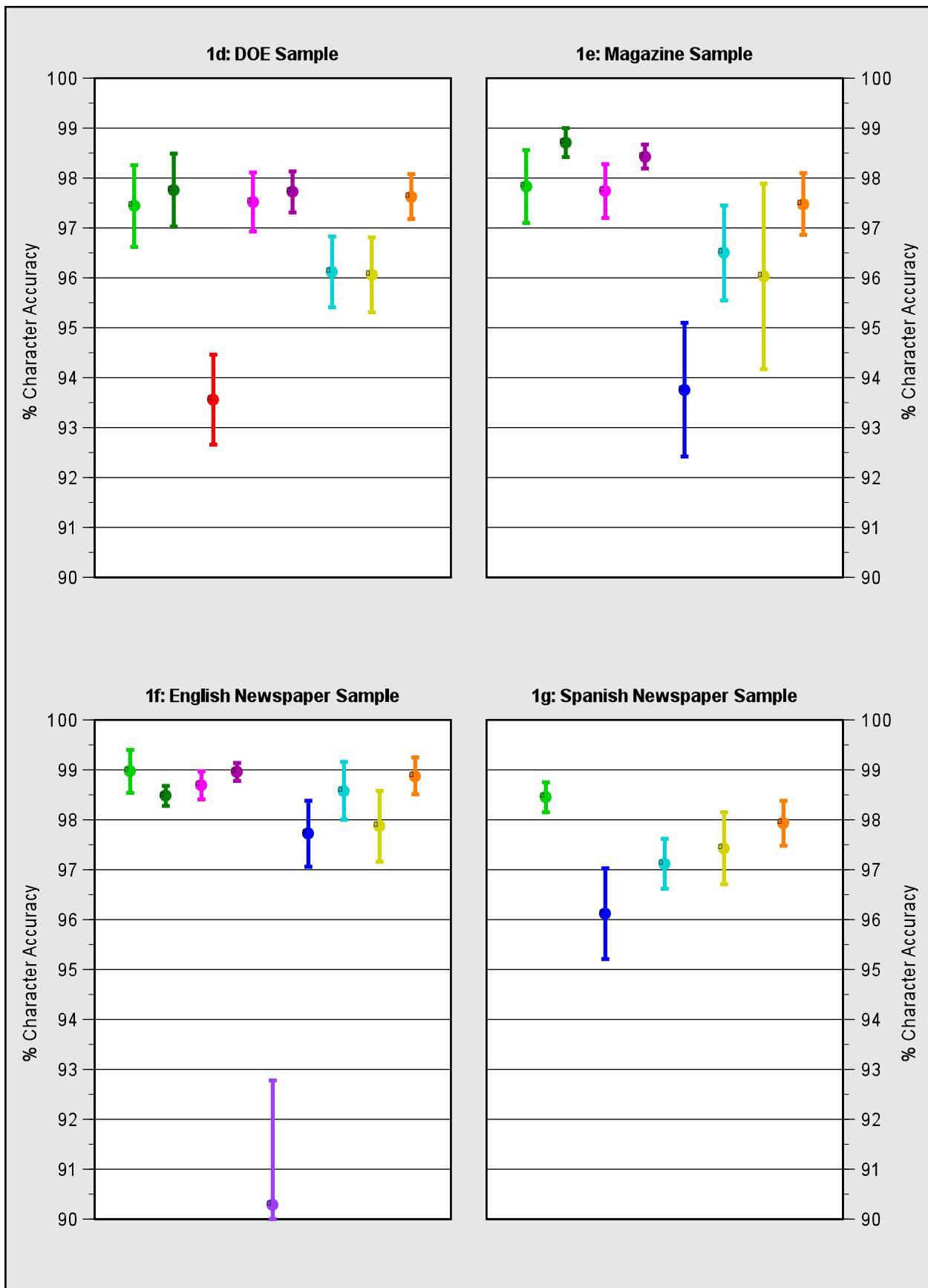


1b: Standard-mode Fax Business Letters



1c: Fine-mode Fax Business Letters

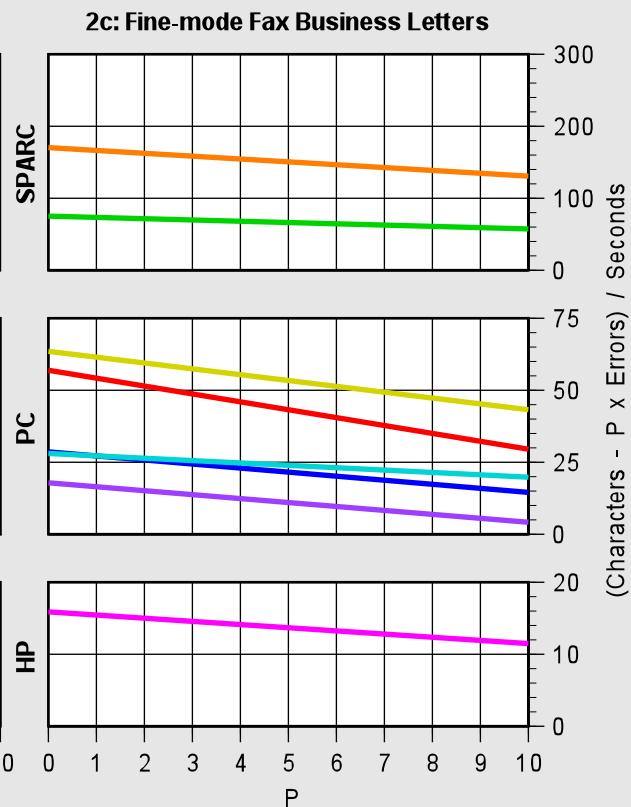
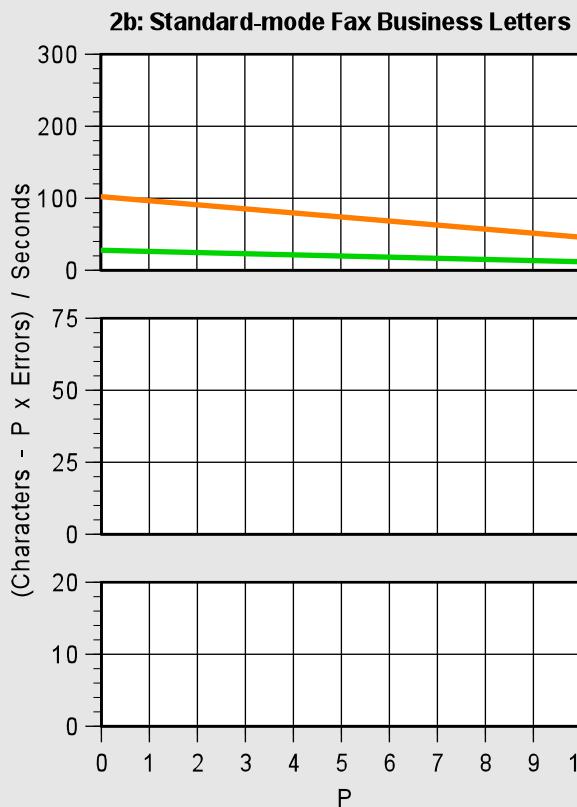
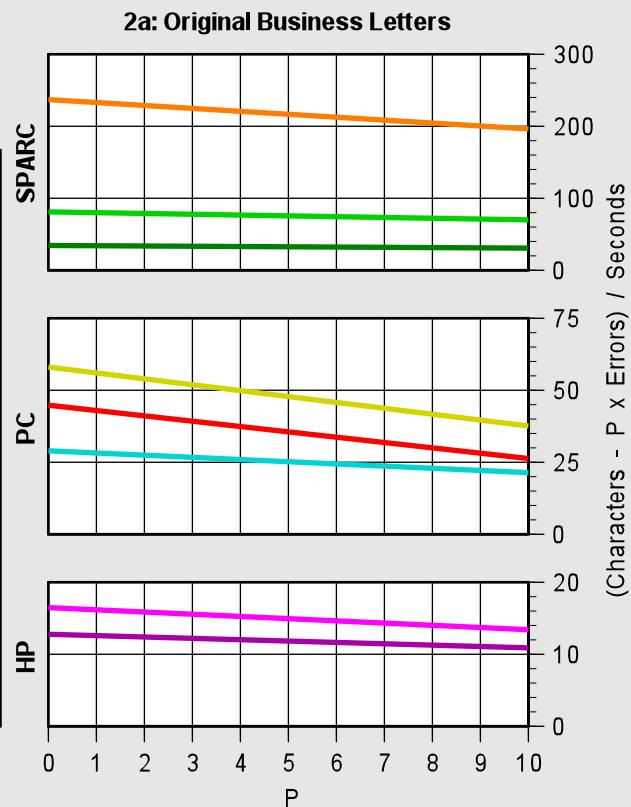


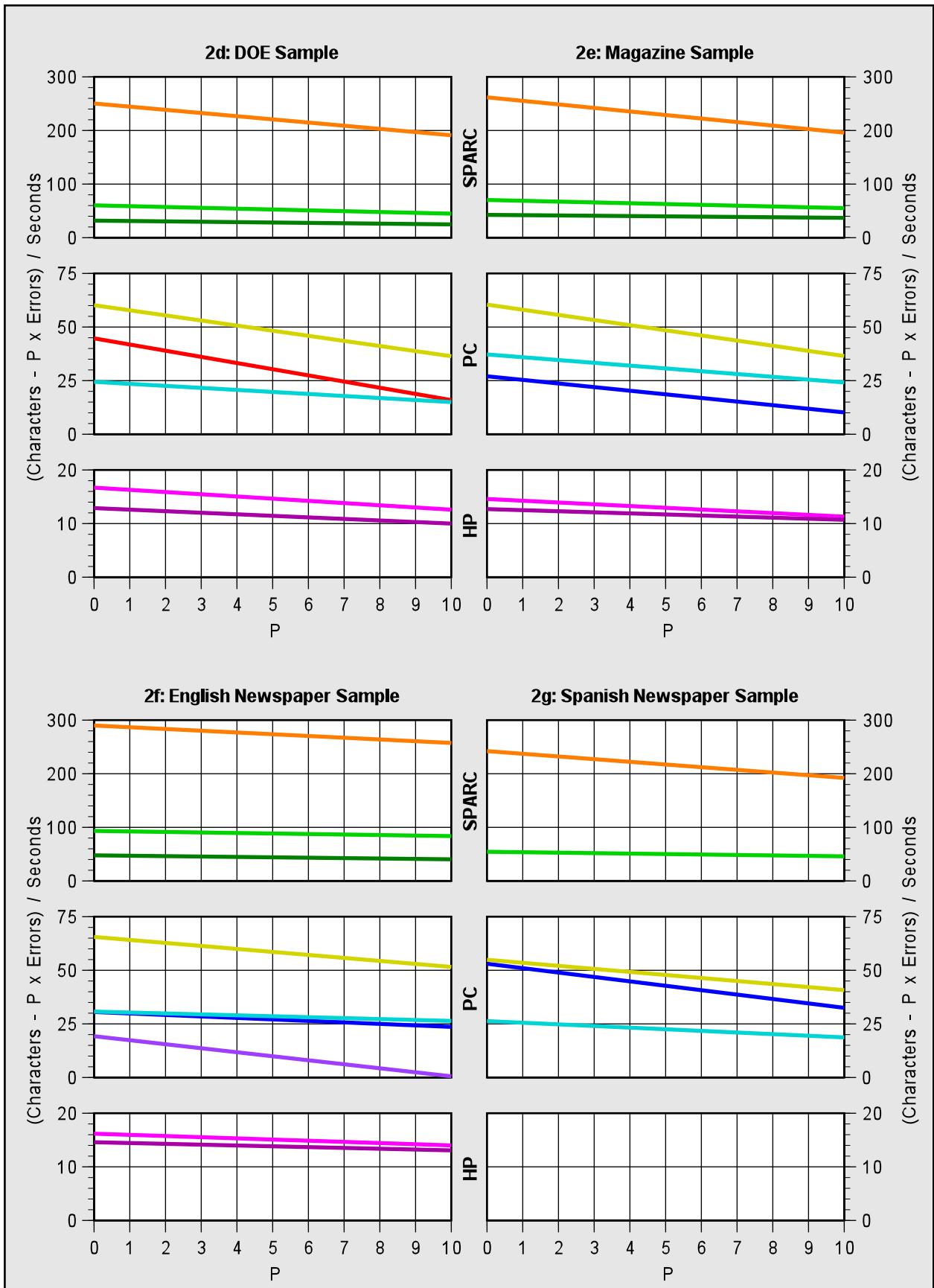


2

Throughput

- 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

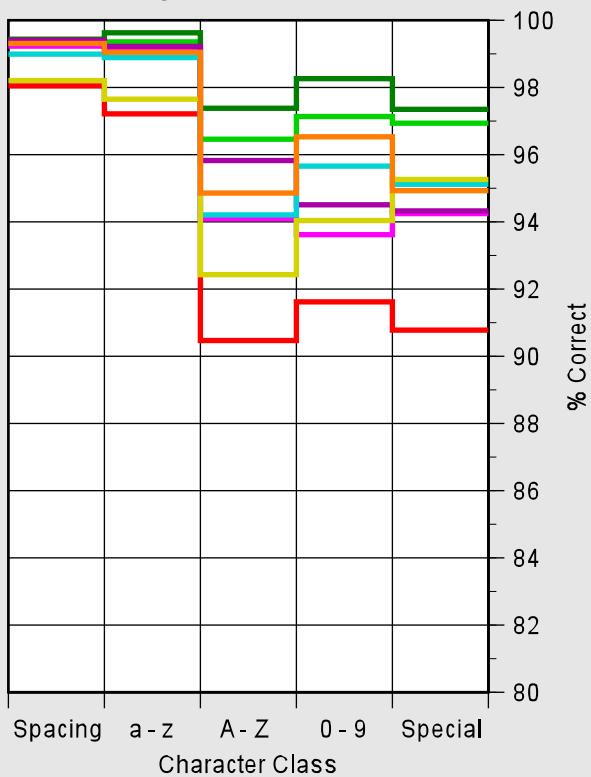




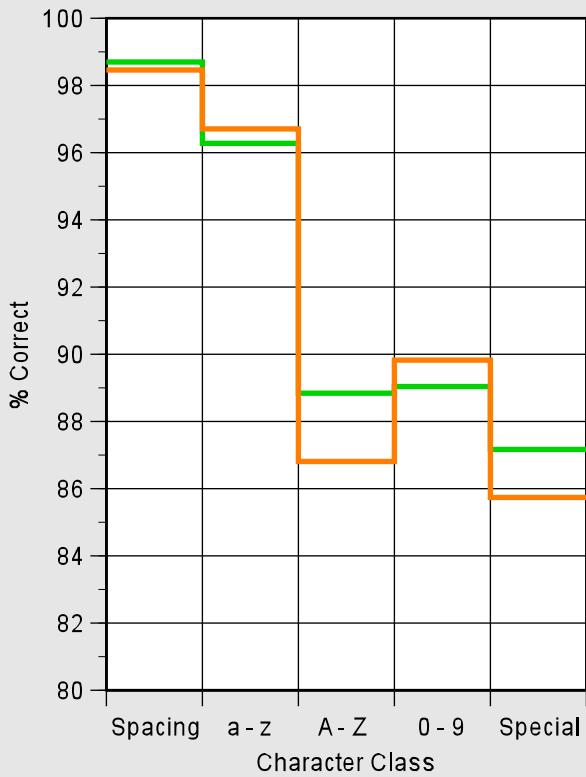
3 Accuracy by Character Class

- 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

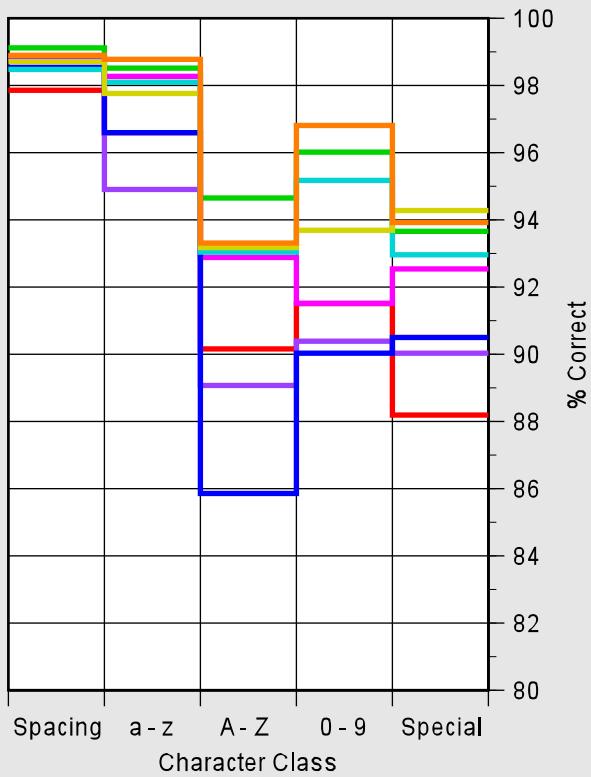
3a: Original Business Letters

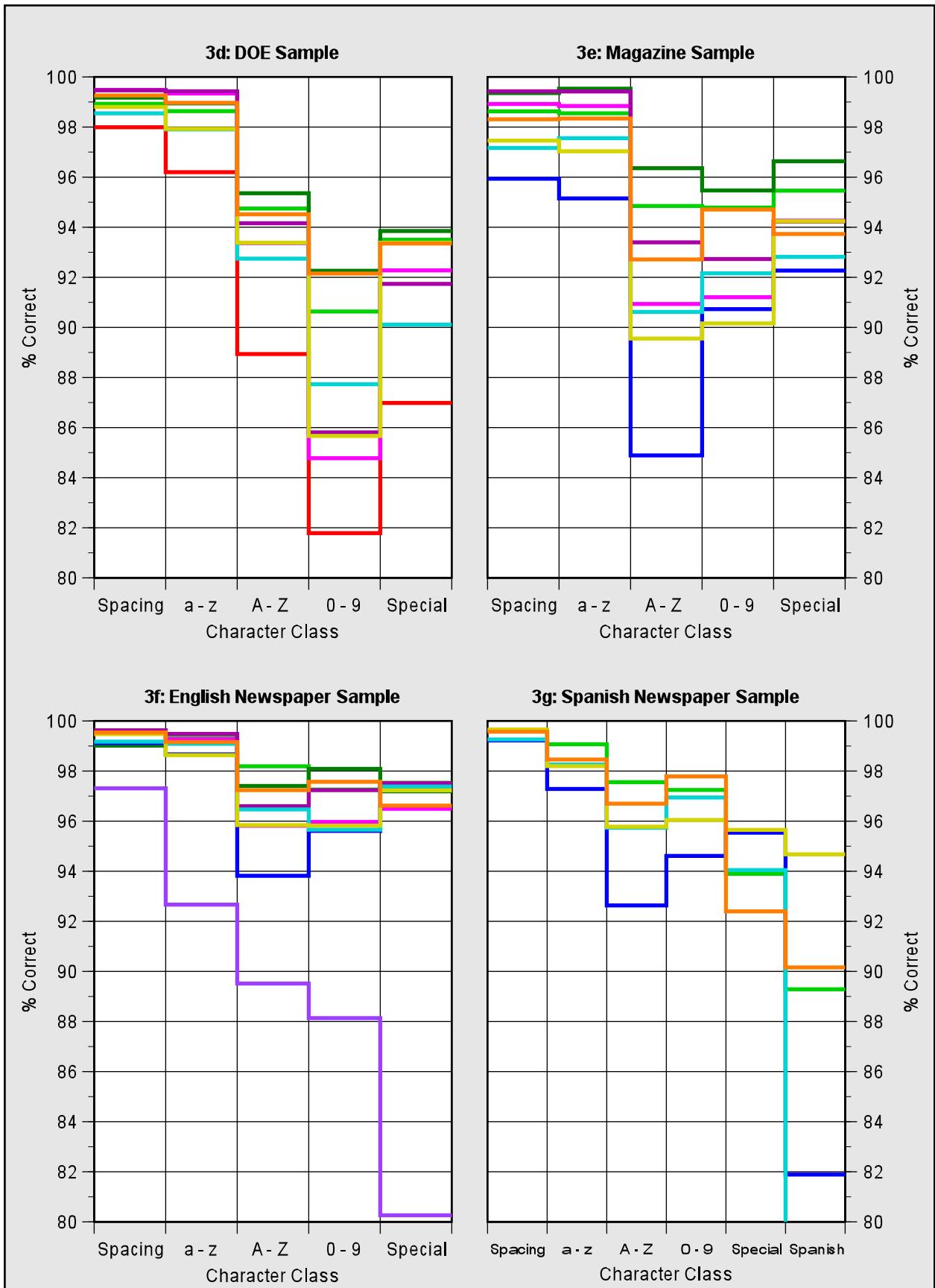


3b: Standard-mode Fax Business Letters

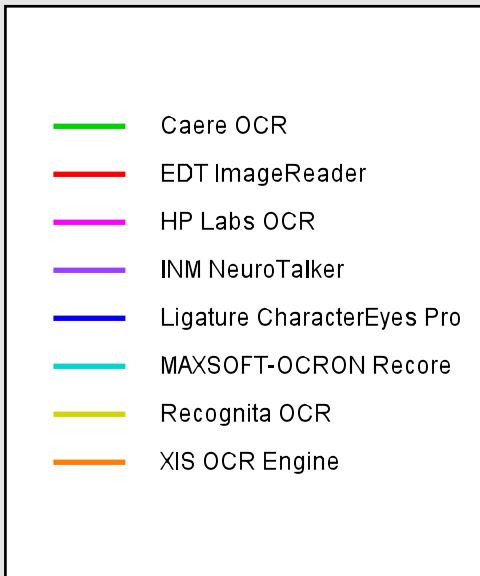


3c: Fine-mode Fax Business Letters

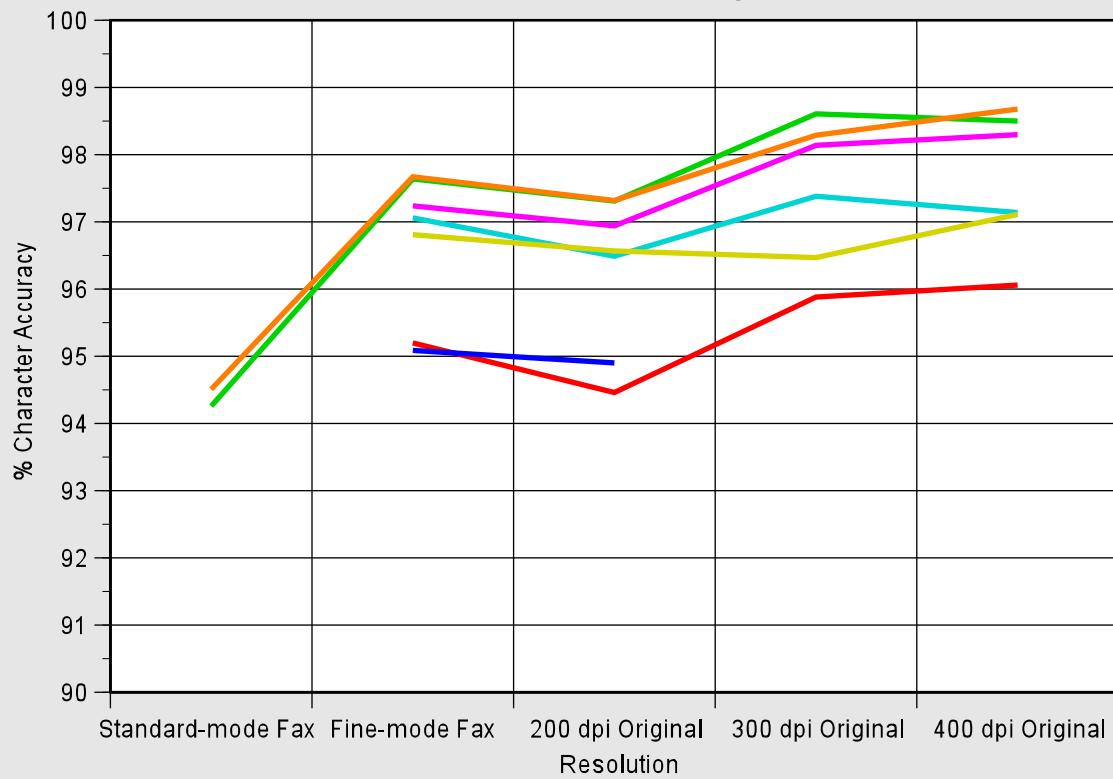


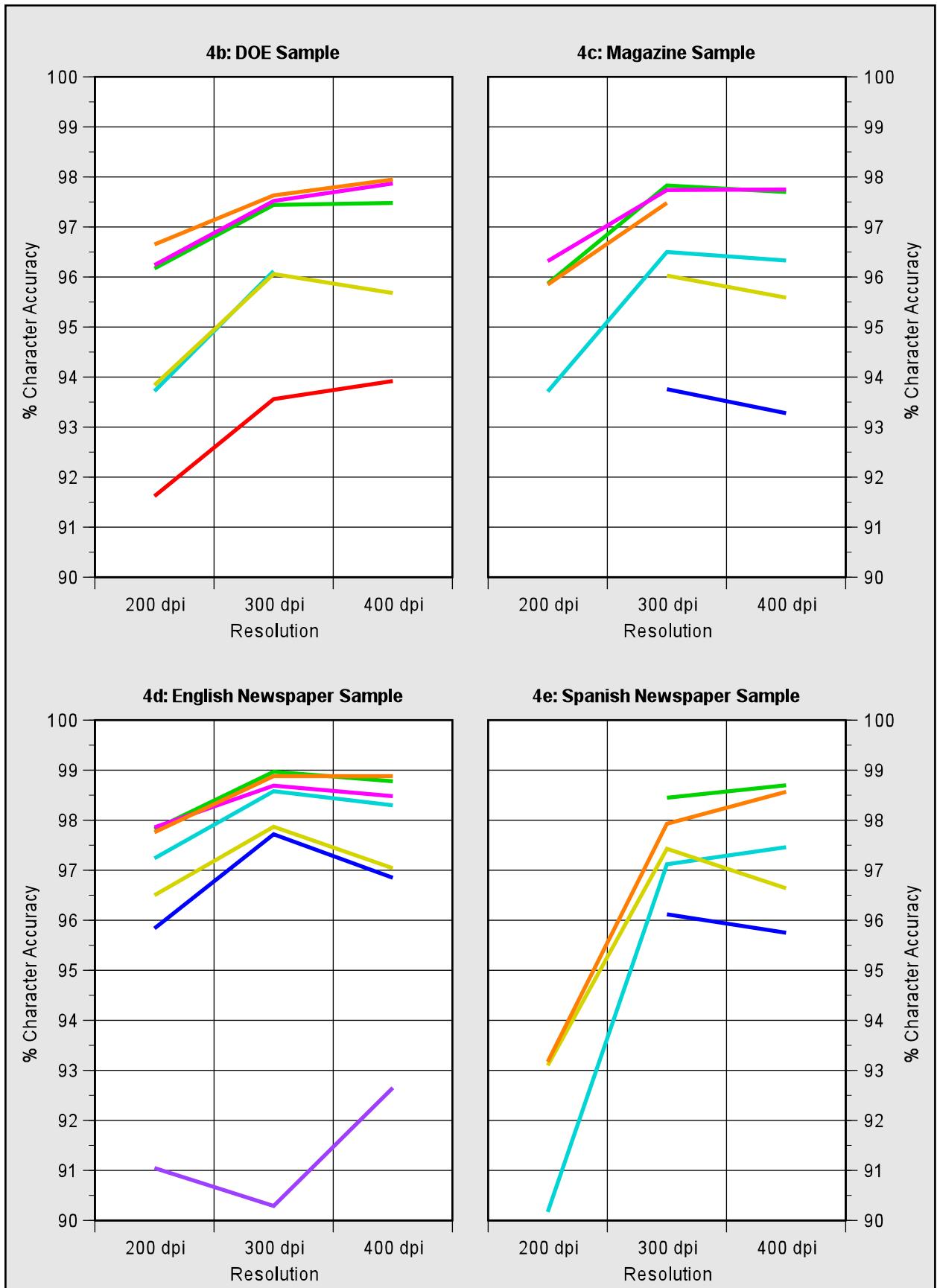


4 Effect of Resolution



4a: Business Letter Sample

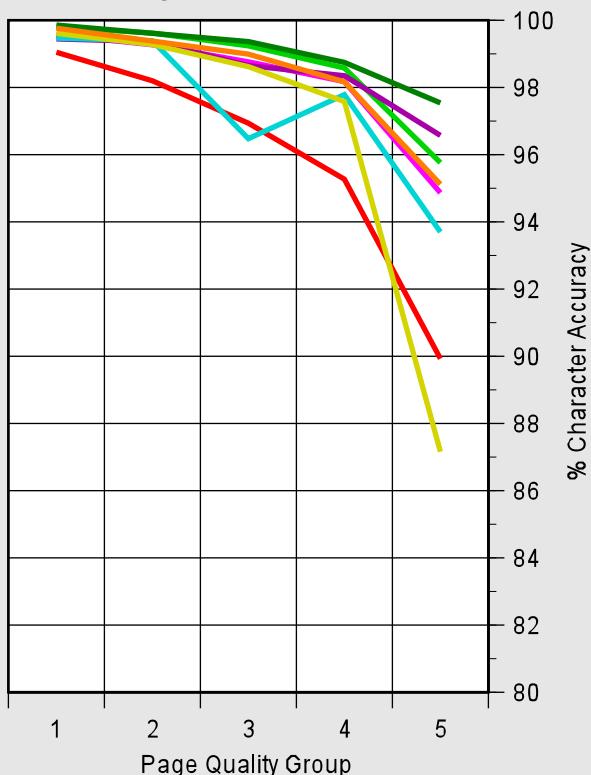




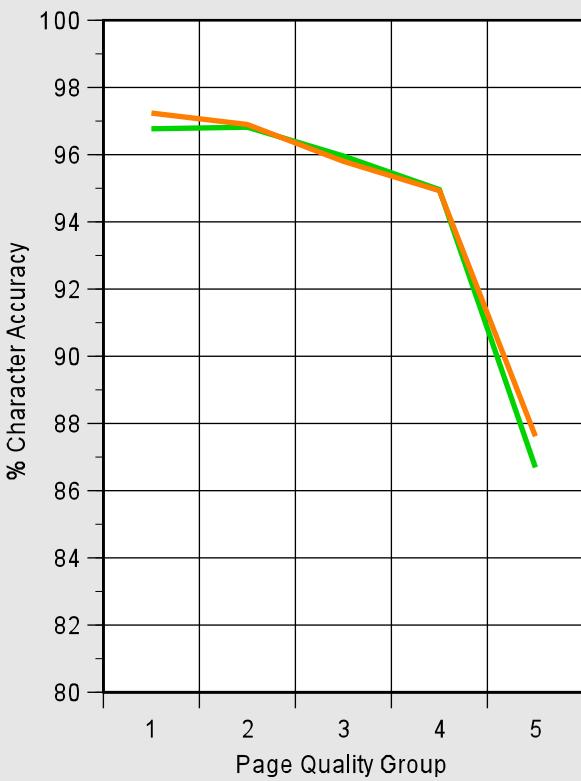
5 Character Accuracy vs. Page Quality

- 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

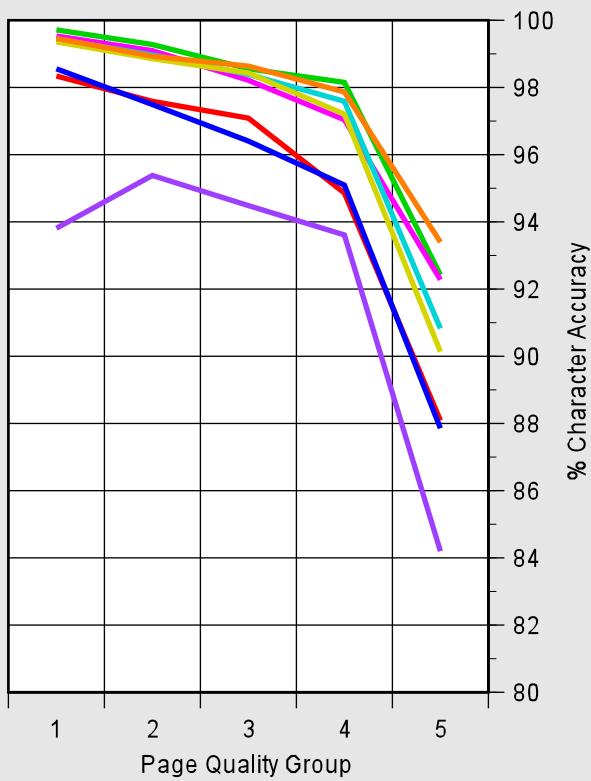
5a: Original Business Letters

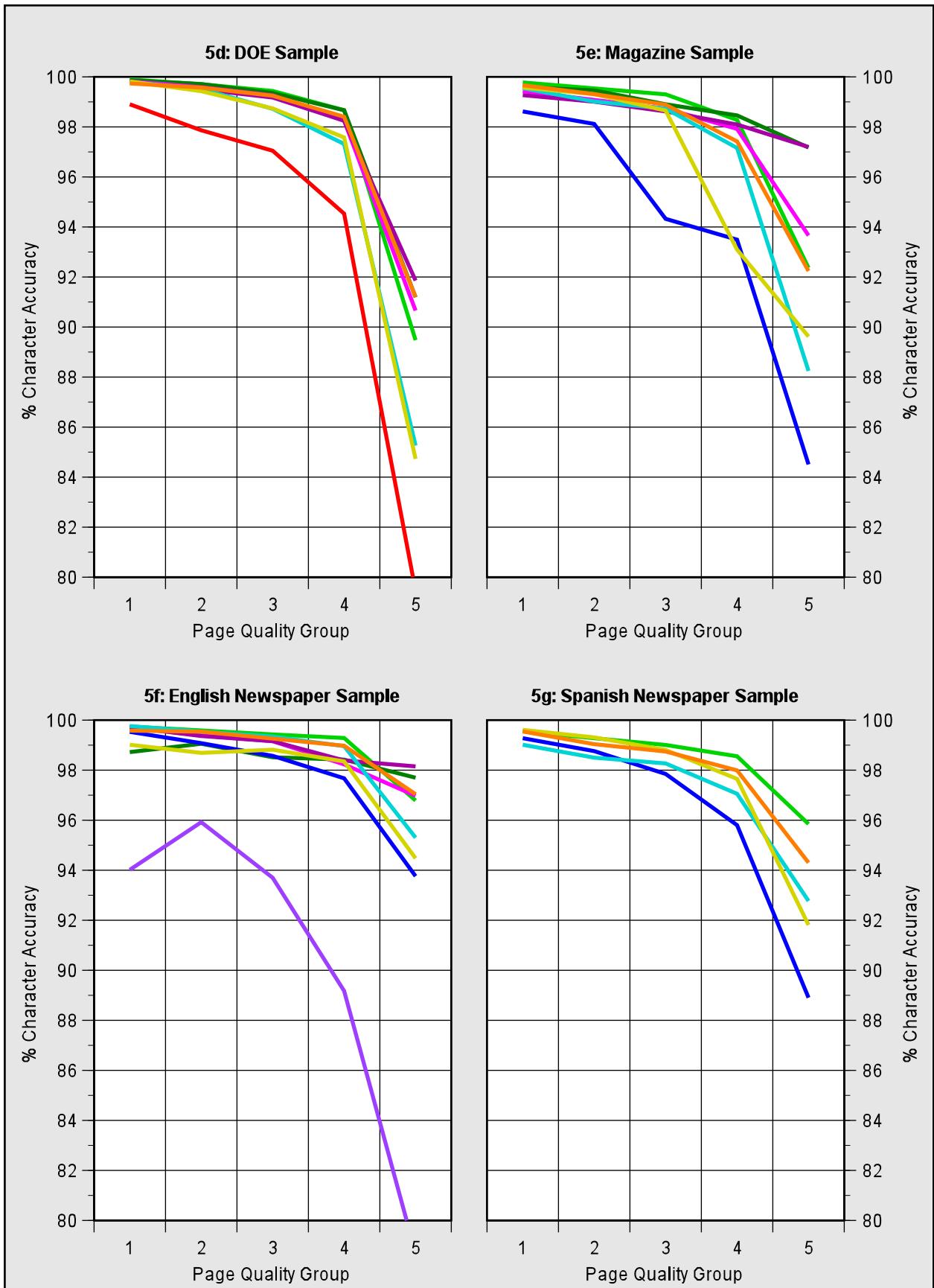


5b: Standard-mode Fax Business Letters



5c: Fine-mode Fax Business Letters





6 Word Accuracy vs. Page Quality

- 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

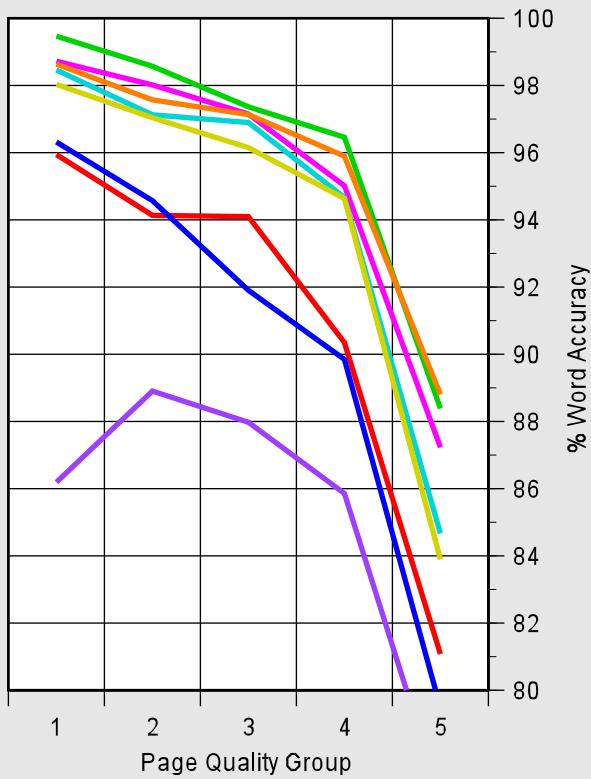
6a: Original Business Letters

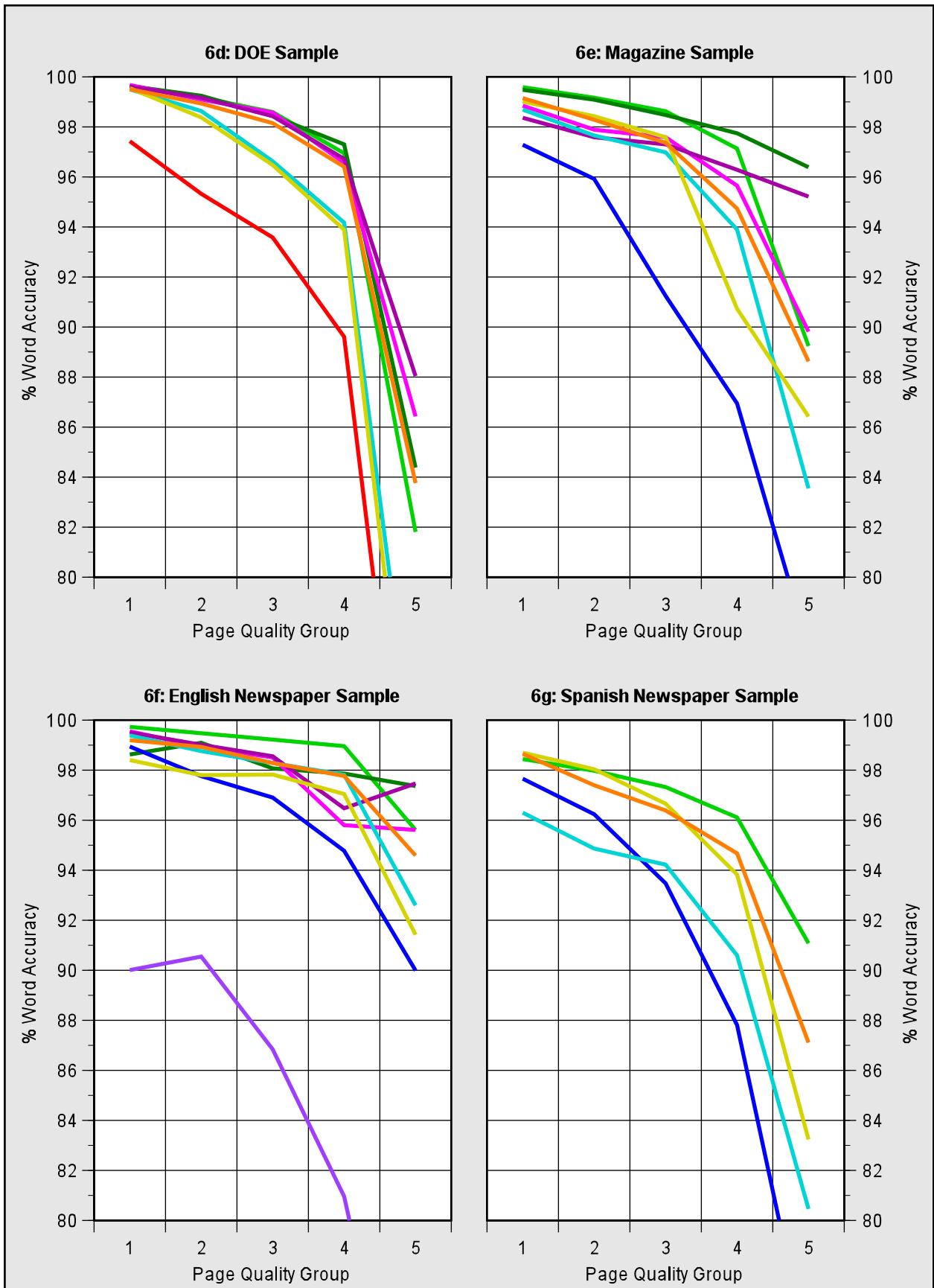


6b: Standard-mode Fax Business Letters



6c: Fine-mode Fax Business Letters

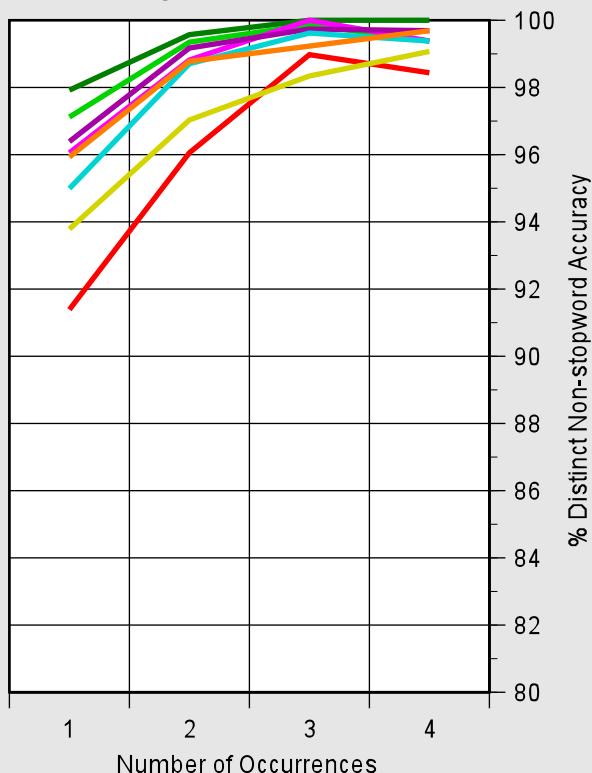




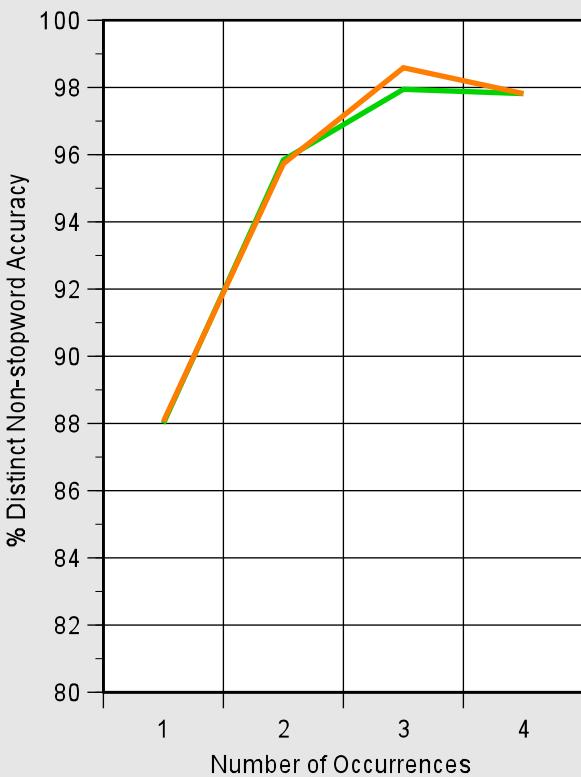
7 Distinct Non-stopword Accuracy

- 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

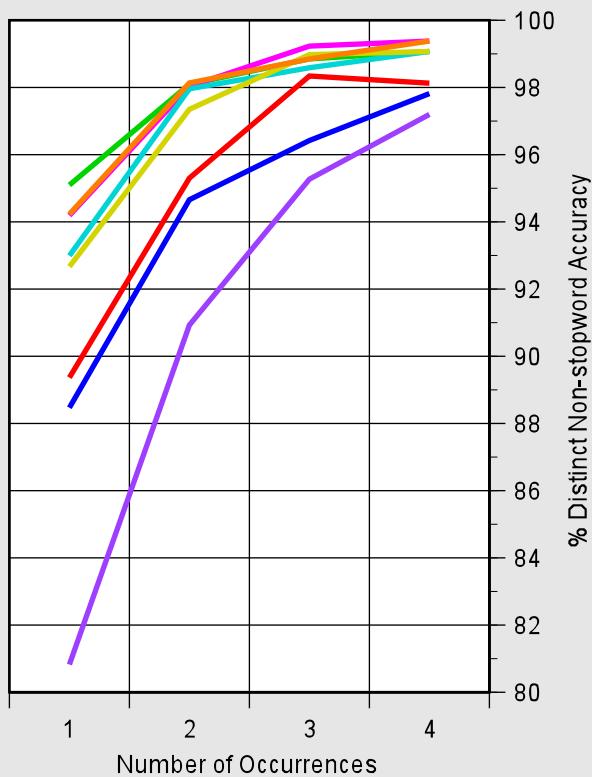
7a: Original Business Letters

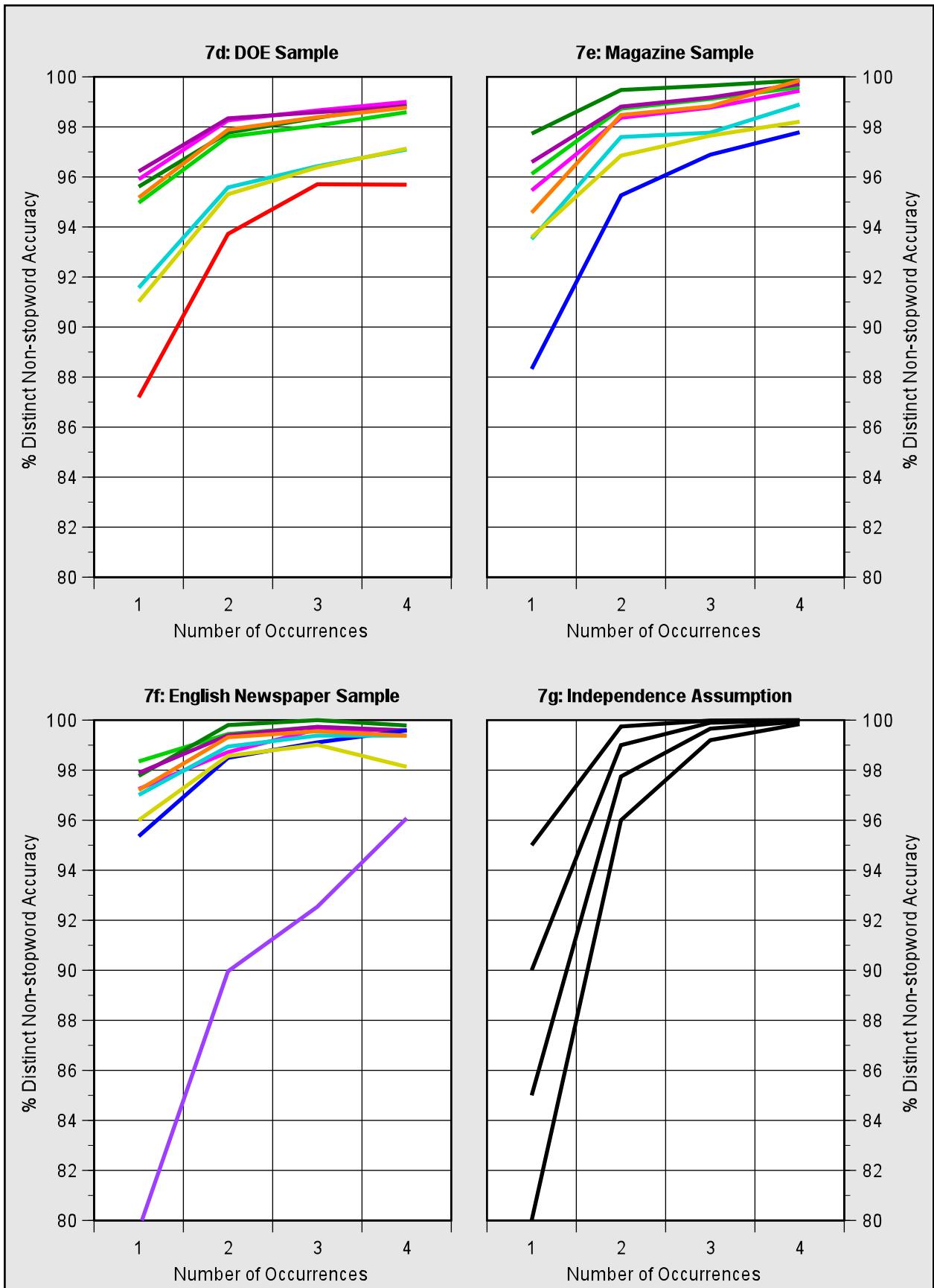


7b: Standard-mode Fax Business Letters



7c: Fine-mode Fax Business Letters

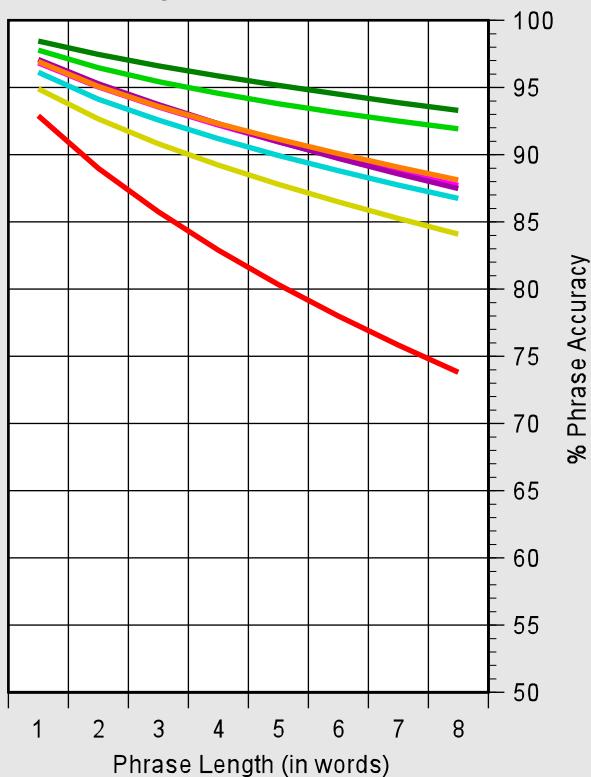




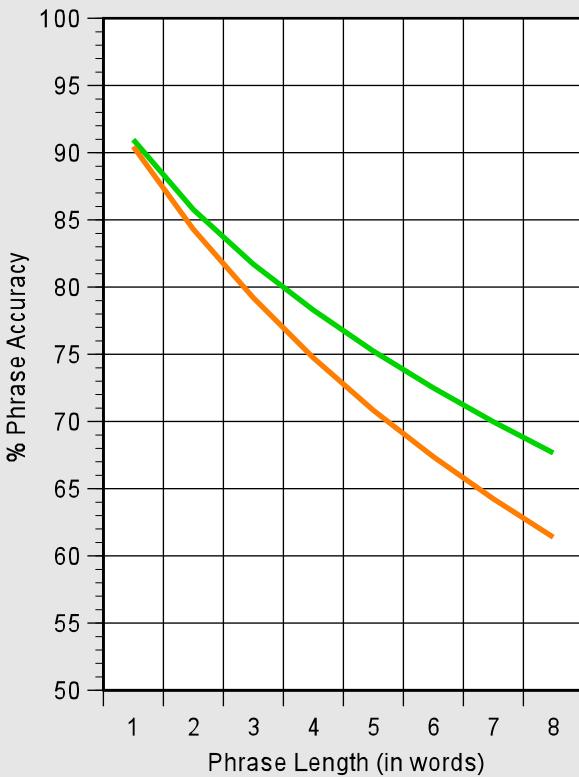
8 Phrase Accuracy

- 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

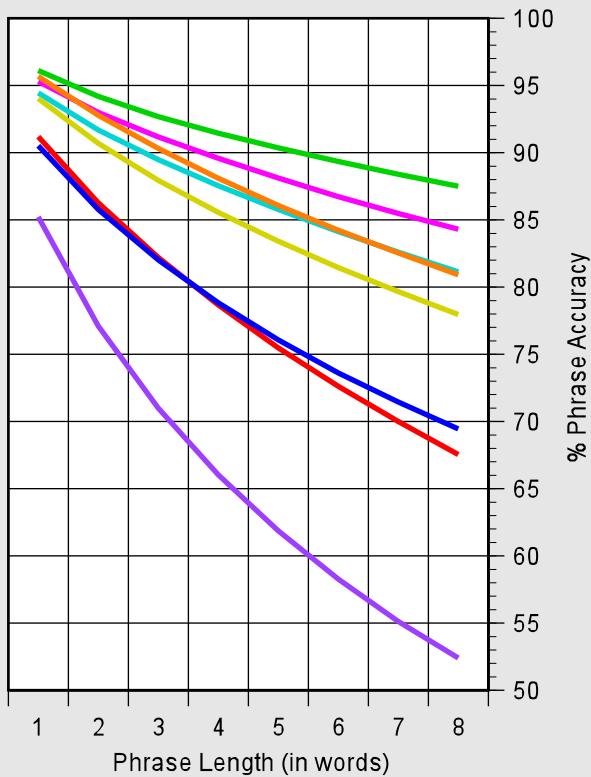
8a: Original Business Letters

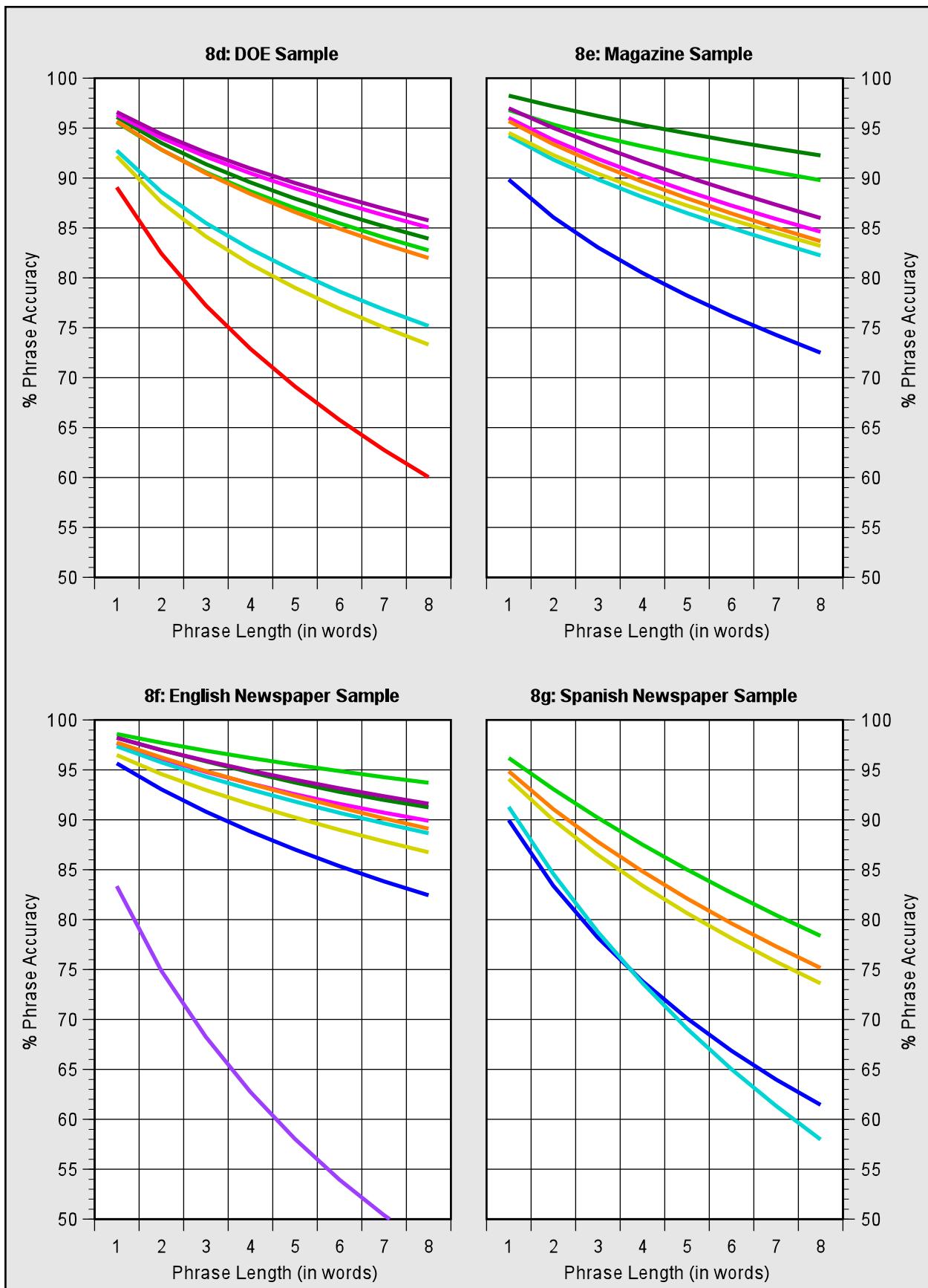


8b: Standard-mode Fax Business Letters



8c: Fine-mode Fax Business Letters

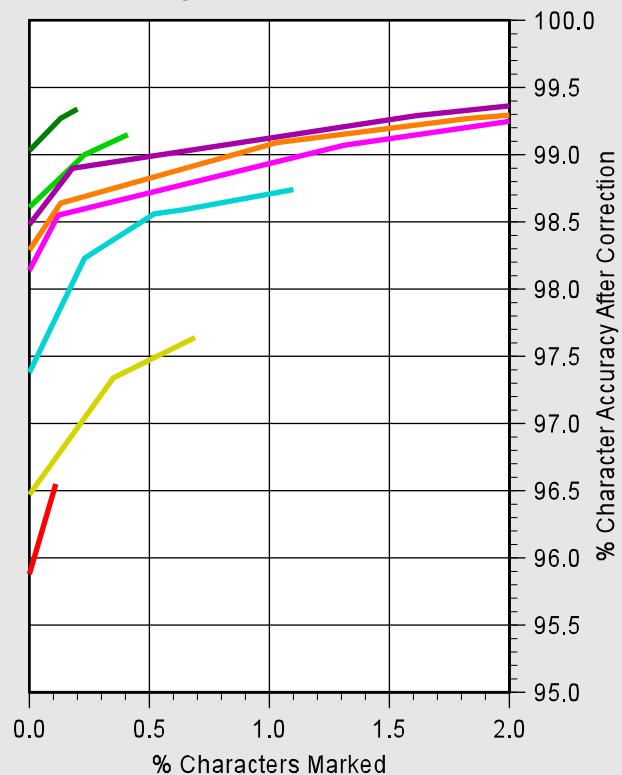




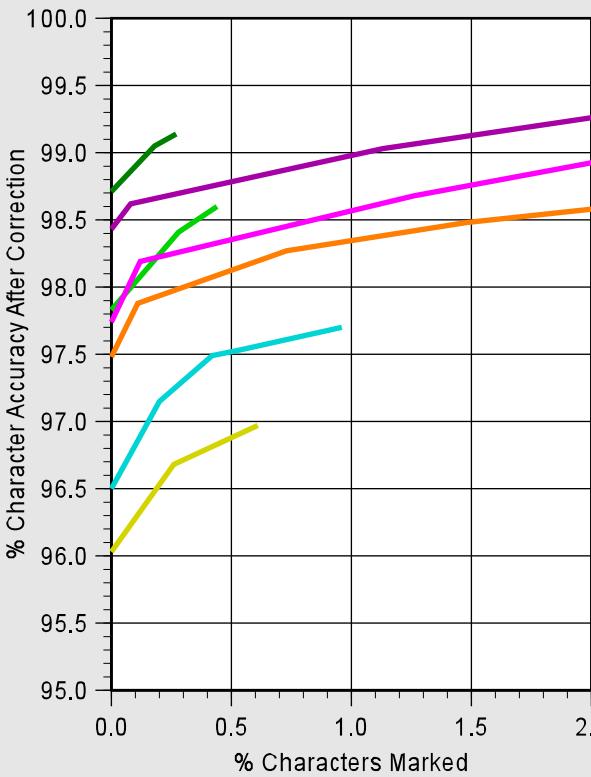
9 Marked Character Efficiency

- Caere OCR
- Caere OCR (gray scale)
- EDT ImageReader
- HP Labs OCR
- HP Labs OCR (gray scale)
- Ligature CharacterEyes Pro
- MAXSOFT-OCRONE Recore
- Recognita OCR
- XIS OCR Engine

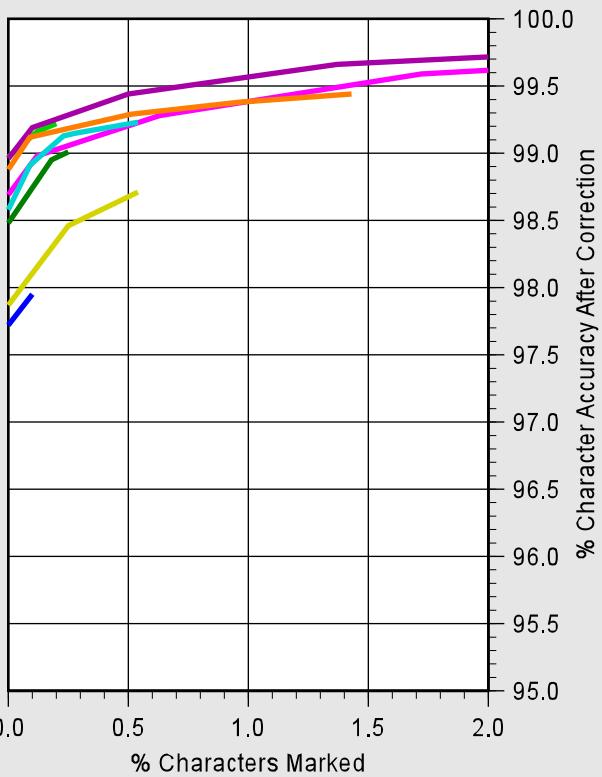
9a: Original Business Letters



9b: Magazine Sample



9c: English Newspaper Sample



10 Automatic Zoning

- Caere OCR
- Caere OCR (gray scale)
- EDT ImageReader
- Ligature CharacterEyes Pro
- MAXSOFT-OCRON Recore
- Recognita OCR
- XIS OCR Engine

