

# Measuring and Delivering 95% Non-Stopword Document Accuracy

ISRI Staff

Information Science Research Institute  
University of Nevada, Las Vegas

## 1 Introduction

The Information Science Research Institute (ISRI) at the University of Nevada, Las Vegas (UNLV) was tasked with evaluating the performance of the Department of Energy (DOE) document conversion system used to prepare documents for the Licensing Support Network (LSN). Their current system applies automatic zoning and OCR followed by automated post-processing using a system developed at ISRI called, *MANICURE*. These documents will be subsequently loaded into an information retrieval (IR) system for document discovery during the licensing proceedings.

We know from previous ISRI studies[1, 5] that *MANICURE* improves OCR text for retrieval by correcting misrecognized non-stopwords. In fact, based on [1], the Nuclear Regulatory Commission (NRC) decided to change its measure of correctness from character accuracy to non-stopword accuracy. The minimum level of non-stopword accuracy determined appropriate for documents submitted to the LSN is 95%. This guidance necessitates a Quality Control (QC) procedure that will determine whether a document meets or exceeds this threshold.

To estimate the non-stopword accuracy of documents converted, statistics about non-stopwords must be collected. For example, the QC procedure must know the number of misspelled and correctly spelled words in a document to calculate its percentage of non-stopword correctness. Since *MANICURE* has access to this data as it processes a document, it made sense to include the QC procedure within *MANICURE*.

The following report substantiates the NRC's 95% threshold criterion and explains the process by which *MANICURE* calculates a document's non-stopword accuracy. Further, we explain the process that we used to validate the QC procedure's correctness.

## 2 The 95% Non-Stopword Accuracy Threshold

For several years, the recommended NRC guidelines for OCR-generated text focused on character accuracy. After studies performed by ISRI in 2002, it was determined that non-stopword accuracy would be a more appropriate measure for documents submitted to LSN for retrieval. The NRC's 95% non-stopword accuracy recommendation was influenced by these studies.

In [1], seventeen documents were selected from DOE's Records Information System (RIS) for the test. Six generations for these seventeen documents were produced. Each subsequent generation was a copy of the one before with the original generation being a direct image capture from Microsoft Word (never scanned or printed). With some subjective analysis, it was determined that the average image quality in the DOE collection was close to the first and second-generation photocopies applied in this test. Knowing that image quality on some of the documents submitted to the LSN would be lower, we ran some statistical tests on the third generation's non-stopword accuracy and calculated the statistics shown in Table 1.

Since the 17 document set was such a small sample, it would not be considered large enough to be representative of the DOE collection. We decided to run the same statistical test on a larger collection of 1055 documents obtained from the RIS to predict the OCR quality of the DOE's collection. The statistics from this larger collection substantiated the numbers in Table 1. Our initial threshold for meeting or exceeding NRC's requirement was set to the sample's confidence low, 0.95615.

mean	0.96235
std deviation	0.012058
confidence range	0.0061998
confidence low	0.95615
confidence high	0.96855
confidence percent	0.95
number of documents	17

Table 1: Non-stopword statistics for generation 3 in 17 document test set

Dictionary Type	Word Count
Ispell (English words)	260,079
GeoDictionary v 0.6b (geological words)	9171
Radiological and Nuclear Medicine words	158
1990 census surname data	88,799
12dicts, v2.0	77,204
LSN thesauri	19,894
LSS dictionary	136,781

Table 2: Dictionaries used in MANICURE v1.11

### 3 About MANICURE QC

The task of document preparation for the LSN begins with character recognition. The DOE is using the Developers Kit 2000 (SDK2000) distributed by the Scansoft Corporation[2]. Because the ultimate purpose for the text produced is to load it into an IR system, DOE also applies MANICURE as part of its document conversion process. MANICURE applies several algorithms to correct misrecognized non-stopwords in a given document. These are well documented in [4, 3, 6, 7]. What we would like to present in this report is the application of the *QC procedure* that was recently added to MANICURE version 1.11.

Implementing a QC procedure for measuring OCR quality may seem a simple task. Equation 1 gives the basic method for calculating non-stopword accuracy.

$$\frac{\text{correct\_words}}{\text{misspellings} + \text{correct\_words}} \tag{1}$$

The difficulty arises when one begins defining correct words and misspellings. A correct word should clearly be any word found in the dictionary and a misspelling, a word that is not. Of course, the completeness of the dictionary will have a huge impact on what gets marked as a misspelling. There are over 400,000 unique words in MANICURE’s dictionary.

MANICURE uses all the dictionaries listed in table 2 tailoring it specifically to the DOE collection.

Still, not all words correctly recognized will be marked as correctly spelled. Some proper nouns and acronyms will not occur in MANICURE’s dictionary. For this reason, MANICURE treats these words specially. MANICURE actually divides words in a document into four groups: correct words, rejected words, proper nouns/acronyms and misspellings. Following is the definition for each of these groups:

<b>Correct words</b>	all words that are found in the dictionary <i>plus</i> acronyms defined in the document’s text.
<b>Rejected words</b>	stopwords (see <code>stopwords</code> file) and any strings three characters or less in length.
<b>Proper nouns/acronyms</b>	words that look like proper nouns or acronyms but were not found in the dictionary or defined in the text.
<b>Misspellings</b>	all remaining words not in the lists above.

Separating words into the four groups above was done to aid correction. Only terms in the correct words list are used to correct misspellings in the document. Rejected words are ignored since stopwords are unimportant and short strings are not good candidates for correction. Words that look like proper nouns or acronyms but were not found in the dictionary are not marked as correct words or misspellings. This is purposeful since unless they are known to be correct, they should not be used to correct other words. On the other hand, it may also be a mistake to mark them as misspellings.

So how do these lists fit into the non-stopword accuracy calculation in Equation 1? Currently, only the correct words list and the misspellings counts are plugged into the formula. The question then becomes, is this an accurate assessment of non-stopword accuracy? We believed it to be, but felt it necessary to verify our position.

The most precise way of calculating non-stopword accuracy is to compare every word in these lists to what appears on the image. We did exactly this for more than fifty documents (1680 pages, 304,000 words). In every case, we found that if the words from the rejected, proper noun/acronym, and misspellings lists were verified for correctness, the non-stopword accuracy of the document *always* increased. In other words, the current method for calculating non-stopword accuracy for a document is a *lower bound*. Since this manual verification, we have lowered the threshold from 95.615% to 95.0% non-stopword accuracy so as not to fail documents that actually meet NRC's accuracy guideline.

## 4 Summary

The QC procedure included in MANICURE is a necessary prerequisite for documents being submitted to the LSN. In some preliminary runs, MANICURE QC identified certain problems with DOE's document conversion process that produced failed documents. Further, we believe that if a document passes MANICURE QC, it meets the NRC 95% non-stopword accuracy requirement and will be retrievable for users of the LSN.

## References

- [1] Tom Nartker and Ron Young. OCR accuracy produced by the current DOE document conversion system. Technical Report 2002-06, Information Science Research Institute, University of Nevada, Las Vegas, April 2002.
- [2] Scansoft, Inc., Peabody, MA. *Recognition API Manual*, v10 edition, 2000.
- [3] Kazem Taghva, Julie Borsack, Bryan Bullard, and Allen Condit. Post-editing through approximation and global correction. *International Journal of Pattern Recognition and Artificial Intelligence*, 9(6):911–923, 1995.
- [4] Kazem Taghva, Julie Borsack, and Allen Condit. An expert system for automatically correcting OCR output. In *Proc. IS&T/SPIE 1994 Intl. Symp. on Electronic Imaging Science and Technology*, pages 270–278, San Jose, CA, February 1994.
- [5] Kazem Taghva, Julie Borsack, Steven Lumos, and Allen Condit. A comparison of automatic and manual zoning: An information retrieval prospective. *International Journal on Document Analysis and Recognition*, January 2003. To Appear.
- [6] Kazem Taghva, Allen Condit, Julie Borsack, John Kilburg, Changshi Wu, and Jeff Gilbreth. The MANICURE document processing system. In *Proc. IS&T/SPIE 1998 Intl. Symp. on Electronic Imaging Science and Technology*, San Jose, CA, January 1998.
- [7] Kazem Taghva and Jeff Gilbreth. Finding acronyms and their definitions. *Int. Journal on Document Analysis and Recognition*, 1(4):191–198, May 1999.