

**OCR ACCURACY
PRODUCED BY
THE CURRENT DOE DOCUMENT CONVERSION SYSTEM**

ISRI Staff

Technical Report 2002-06
Information Science Research Institute
University of Nevada, Las Vegas
Las Vegas, Nevada

May 2002

CONTENTS

1. INTRODUCTION	1
2. MEASURING CHARACTER ACCURACY OF TEXTUAL OUTPUT	1
3. CONVENTIONAL CHARACTER ACCURACY AS A MEASURE OF RETRIEVABILITY.	2
4. NON-STOPWORD ACCURACY OF DOE DOCUMENTS	3
5. SUMMARY	5
APPENDIX A. Non-stopword accuracy's and unique non-stopword accuracy's for OCR output	
APPENDIX B. Non-stopword accuracy's and unique non-stopword accuracy's for MANICURE output	

OCR ACCURACY PRODUCED BY THE CURRENT DOE DOCUMENT CONVERSION SYSTEM

ISRI Staff
Information Science Research Institute
University of Nevada, Las Vegas

May 2002

1. INTRODUCTION

The technical requirements for document collections submitted to the Licensing Support Network (LSN) have been set forth by the Nuclear Regulatory Commission (NRC). The current accuracy requirements for OCR'd document collections are a "target character accuracy of 99.5% with a 98.5% character accuracy target for each individual page."⁽¹⁾

The Department of Energy (DOE) has selected and installed the best available OCR technology for converting its document collection for submission to the LSN. It is using the "Developers Kit 2000" (SDK2000) distributed by the Scansoft Corporation. SDK2000 is based on combined technologies developed by the Calera, Caere, and Recognita Corporations, and is the best available page reading engine for general purpose use. In installing this system, care has been taken in determining operating parameters that maximize the quality of the output.

Having setup a document conversion system based on SDK2000, the DOE has tasked the Information Science Research Institute (ISRI) at the University of Nevada, Las Vegas (UNLV) to measure the accuracy of the output produced by this system when converting documents from the current DOE collection. This report describes the tests conducted by ISRI staff and gives the average accuracy's measured.

Because the ultimate purpose for the text produced by the OCR engine was to enter it into an Information Retrieval (IR) system⁽²⁾, DOE has chosen to use the ISRI designed MANICURE post-processing system. MANICURE was designed to accept the document text output from an OCR engine and to perform operations on it that would improve each documents retrievability. The operations applied by MANICURE include "garbage string removal" and spell checking and correction based on document level and collection level dictionaries that are built dynamically. Thus, the DOE document conversion system is a combination of the SDK2000 OCR engine and the MANICURE post-processing system. For this reason, both the accuracy's of OCR output text and of MANICURE output text are reported.

2. MEASURING CHARACTER ACCURACY OF TEXTUAL OUTPUT

The task of measuring the accuracy of textual output is complicated by several factors. First, in order to measure the accuracy of a text stream, it is necessary to have a "correct" text stream for comparison. In most cases, the cost of producing the correct, or ground-truth, character stream is very high⁽³⁾. Second, it is necessary to conduct such tests with large numbers of pages. No test of 5, or even 10 pages can be expected to produce statistically significant results. In general, it is preferable to have hundreds of test pages in order to insure significance of the measured results. Third, it is not a trivial issue to determine exactly which accuracy measure is most appropriate for a particular application. The standard measure, which we refer to as "conventional" character accuracy, measures the correctness of every ASCII character on each page. Correctness is defined as the number of total characters minus the number of character errors, divided by the total number of characters.

$$\text{Character Accuracy} = \frac{\text{Total Characters} - \text{Character Errors}}{\text{Total Characters}} \quad (\text{E1})$$

(1) Final Licensing Support Network Guidelines, April 2002, NRC

(2) The NRC has chosen the Autonomy search system for use in the LSN

(3) Not only must each character of the document be manually retyped, but each character must be checked and rechecked for correctness.

Character errors are the sum of character insertions, deletions, and substitutions that are necessary to convert an output character string into the exact ground-truth string.

In Section 3 of this report, we discuss the relevance of “conventional” character accuracy as a measure of goodness of the output of an OCR system. In Section 4, we describe a test to measure a different set of accuracy metrics. In particular, we measure word accuracy’s and the accuracy of the characters in words, produced by the current DOE system. For both of these metrics, we measure the accuracy of both the OCR output text and of the MANICURE output text.

3. CONVENTIONAL CHARACTER ACCURACY AS A MEASURE OF RETRIEVABILITY

The most important part of measuring the accuracy of any document conversion system is to determine what accuracy metric is most appropriate. There are many different performance metrics of conversion systems. The appropriate choice is the metric (or metrics) that best reflect improvement in the usage of the textual output. In this case, the output text will be used to build an index for the Autonomy search engine. Subsequently, the Autonomy engine will be used to retrieve documents of interest. Thus, it is the “retrievability” of documents that is most important.

Although the character accuracy of output text is related to retrievability, the conventional definition of character accuracy is not a good measure of retrievability. For example, OCR technologies typically output one or more characters for any set of black pixels on a page, even though these pixels do not resemble an ASCII character. Manufacturers of these technologies take the position that the user can easily delete such characters if they were generated because of stray marks. If these pixels were ignored by the system, it most certainly would not be noticed by the user. Just in case important information is represented, it is deemed better to draw the users’ attention (and require a delete operation) rather than risk losing important information. This phenomenon is especially noticeable when converting documents that are photocopies.

The overall result can be that a large number of delete operations are required to convert the output character string into the exact ground-truth string. Remember that each such delete operation is counted as a character error (see equation E1). Although the MANICURE system was designed to remove such noise, “conventional” character accuracy of OCR output will be affected by these delete operations.

Furthermore, since the conversion output is to be used to build the index in an IR system, it is the accuracy of the words to be indexed that better reflects retrievability. “Word” accuracy is defined as follows:

$$\text{Word Accuracy} = \frac{\text{Total Words} - \text{Number of Incorrect Words}}{\text{Total Words}} \quad (\text{E2})$$

In fact, since IR systems normally are setup to ignore some specific words, called stopwords (such as “the” & “and”), “non-stopword” accuracy is yet a better measure of retrievability. Equation E2 can also be used to calculate non-stopword accuracy by substituting non-stopwords for words.

The major point here is that print noise (or any stray marks), numbers, and punctuation marks in a document are NOT indexed by IR systems and thus, do not affect retrievability. Since “conventional” character accuracy can be profoundly affected by these kinds of characters, it is clear that “non-stopword” accuracy is a much better measure of retrievability. One possible alternative is to measure just the accuracy of the characters used to make up non-stopwords as an alternative to the “conventional” definition of character accuracy. We refer to this measure as the “character accuracy of non-stopwords” and use equation E1 replacing “characters” with “characters in non-stopwords.”

We thus undertook the task of conducting a test to measure the average “non-stopword” accuracy (and the character accuracy of these non-stopwords) that is produced by the current OCR/MANICURE system. To ensure that the full benefit of using MANICURE is measured, complete documents must be used in these tests. Our goal was to measure “non-stopword” accuracy from a set of “documents” selected at random from the DOE collection.

4. NON-STOPWORD ACCURACY OF DOE DOCUMENTS

The major impediment to document level tests of OCR accuracy is the cost of producing the “correct,” or ground-truth, copy of each page to use in calculating accuracy’s. The cost of producing accurate ground-truth for even two or three 80 page documents is extremely high. Thus, finding a low cost method of producing the ground-truth needed was a dominant part of conducting document level tests.

To solve this problem, we selected 17 documents at random from the DOE collection that had Microsoft Word based native files. The accession number, the total number of non-stopwords, and the number of characters in each of these non-stopwords are shown in Table 1. We developed a process to capture the correct output text directly from the Microsoft Word system. We also parsed this text to remove all punctuation, most of the digits, and all stopwords⁽¹⁾. A concerted effort was made to retain document identifiers and other “project words” containing digits.⁽²⁾ Thus, the text remaining contained only English non-stopwords (and project related non-stopwords that might not be in a normal dictionary) and formed the basis for computing accuracy’s⁽³⁾.

Table 1. Number of Non-stopwords and Characters in the 17 Document Sample				
Document Accession Number	Total Number of Non-Stopwords	Number of Characters in Non-Stopwords	Number of Unique Non-Stopwords	Characters in Unique Non-Stopwords
mol199907200407	6974	56611	912	7880
mol199911010207	9641	80684	1720	15199
mol200002170216	9595	79777	1796	16064
mol200002280529	5572	47005	1088	9674
mol200004130692	4115	33855	778	6641
mol200004140874	12379	101435	2131	18691
mol200005230155	4381	37387	769	7033
mol200005250378	13920	113193	2465	21669
mol200005260336	6318	51210	1193	10743
mol200006090266	5112	44172	1175	10750
mol200006270254	7792	62586	1307	11738
mol200007250453	36713	302763	4198	39005
mol200011220005	8440	70020	1611	14221
mol200012080086	4523	37402	959	8568
mol200101250233	14247	120653	1883	16925
mol200103160002	8441	69501	1338	12360
mol200104160088	6320	52870	907	7874
Average of all 17 documents	9675.47	80066.12	1542.94	13825.59

Because the images extracted from native Microsoft Word documents were never printed or scanned, they were completely free of defects associated with either the printing or scanning process. Although the cost of generating this ground-truth data was reasonable, tests of OCR output accuracy from these images would not produce results that were typical of the current DOE conversion operation. Even if each document were printed and scanned, the images produced would be of higher quality than the average image from the DOE collection.

- (1) The stopwords removed were the Brown Corpus list of 450 stopwords.
- (2) The criterion for “project-words” that were retained was the same as that used by the MANICURE system. Equations, tables, graphs, and other non-textual material were manually removed.
- (3) Note that the total number of characters of test data is over 1.3 million characters.

We therefore chose not only to print and scan each page of these documents, but to produce several generations of photocopies of each document. Our best judgement of the average quality of images in the DOE collection is somewhere between a first and second-generation photocopy. Thus, we chose to measure non-stopword accuracy's not only from original images, but also from the first printed and scanned image and from the first, second, third, and fourth generation photocopies of these images. Care was taken to use the same photocopy engine to produce all successive copies. To insure that all accuracy's measured were typical of current DOE conversion operations, all image copies were processed by BSC operations staff and the resulting OCR & MANICURE output was transmitted to UNLV on CD-rom

In designing this test of OCR output accuracy based on the characters of non-stopwords in documents, one other important issue was considered. If an OCR engine mis-recognized any of the characters of a non-stopword in a document, that document might still be retrieved by a retrieval engine. Since most non-stopwords exist several times in a document, if any one of these were recognized correctly, the document could still be retrieved by a word search. Thus, since retrievability is the important issue, it is also desirable to measure non-stopword accuracy based on "unique" non-stopwords. The idea is that character errors made in recognizing non-stopword A are not significant as long as one correct occurrence of A is generated. The number of unique non-stopwords in each document is also shown in Table 1.

Therefore, we constructed a program to measure both the average non-stopword accuracy (and the average accuracy of the characters of non-stopwords) for all non-stopwords and for "unique" non-stopwords from input documents. Although only 17 documents were involved, six different images of each document were tested. The first image was the "original" image extracted from Microsoft Word. The second image was the first-printed and scanned image and we refer to this as "generation 0". The third through sixth images are the first through fourth generation photocopies of generation 0 of these documents. The average OCR output accuracy's measured for all of these images for each document is shown in Appendix A. The average MANICURE output accuracy's measured for all of these images for each document is shown in Appendix B.

4.1 Summary of the Results of Non-stopword Accuracy Tests

The average character accuracy's for all 17 documents are shown in Table 2 below. The top two rows show accuracies from raw OCR output. The bottom two rows show accuracies from MANICURE output. The first and third row show accuracy's for all non-stopwords and the second and fourth rows show accuracies for unique non-stopwords.

TABLE 2. AVERAGE CHARACTER ACCURACY FOR ALL 17 DOCUMENTS							
System	Character Accuracy Of	Orig.	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4
RAW OCR OUTPUT	All non-stopwords	99.37	99.40	99.20	98.67	97.89	97.79
	Unique non-stopwords	99.65	99.58	99.46	99.22	98.93	98.80
MANICURE OUTPUT	All non-stopwords	99.50	99.47	99.30	98.83	98.16	98.06
	Unique non-stopwords	99.66	99.57	99.44	99.27	98.93	98.86

Table 2 shows that the character accuracy of all non-stopwords from the MANICURE output is slightly better than the raw OCR output. This improvement is more profound for higher generation photocopies (i.e., 97.79% for raw OCR and 98.06% for MANICURE for the fourth generation copy). It is interesting that this improvement is not reflected in the unique non-stopword results. Again the "improvement" of MANICURE output accuracy over raw OCR output accuracy is greatest for the fourth generation copy, but even then is only 0.06% (i.e., 98.80 for raw OCR and 98.86 for MANICURE). In general, this result shows that the MANICURE post-processing system does improve the accuracy of "all" non-stopwords in a document but does not significantly improve the accuracy of "unique" non-stopwords.

If the average image quality from the DOE collection is between a first and second-generation photocopy, then the most appropriate character accuracy (i.e., unique non-stopwords) is between 99.44% and 99.27%.

TABLE 3. AVERAGE NON-STOPWORD ACCURACY FOR ALL 17 DOCUMENTS

System	Word Accuracy Of	Orig.	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4
RAW OCR OUTPUT	All non-stopwords	97.44	97.03	96.45	95.05	92.78	91.91
	Unique non-stopwords	97.89	97.56	97.28	96.77	95.64	95.46
MANICURE OUTPUT	All non-stopwords	98.01	97.54	97.23	96.15	94.64	94.14
	Unique non-stopwords	98.74	98.36	98.15	97.65	96.61	96.51

The results shown in Table 3 above are much more significant. The word accuracy improvement of MANICURE output over raw OCR output for all non-stopword output ranges from 0.51% for generation 0 to 2.55% for generation 4. A one percent improvement in this accuracy measure is very significant ⁽¹⁾. The word accuracy improvement of MANICURE output over raw OCR output for unique non-stopword output ranges from 0.80% for generation 0 to 1.05% for generation 4.

These results show the benefit of applying the MANICURE post-processing system. Even an 0.8% improvement in word accuracy is significant. In addition, these results show that the improvement provided by MANICURE post-processing increases as page quality decreases. This is exactly as we expected, since MANICURE was designed to improve the accuracy of the non-stopwords in OCR output leading to improved overall document retrievability.

5. SUMMARY

We believe the unique non-stopword accuracy's between 98.15 for generation 1 and 97.65 for generation 2, as shown in Table 2, to be very high. It is important to note that word (and non-stopword) accuracy's are always lower than character accuracy's. ⁽²⁾

In terms of character accuracy, the character accuracy of unique non-stopwords for first and second-generation copies, as shown in Table 1, is between 99.44% and 99.27%. Although these accuracies are not quite at the 99.5% level, they are carefully measured results based on realistic documents and over 235,000 characters.

Overall, using the accuracy metrics ISRI believes are most appropriate, these results indicate that the character accuracy's produced by the current DOE document conversion system are very close to NRC requirements.

Finally, because retrievability of documents from the LSN will be the primary use of the text produced by the DOE, it seems clear that a test measuring retrieval effectiveness is at least as important; in fact, it is more important than the tests described above. ISRI has recommended that tests measuring retrievability (i.e., precision and recall) of documents from the Autonomy retrieval system be conducted.

⁽¹⁾ A one percent improvement over 97% correct words corresponds to eliminating 1/3 of the word errors.

⁽²⁾ Because one incorrect character in a word causes the whole word to be in error, and because character errors tend to be spread among different words, character accuracies are uniformly higher than word accuracy's. This phenomena has been repeated in every OCR test conducted by ISRI over the past 10 years.

APPENDIX A.

Non-stopword accuracy's and unique non-stopword accuracy's for OCR output

APPENDIX B.

Non-stopword accuracy's and unique non-stopword accuracy's for MANICURE output