

# FALCon: Evaluation of OCR and Machine Translation Paradigms



Written by:

Kathleen Swam  
Aberdeen High School

Mentor:

Ann Brodeen  
Information Science and Technology Directorate  
US Army Research Laboratory (ARL)

13 August 1999

## ABSTRACT

We evaluated the performance of the Forward Area Language Converter's (FALCon) embedded commercial-off-the-shelf (COTS) optical character recognition (OCR) and machine translation (MT) software for the Spanish language. We identified four critical factors for evaluation. The OCR's software performance was independent of the factors we evaluated; the MT software performance increased with a decrease in the length of the document.

### 1. INTRODUCTION

#### 1.1 Purpose

We evaluated the Forward Area Language Converter's (FALCon) optical character recognition (OCR) and machine translation (MT) software paradigms for Spanish. The purpose of evaluating the software was to determine where performance "bottlenecks" occurred. We believe the problems in the OCR may result from the documents' font size, quality, length, and type. Our experiments were designed to determine which of these factors significantly contribute to the problems with the OCR. Feedback from users in the field indicated that most of the problems with the MT software result from errors in the OCR. However, we also want to find some other common reasons for the MT errors that will be used in the development of a quantitative method for evaluating embedded MT systems.

#### 1.2 Background

FALCon is used by United States troops in foreign countries to translate foreign language documents into English. The troops capture documents in the field, obtain a rough translation via FALCon and decide whether to send them to linguists who carefully translate and analyze the documents. Several prototypes are currently in Bosnia. A prototype includes a laptop computer, paper scanner, and multiple communication links all enclosed in a specialized case. At this time, FALCon can translate Russian, French, Italian, Spanish, Portuguese, German, and Serbo-Croatian. Researchers are currently working on adding Arabic and Korean to its languages and engineering a lighter, modular configuration.

Many steps must be taken before a full translation is obtained. Once the documents are captured in the field they must be scanned into the laptop computer via the paper scanner. The scan is passed through the OCR software where it groups the dots (pixels) to characters and groups the characters into words. It compares the words to the built-in dictionary and highlights all the possible errors. Then the document is sent through the MT software. The first dictionary check is for the general language, the second dictionary check looks for words relating to the military that could give the troops an idea of whether it contains useful information and needs to be sent to the linguist for further evaluation.

## 2. OPTICAL CHARACTER RECONGITION

We began our evaluation with the COTS OCR software. This software takes a scanned image and groups the dots (pixels) into characters that are then compared to a template of characters for a specific alphabet (*e.g.*, Latinic, Cyrillic). The characters are grouped into words. The words are compared to the dictionary and possible errors are highlighted. The OCR'd document may be saved as a Microsoft Word Rich Text Format (RTF) file.

### 2.1 Factors Affecting OCR Quality

Initially, several critical factors were identified that could affect OCR accuracy: scanning process, document quality, font style, font size, document length, FALCon system, imaging parameters, document type and zoning. Many of the factors had to be overlooked due to time constraints. The following four factors were selected for evaluation: document quality, font size, document length, and document type.

#### 2.1.1 Scanning Process

The scanning process was partially controlled. A Visioneer PaperPort Strobe sheet-fed scanner was used to scan all documents into the computer at 300dpi (dots per inch) under the black and white, filing business card setting. The documents were scanned at different brightness levels. We observed that the scanning process, independent of the quality, affected the OCR more than previously thought.

#### 2.1.2 Document Quality

Document quality has three levels: low, medium, and high. Low quality document characteristics include yellowing or thin paper, multiple generation copies, and documents with speckle and touching or broken characters. The University of Nevada at Las Vegas (UNLV) Information Science Research Institute's (ISRI) 1994 study on OCR reported touching and broken characters as the most significant factors in determining document quality. Medium quality documents are the hardest to determine as it is strictly a judgement call. An example is a document with an extremely small font size and recognition is still difficult because there are broken and touching characters, but fewer than in a low quality document. High quality documents are the easiest to determine. They are usually printed on laser printers with white paper, and have little/no speckle and few/no broken or touching characters.

#### 2.1.3 Font Style

Most font styles can be classified as serif and sans serif. Due to time constraints, it was decided to look at only serif font styles.

#### 2.1.4 Font Size

Font sizes of 10, 12, and 14 were chosen as they are common sizes for the types of documents collected for the experiment.

### 2.1.5 Document Length

Three different document lengths selected for evaluation: paragraph, half page, and full page. Document length is directly proportional to the number of characters. A paragraph is approximately 500 characters, a half page is approximately 1250, and a full page is approximately 2000 characters. Length should be determined by the number of characters as different font sizes take up varying amounts of space on the page. What looks like a full page document at size 14 may have the same number of characters as a half page at size 10. Using the number of characters made the evaluation more precise and accurate.

### 2.1.6 System

Only one FALCon system was used during the experiment to eliminate variation due to differences in systems.

### 2.1.7 Imaging Parameters

Possible imaging parameters include color, grayscale, and fax mode. For this experiment we looked only at grayscale images.

### 2.1.8 Document Types

Document types include, but are not limited to, letters, faxes, newspapers, magazines, advertisements and technical articles. Only letters, newspapers, and magazines made up the document sample.

### 2.1.9 Zones

Zones refer to the different sections of a document (*e.g.*, header, caption, footer, main body, advertisement, and signature block). Only main body text was evaluated. Once a document was scanned all zones except the main body text were cropped and all text forced into a single column. It is anticipated the OCR will have problems with different zones and multiple column documents and these should be investigated at a later date.

## 2.2 Experimental Design

A full factorial design was initially selected for the study. Factorial designs are used in experiments with several factors where it's necessary to see the joint effect of the factors in the response (*i.e.*, each level of one factor is compared with each level of every other factor). A full factorial would have required 6480 documents. Finding this many documents would have been impossible. We began eliminating factors until we got the previously listed factors. It was narrowed down to 81 documents for a full factorial with the factors we chose. However, we still didn't have enough time to evaluate so many documents. We used the DESIGN-EXPERT software to narrow down the search. It has a variety of blocking patterns for use with central composite designs. We selected the Box-Behnken Design which can be used for three to seven

factors. Four factors, document length, document quality, document type, and font size, were to be evaluated; this design best suited our needs. It created a design with properties desirable for statistical analysis, but only required a fraction of the experiments for a full factorial. This particular Box-behnken design required only 25 documents, with the center point repeated 4 times. Collecting 25 documents was much more manageable for two people in the time allotted for the study.

## 2.3 Document Collection

Finding the documents needed to fit the design parameters turned out to be harder than expected. To evaluate the software properly we needed documents from different sources. We couldn't find all the documents to fit the parameters so some of them were manipulated (*i.e.*, documents were typed in at the needed font style or size, while some of the qualities were simulated via the scanning process).

### 2.3.1 Newspapers

The newspaper articles came primarily from Spanish newspapers printed in the United States, Mexico, and Puerto Rico. Some of the newspapers were old and many had started to yellow; the paper was thin and there was evidence of "bleed through". Most of these articles were low quality. Some articles came from newspapers on the Internet. These articles were printed on white paper utilizing a DeskJet printer. These articles were considered high quality.

### 2.3.2 Magazines

Our magazine articles came from Spanish versions of People Magazine and Reader's Digest published in Mexico. The magazine articles were on white glossy, but thin paper. Many had "bleed-through". Although some photo backgrounds were colored, due to the time constraints we only looked at body text so they were cropped out. Most of these articles were medium quality, but overall they varied.

### 2.3.3 Letters

Our letters were personal letters. One was from a friend who e-mailed us in Spanish. The others came from a school pen pal that wrote to us from Argentina. These were all on white paper and printed on laser printers. For the most part, these were high quality documents.

## 2.4 Evaluation Preparation

Once the required documents were collected, word and character counts were computed for all the documents. A groundtruth version of each authentic document was typed into Microsoft Word as a model for each original. The groundtruth should mimic the original - mistakes and all. Upon completion of the groundtruthing task, the originals needed to be scanned into the computer. The documents were scanned using various brightness levels depending on the document quality. Some of the documents were lightened or darkened to fit the quality level

needed for the experiment. Darkening the documents degraded their quality by adding speckle and touching characters.

## 2.5 Evaluation

Once the documents were scanned in at the correct brightness, they were passed through the OCR software module. Once the software recognized the letters and words, it made a split screen showing the original scanned document and the OCR document so you could see the differences. The OCR'd documents were saved as Microsoft Word RTF files. The scanned original and MS Word groundtruth documents were then passed through the automated evaluation software for comparison .

The Department of Defense scoring software was used for automatically scoring the character accuracy of the OCR. The character accuracy was calculated from the number of errors as follows:

$$\frac{n - (\text{number of errors})}{n},$$

where  $n$  is the total number of characters in the groundtruth file. Every character inserted, substituted, or deleted to correct the OCR generated text to make it like the groundtruthed text is counted as an error. The overall accuracy of the 25 documents was 96.51%.

The accuracy reports showed the total number of characters compared to the total number of errors and the accuracy percent (see Appendix A). A report was generated for each document as well as a cumulative report. It listed all the confusions including what was generated, what was correct and the number of errors for each. It also listed all the possible characters, the number of those characters in the document, the number missed, and the percentage right. We generated a list of the most common OCR errors, or confusions, in the 25 documents. The report broke down the confusions into one-to-one, one-to-two, two-to-one, and two-to-two, confusions (*i.e.*, one character was confused with another character, one character was confused with two characters, etc.). We charted the ten most common one-to-one confusions, five most common one-to-two confusions, ten most common two-to-one confusions, and five most common two-to-two confusion (see Appendix B).

## 2.6 Stop Word List

Stop words are common words like “the,” “and,” “of,” “in,” etc. These words are typically not indexed in information retrieval systems. We translated an English stopword list from Cornell University into Spanish. We encountered many problems trying to translate the list. Most of the stop words have several different meanings depending on the context in which it is used. Also, when translating from English to Spanish, or vice versa, the words aren't always one-to-one mappings. Many times a word in English is translated to a phrase in Spanish. This created problems because the scoring software would only accept one-to-one mappings. We did not use the stop word list created as we weren't investigating the information retrieval component of the system. It is available for use at a later time.

### 3. MACHINE TRANSLATION

Once the documents are OCR'd, they are passed through the MT software. This program also showed a split screen with the original Spanish document and its English translation. We saved the translated documents as Microsoft Word RTF files. We initially began the evaluation by comparing the original Spanish document and the English MT version and tallying the mistakes in the MT. However, we found this to be very difficult and time consuming. We decided to try a new approach by translating all the documents ourselves and comparing our translations against the MT versions. This was a much easier approach. While my colleague, a native speaker, translated the original Spanish documents into English, I began comparing the MT translations to her translations and logging the errors. We made tally sheets to document the errors. The sheet showed the MT error and its correction, as well as the reason for the errors. Once all the errors were found they were tallied and broken down according to the reason for the error (see Appendix C).

#### 3.1 Common Errors

When we started looking at the MT component, we didn't really know how to go about evaluating its performance. Not having much past research to go by, we began by checking the translations. Once we realized this approach wouldn't work, we translated the documents ourselves and checked the human translations against the MT versions. This allowed us to compile a list of the most common types of errors that occurred.

##### 3.1.1 Word Order Errors

Word order errors occurred when the translated words were put in an arrangement that made no sense. This is a big problem when translating from Spanish to English, or vice versa, as some words and phrases are said in a different order.

##### 3.1.2 Context Errors

These errors occurred when a word with multiple meanings was used in the wrong context. Many words in Spanish have many different meanings depending on how they are used. It is very hard for the MT software to understand this as, in most cases, its "memory" consists of only the previous sentence.

##### 3.1.3 Pronoun Errors

Pronoun errors were very common mistakes in our study. The software frequently misplaced he, she, it, them, etc. The probable cause is that English is not a gender specific language, while Spanish is.

##### 3.1.4 Dictionary Errors

Dictionary errors occurred when the software simply used the wrong word. This wasn't a huge problem, but it did happen.

### 3.1.5 OCR Errors

OCR errors occurred because the OCR didn't recognize certain words which filtered down to the MT process. This was a common problem albeit not the MT software's problem.

### 3.1.6 Missing Word Errors

Missing word errors occurred when the software "skipped" words or failed to translate certain words to English.

### 3.1.7 Extra Word Errors

Extra word errors added unnecessary text, causing the translation to be incorrect.

### 3.1.8 Translation Errors

If in the English translation the word was still in Spanish, it was classified as a translation error.

### 3.1.9 Proper Name Errors

Proper name errors occurred when the MT software didn't recognize a proper name, but translated it to something else which confused the meaning.

## 3.2 Evaluation

When we began to work with the MT software we discovered, there was no quantifiable method for evaluation. Creating an exact method for MT is difficult due to the many differences in the languages. Our evaluation had to be performed manually as currently no automated scoring software for MT exists. It was decided the best way to evaluate the performance of the software was to calculate the percentage of translated words, similar to the OCR evaluation. The word accuracy was calculated as:

$$\frac{n - (\# \text{ of } n \text{ wrong})}{n}$$

where  $n$  was the number of words in the English MT. Every word in the MT that was different from the human translation was counted as an error. The overall accuracy of the documents was 55.17%. (see Appendix D.)

## 4. CONCLUSIONS

### 4.1 Scanning Process

We discovered that the scanning process significantly affected the OCR and, subsequently, the MT more than previously thought. The quality of the scan plays a significant role in the quality of the OCR accuracy. To improve the scan and OCR accuracy, we recommend you remove all stray dots and straighten the page for every document. We also recommend you pay careful attention to the brightness level when scanning the documents. Adjust the settings as needed to achieve the best scan possible. Unfortunately, the soldiers in the field probably don't have the time to carefully scan in all the documents.

### 4.2 OCR

OCR accuracy was shown to be independent of the factors evaluated. The character accuracy ranged from 90.79% – 99.76% across the sample documents. The overall accuracy was 98.10%.

### 4.3 Machine Translation

The word translation accuracy ranged from 43.09% to 84.74% across the sample documents. The overall accuracy was 55.17%. Although the accuracies seem low, the translations should be adequate to determine the military relevance of the documents.

## **Acknowledgements**

The author would like to thank Ann Brodeen, her mentor, for her endless help and guidance with the research, report and presentation; Fred Brundick for running the automated scoring software used for evaluating OCR accuracy; Dawn French for getting us the pictures of FALCon; Mrs. Burgos for her help with the human translations; and Raquel Burgos, her partner, for all her help and support with the research and presentation.

## APPENDIX A

### OCR Character Accuracy

Document Number	Document Length	Document Quality	Document Type	Font Size	Number of Characters	Number of Errors	Accuracy Percentage
1	paragraph	medium	letter	12	671	42	93.74
2	half page	low	letter	12	1537	58	96.23
3	half page	medium	letter	10	1598	13	99.19
4	half page	medium	letter	14	1537	45	97.07
5	half page	high	letter	12	1598	6	99.62
6	full page	medium	letter	12	2025	33	98.37
7	paragraph	low	newspaper	12	435	17	96.09
8	paragraph	medium	newspaper	10	731	34	95.35
9	paragraph	medium	newspaper	14	1111	17	98.47
<b>10</b>	<b>paragraph</b>	<b>high</b>	<b>newspaper</b>	<b>12</b>	<b>1259</b>	<b>3</b>	<b>99.76</b>
11	half page	low	newspaper	10	1341	120	91.05
12	half page	low	newspaper	14	1135	15	98.69
13	half page	medium	newspaper	12	1070	13	98.79
14	half page	high	newspaper	10	1517	21	98.62
15	half page	high	newspaper	14	1329	18	98.65
16	full page	low	newspaper	12	2304	69	97.01
17	full page	medium	newspaper	10	3258	155	95.24
<b>18</b>	<b>full page</b>	<b>medium</b>	<b>newspaper</b>	<b>14</b>	<b>1889</b>	<b>174</b>	<b>90.79</b>
19	full page	high	newspaper	12	1896	93	95.09
20	paragraph	low	magazine	12	494	11	97.77
21	half page	low	magazine	12	1496	131	91.24
22	half page	medium	magazine	10	988	84	91.50
23	half page	medium	magazine	14	1421	32	97.75
24	half page	high	magazine	12	1436	9	99.37
25	full page	medium	magazine	12	2246	10	99.55

## APPENDIX B

### Most Common One to One Confusions

	Correct	Generated	Number
1	l	l	15
2	o	a	11
3	a	s	8
4	a	e	8
5	n	a	7
6	e	c	5
7	a	o	5
8	a	?	5
9	I	H	3
10	i	?	3

### Most Common One to Two Confusions

	Correct	Generated	Number
1	m	rn	2
2	w	vr	2
3	n	r.	2
4	h	l.	2
5	d	?l	2

### Most Common Two to One Confusions

	Correct	Generated	Number
1	fi	6	24
2	fr	&	20
3	la	h	10
4	ll	H	8
5	ro	m	8
6	fi	s	4
7	fr	k	4
8	re	m	4
9	FI	H	4
10	en	m	4

---

### Most Common Two to Two Confusions

	Correct	Generated	Number
1	su	SU	2

2	im	?n	2
3	qu	v	2
4	no	rm	2
5	da	ck	2

Note: (?)= Unrecognizable ( \_ ) = space

## APPENDIX C

### Totals of Common Errors in Machine Translation

	Word Order	OCR	Extra	Missing	Pronoun	Context	Translation	Dictionary	Conjugation	Proper Name
1	4	8	9	4	0	8	6	3	5	1
2	14	19	22	8	6	13	7	7	3	1
3	3	9	17	12	11	8	5	9	3	1
4	13	3	26	2	7	25	3	10	4	2
5	6	7	31	12	10	19	4	13	2	0
6	16	10	35	16	20	36	3	11	6	0
7	1	10	3	0	2	2	2	6	2	1
8	1	5	3	2	1	4	0	7	2	1
9	5	13	14	12	3	11	2	4	3	1
10	7	15	32	12	6	17	2	8	5	0
11	7	32	12	10	4	10	8	7	2	2
12	5	14	28	4	1	21	8	9	3	2
13	3	1	14	2	1	11	2	3	7	1
14	5	14	24	5	5	24	1	12	12	3
15	8	5	14	7	14	17	1	7	3	0
16	3	22	7	8	6	3	4	2	3	3
17	16	38	19	18	2	12	2	5	1	0
18	8	26	21	22	6	27	4	7	13	2
19	12	2	21	22	6	24	2	6	8	1
20	4	6	14	4	3	4	2	1	1	0
21	10	27	21	4	10	16	5	16	6	3
22	2	6	13	4	2	16	2	8	3	1
23	6	19	33	9	11	23	5	10	6	1
24	8	15	16	8	10	23	8	8	4	4
25	16	11	11	12	11	16	4	8	4	8
T	183	337	474	206	158	390	92	187	111	39
%	8.41	15.48	21.77	9.46	7.26	17.91	4.23	8.59	5.10	1.79

## APPENDIX D

### Machine Translation Word Accuracy

Document Number	Document Length	Document Quality	Document Type	Font Size	Number of Words	Number of Errors	Accuracy Percentage
1	paragraph	medium	letter	12	140	55	60.71
2	half page	low	letter	12	275	123	55.27
3	half page	medium	letter	10	324	108	66.67
4	half page	medium	letter	14	283	110	61.13
5	half page	high	letter	12	322	112	65.23
6	full page	medium	letter	12	405	174	57.04
7	paragraph	low	newspaper	12	68	27	60.29
8	paragraph	medium	newspaper	10	121	24	80.17
9	paragraph	medium	newspaper	14	192	77	59.90
<b>10</b>	<b>paragraph</b>	<b>high</b>	<b>newspaper</b>	<b>12</b>	<b>246</b>	<b>140</b>	<b>43.09</b>
11	half page	low	newspaper	10	252	116	53.97
12	half page	low	newspaper	14	199	110	44.72
13	half page	medium	newspaper	12	179	55	69.27
14	half page	high	newspaper	10	285	125	56.14
15	half page	high	newspaper	14	236	78	66.95
<b>16</b>	<b>full page</b>	<b>low</b>	<b>newspaper</b>	<b>12</b>	<b>426</b>	<b>65</b>	<b>84.74</b>
17	full page	medium	newspaper	10	583	151	68.99
18	full page	medium	newspaper	14	328	178	43.73
19	full page	high	newspaper	12	306	138	54.90
20	paragraph	low	magazine	12	98	42	57.14
21	half page	low	magazine	12	285	144	49.47
22	half page	medium	magazine	10	182	69	62.09
23	half page	medium	magazine	14	278	142	48.92
24	half page	high	magazine	12	279	136	51.23
25	full page	medium	magazine	12	398	119	70.10

## Bibliography

Broome, B. "FALCon (Forward Area Language Converter): Translingual Help for U.S. Troops." U.S. Army Research Laboratory, Aberdeen Proving Ground, MD, 1999.

Design-Expert Software, Version 5.0 User's Guide. "Statistical Details: Design Selection." STAT-EASE, Inc.

Myers, R. H. and D.C. Montgomery. Response Surface Methodology: Process and Product Optimization Using Designed Experiments. John Wiley & Sons, Inc: New York, 1995.

Rice, S.V., J.Kanai, and T.A. Nartker. UNLV Information Science Research Institute 1994 Annual Report. "The Third Annual Test of OCR Accuracy." Information Science Research Institute: Las Vegas, 1994.