Online Myanmar Handwritten Compound Words Recognition and Erratum Detection with MICR

Dr. Yadana Thein
University of Computer Studies, Yangon (UCSY)
3 (B), Yankin Tsp., Yangon, Myanmar

San Su Su Yee
University of Computer Studies, Yangon (UCSY)
312 (B), SouthOkkalapa Tsp., Yangon, Myanmar

ABSTRACT
This paper describes an effective recognition and detection erratum approach for Myanmar handwritten compound words. In this article, ***MICR*** (Myanmar Intelligent Character Recognition) method is used for the character recognition. The method is composed of statistical/semantic information and the final decision is made by the voting system. MICR has been successfully applied for many applications in online Myanmar character recognition field. The method recognizes isolated characters only and not compound words or phrases.

Erratum Detection is a new technique, it detected irregularly form of isolated compound word in string texts. It dependent on the language set of string substitutions reflects the surface form of errors that result from cognitive, typographical mistakes, or mistyping. A robust erratum detection technique is needed to cover above all situation. The system index pair possible extended/medial code and then provides a pair code from the database for detect complete compound word. In detection, it has three situations: twice the same extended, extended/ medial pair not matching error, not compound word in real. Then the final output, Myanmar compound words will be produced editable text with highlight color in each error word.

General Terms
Character recognition, Image Processing, ICR (Intelligent Character Recognition), OCR (Optical Character Recognition), Statistical, Semantic

Keywords
MICR (Myanmar Intelligent Character Recognition), Erratum Detection, Cognitive mistakes, Typographical mistakes, Mistyping

1. INTRODUCTION
In all over the world, there are different techniques that can be used to recognize characters. Among them, Optical Character Recognition and Intelligent Character Recognition are two basic techniques for character recognition. OCR typically involves the process of translating digitized images of text (usually created by a scanner) into a machine-readable format (such as ASCII or Unicode). But Myanmar characters and digits are round shapes in nature and have similar forms so that OCR occur error such as misrecognition, inconvenient, etc. ICR can successfully overcome these problems.

Myanmar Intelligent Character Recognition (MICR) is a technique based on ICR. High speed recognition rates can be gained by using MICR. It can recognize both type-face and handwritten characters. It is used to recognize effectively hand-printed characters.

Writing is very important because it represent the language. In the world, many countries have their own language and native language writing system. The concepts of writing errors are a fuzzy one. The errors others make in Myanmar writing differ according to the characteristics of other language. In erratum detection system, needs to detect each complete compound word.

Familiar erratum detection approaches are often based on language knowledge, and mainly include rule-based method. Rule-based methods use rule sets, which describe some exact dictionary knowledge such as word or character frequency, etc.

Myanmar characters are complexity and widely can be seen in this observation that two of the most common reasons for miswriting are (1) the difference between writing representation and phonetic utterances and (2) phonetic similarity of Myanmar characters. So, we using more knowledge from language itself are required to develop Natural Language Processing.

The remainder of the paper is structured as follows: section 2 gives the background history of our mother language and characteristics of compound word. In section 3, we show up the implementation of the system and we give explanation for MICR method. In section 4, expresses about erratum detection system. In section 5, show the output. Experimental results and conclusion are in section 6 and 7, respectively.

2. HISTORY OF MYANMAR LANGUAGE
The Myanmar language belongs to the Sino-Tibetan family of languages of which the Tibetan-Myanmar (Tibeto-Burman) subfamily forms a part. It has been classified by linguists as a monosyllabic or isolating language with agglutinative features. It is a tonal and analytic language. There are different types of language in Myanmar such as Myanmar, Karen, Rakhine, Chin, Mon, Shan, etc. But, Myanmar language is the mother language in Myanmar.

The Myanmar language is the official language of Myanmar and is more than one thousand years old. Texts in the Myanmar language use the Myanmar script, which derives from a Brahmi-related script borrowed from South India in about the eight century for the Mon language. The first inscription in Burmese dates from the following years and is written in an alphabet almost identical with Mon inscriptions. The earliest Myanmar and Mon language can be seen in MyaZeDi Stone inscription.

2.1 Myanmar Language Characteristics

Myanmar alphabet consists of 33 consonants, 12 vowels, 4 medials and 10 digits. In Pali alphabet consists of 41 letters: (8) vowels and (33) consonants. The consonants in Pali can be grouped into aspirated and non-aspirated consonants, as shown in Figure 1.

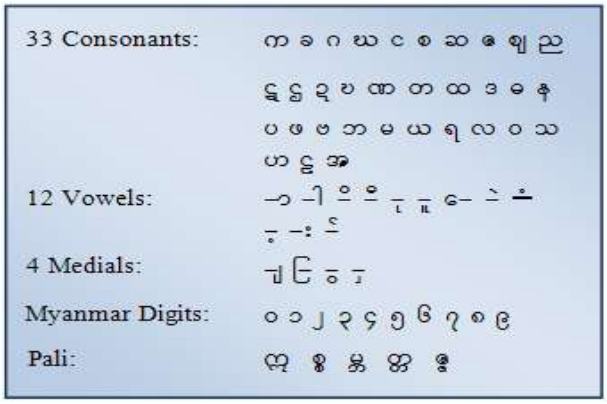

Figure 1. Patterns of Myanmar alphabet

Fortunately, our study found out that standardization and latinization of Myanmar document styles would be a lot easier than expected, because Myanmar writing direction, just like Latin-based scripts, goes horizontally from left to right, then top to bottom.

Some Myanmar characters can stand only one (u? c? r) or combined with other extended characters to become meaningful words (udk? awmf). Myanmar script is written from left to right, as shown in Figure 2. The rounder forms were without tearing the writing surface of the leaf. There are no spaces between words or between syllables, although informed writing developed to permit writing on palm leaves often contains spaces after each clause.

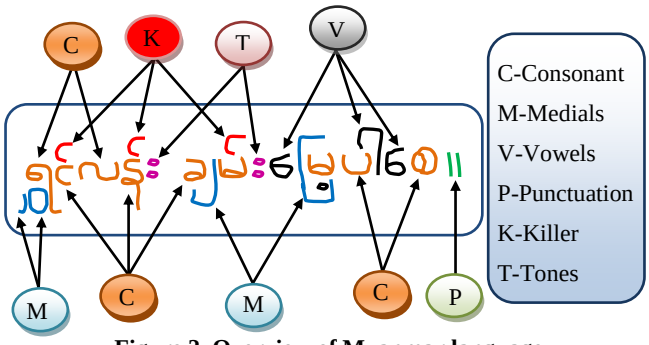

Figure 2. Overview of Myanmar language

Myanmar syllable can be composed of multiple characters. Each consisting of two or more stems joined together is known as a compound word. Myanmar writing language can be distinguished mainly seven kinds of compound words, as shown in Table 1. In this proposed system, it can recognize and detect all kinds of compound words.

Table 1. Structure of compound word

Symbols of Some Compound Words		
No.	**Compound Words**	**No. of characters**
1	pm? cg? ul? yk? aZ? As? jr? vS	2
2	"m;? [D;? tdk? 0wf? rl? oQ	3
3	aqmf? zdk;? 0d*f? aoOf? *gwf? ‘'def	4
4	vQdK? ausmh? Nidrf? qdkif? ayguf	5
5	AsdKuf? a[mif;? tdkif;? ajAmif	6
6	aESmifh? avQmuf? ajr§muf? aoQmif	7
7	ajr§mifh? ajr§mif;	8

3. SYSTEM FRAMEWORK

The basic architecture of the proposed system in this paper is shown in Figure 3. In this system includes five stages: Data acquisition, Pre-processing, MICR method, Erratum detection and Output.

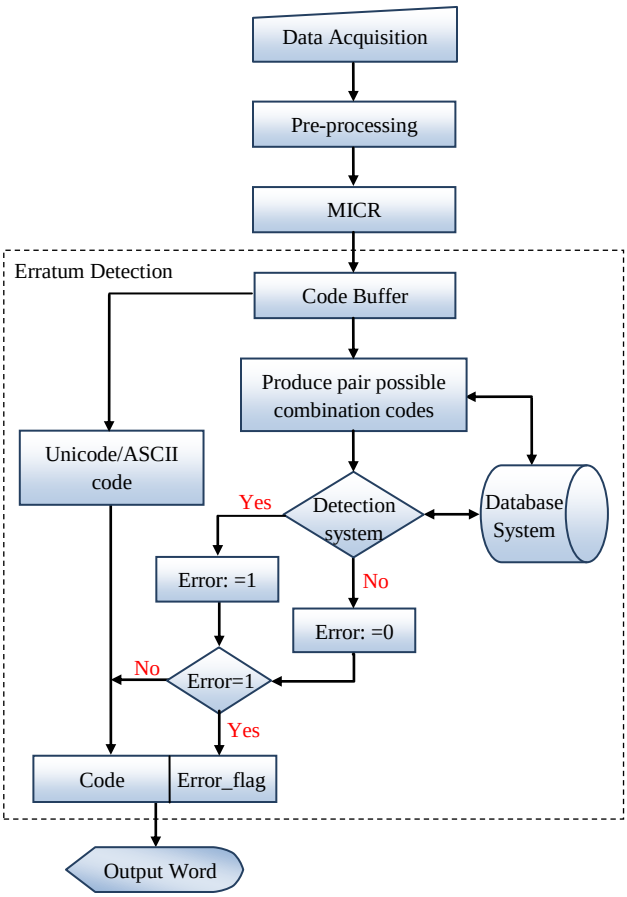

Figure 3. Proposed system design

Two different types of data input method: online and offline. In the stage of Data Acquisition, the proposed system can handle on only online data input by users. Isolated characters are needed to process the image.

Various preprocessing operation are: Gray Scale Converting, Noise Filtering, Binarization and Extraction. Firstly, convert the incoming original image into gray level image and then filtering the noise of the image result from gray scale conversion of image. If conversion of a gray-scale image into a binary image, we extract row and column for each character recognition. And then, labeling scheme is used in this system for the one character lonely.

3.1 MICR

MICR (Myanmar Intelligent Character Recognition) system is based on ICR (Intelligent Character Recognition). MICR was trained in both typeface and handwritten characters, so it can recognize both online and offline characters. But it is more convenient for noise free images and isolated characters to improve accuracy rate. This system used statistical and semantic approach to collect information. That information includes the data of width and height ratio, horizontal and vertical black stroke count, number of loops, end point, open direction, histogram values and character type, etc.

After collecting this required information for each character, we put them on the properties array to record them. Properties of each character are compared with Pre-Defined Database: Basic characters (B-database), Extended characters (E-database), Medials (M-database). When the incoming character matches with the database, the voting system is used to make the final decision of the image on that information (see Figure 4). Then, the output code numbers are stored in the code buffer.

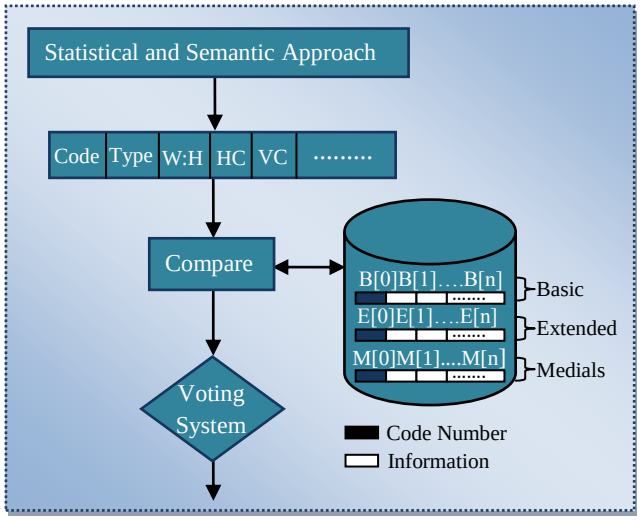

Figure 4. Basic architecture of the MICR

3.1.1 *Statistical and Semantic Approach*

A statistical approach looks for a typical spatial distribution of the pixel values that characterize each character. It is searching for the statistical characteristics of various characters. These characteristics could be very simple, like the ratio of black pixels to white pixels, width and height ratio, histogram, etc.

Some of handwritten characters indeed consist of pixels. Statistical methods ignore is that the pixels also form lines and contours. A semantic approach recognizes the way in which the contours of the characters are reflected in the pixels that represent them and try to find out typical characteristics for each character. Semantic data: black stroke count, loop, open, end point, etc.

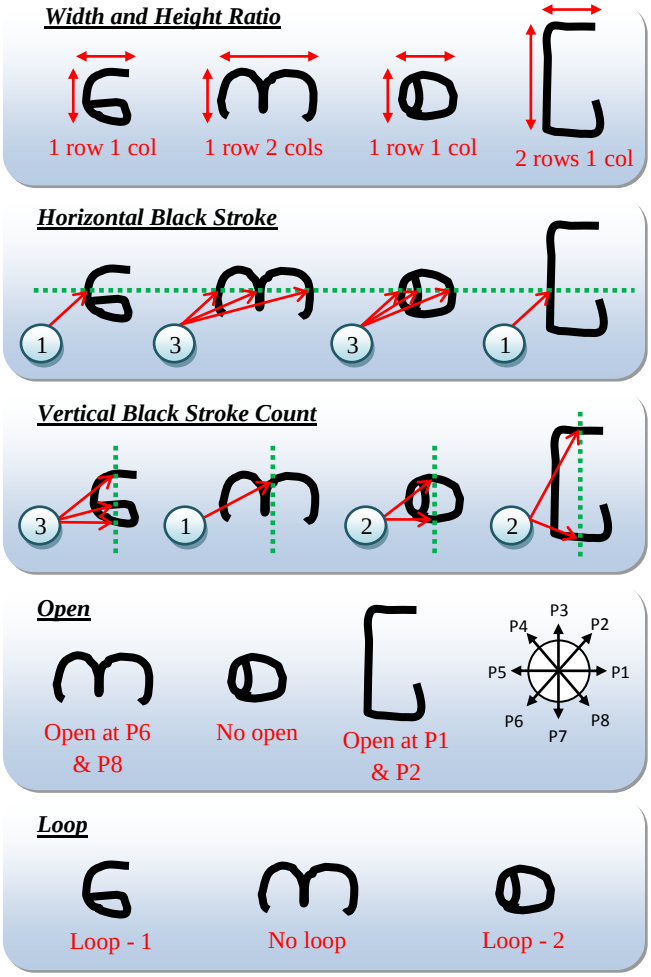

Figure 5. Statistical and semantic information

3.1.2 *Previous Application of MICR*

MICR has been successfully applied in a lot of application such as:

Using MICR

- Speed limited road signs recognition
- Car license plate reader
- Recognition of Myanmar basic characters and compound words(vowels)
- On-line Handwritten Myanmar Pali Character recognition
- Online Myanmar Medial Hand-Printed Characters Into Machine Editable Text
- Handwritten English Characters to Machine editable text by applying MICR
- Converting Myanmar Portable Document format to machine editable text with format

Using both MICR and MVM (Myanmar Voice Mixer)

- Voice production of Handwritten Myanmar Compound Words
- Enhancing the Myanmar Pali Recognition based on MVM

4. ERRATUM DETECTION SYSTEM

Firstly, "why erratum (printing or typing error) is becoming" is presented. Erratum can be of many types, such as typographical error, cognitive error, etc. Myanmar language writing breaks down possible human typing errors into two classes, typographical error and cognitive errors. Typographical errors (e.g., misspelling 'eef' instead of 'eef;') generally occur due to people's mistakes while typing. Cognitive errors (e.g., misspelling 'AdGKif' instead of 'bdGKif') are caused by writers who do not know how to spell the word.

In Myanmar syllable structure, syllables or compound words are formed by consonants combining with vowels or medials. However, some syllables can be formed by just consonants, without any vowel (e.g., rr 00). Myanmar writing language can be distinguished into two parts: actual writing language and general writing language, as shown in Figure 6. Actual writing language, it has 1864 words for all combination of consonants and extended characters in Myanmar Reference Spelling Book as real compound words. In general writing language, there are lots of words which are described in published Myanmar Dictionaries. Some of compound words are not exist in Actual writing because they has been adapted from other language. It has much type of words: adaption words, phonetic tone words, dialect words, etc. In this paper, the system can detect all form.

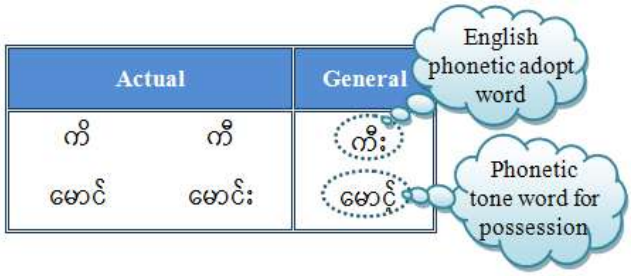

Figure 6. Example words of Myanmar writing language

After the MICR method, the next step is to detect erratum the incoming compound words. This system consists of three main parts: Produce Possible Pair Codes, Detection System, and Database System.

4.1 Produce Possible combination Pair Code

The architecture of the produce pair codes design as shown in Figure 7. In this part, the code numbers of compound words has been rearranged because it needs to index the consonant code number to produce pair code for extended/medial codes.

Then, extract the extended/medial code. According the sequential code numbers result that got this stage, possible pair code will be produced. By producing pair code number, it uses possible combination of extended/medial database. The possible combination pairs are two to seven characters. There are (6) rules that need to follow by producing possible pair code for each compound words. Figure 8 is illustrated by using (4) rules for (7) characters compound words pair.

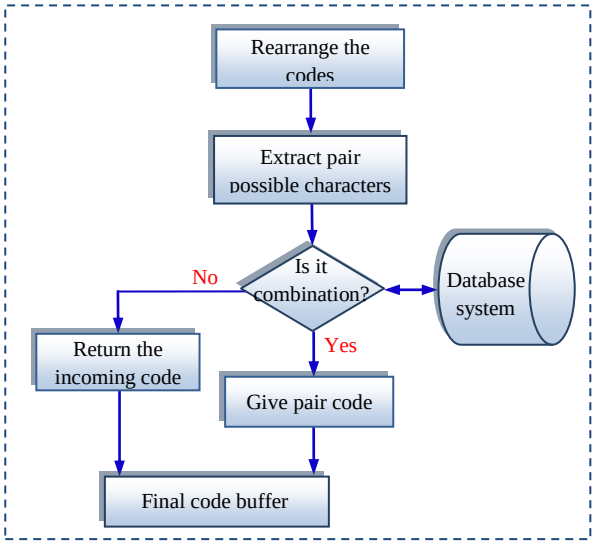

Figure 7. Producing pair code design

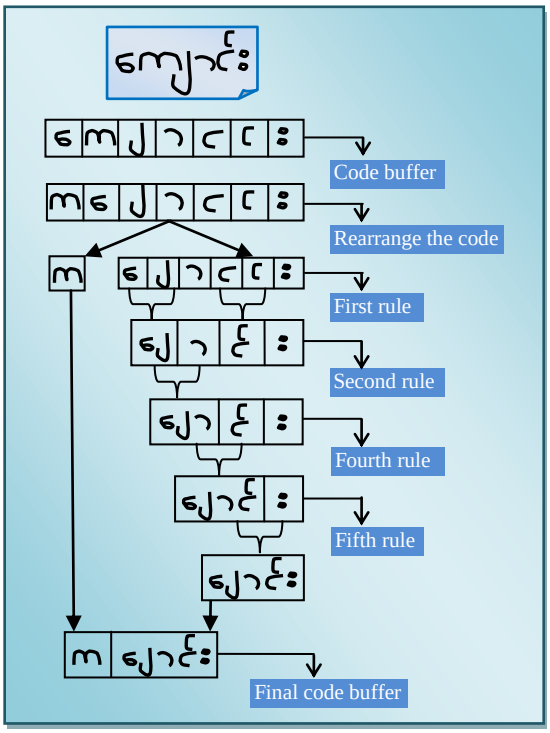

Figure 8. Step by step execution by using 4-Rules for 7 characters compound word

4.1.1 First Rule

In this rule, it indexes the first pair extended/medials characters codes (e.g., a-s / a-: / -dk) or consonant/ extended character code (e.g., rf / if / of / [f) to produce pair code. These codes are compared the database. If these codes are really combined in the writing system, the pair code is produced for error detection. When the two codes aren't combined, the system returned the incoming two codes. And then, these codes are stored in the final code buffer.

4.1.2 Executing another Rules

It indexes the previous rule produce pair code and next extended/medial code and combines them to perform new pair code.

4.2 Detection System

In this system, it detects three situations: twice the same extended error, extended/medial pair not matching error, not compound word in real.

4.2.1 Twice the Same Extended Error

This error is performed when the writer producing text containing the same extended character or medial character by twice. But, this system allows the consonant twice the same, as shown in Figure 9.

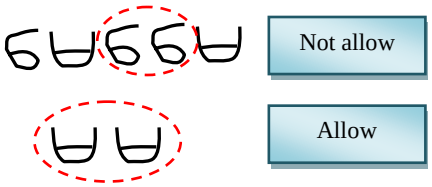

Figure 9. Detect error for twice the same extended error

4.2.2 Extended/Medial Pair not Matching Error

When the extended or medial character can't combine each other, this error is happen, as shown in Figure 10.

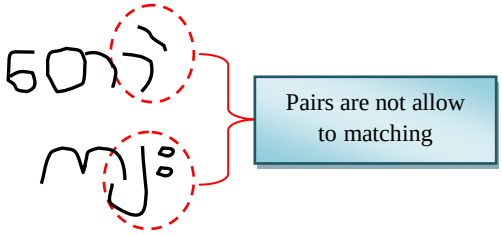

Figure 10. Detect error for pair not matching error

4.2.3 Not Real Word Error

Humans often make errors during communication, in either spoken or written language. In this error includes typographical errors and cognitive errors. The typographical errors involve regular forms of mistyping rather than cognitive errors. Some compound words not exist in language writing system, as shown in Figure 11.

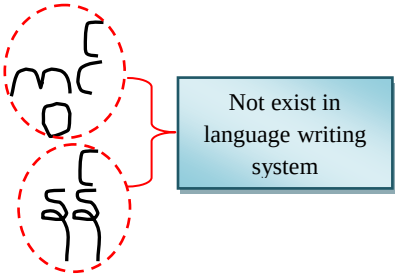

Figure 11. Detect error for not real compound word error

4.3 Database System

In this system, it collects all of the databases for produce pair code and detection system. Produce pair code system, it is used (6) databases for (6) rules. In Figure 12, it shows by using two possible pair code database to produce pair code. In detection system, it uses (7) databases for each writing system (actual writing and general writing) to compare compound word. They are two characters compound words, three characters compound words, four characters compound words, five characters compound words, six characters compound words, seven characters compound words and eight characters compound words.

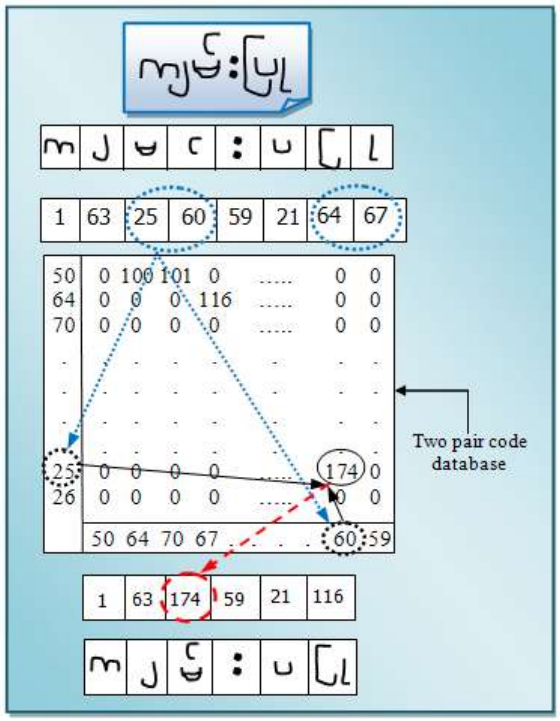

Figure 12. Produce pair code by using two possible pair code database

5. OUTPUT

After that, the recognized combined words are produced as output. This output can be shown in the Microsoft Word file as the editable text format and incorrect compound words with color.

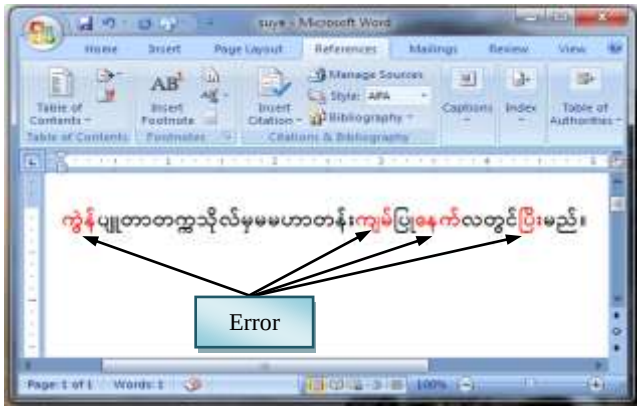

Figure 13. Output data with color

6. EXPERIMENTAL RESULT

In this paper, MICR method is used for the handwritten recognition. But MICR can recognize not only handwritten but also type-face. By using MICR method, we can achieve the high accuracy rates. The performance of the detection is totally depending on the MICR.

Table 2. Recognition and detection result for type-face character

Type-face characters		
Sample	Recognition accuracy rate	Erratum detection accuracy rate
10 words	100%	99.10%
30 words	99%	97.80%
50 words	98%	96%
Over 50 words	97%	94%

Table 3. Recognition and detection result for handwritten character

Handwritten characters		
Sample	Recognition accuracy rate	Erratum detection accuracy rate
10 words	98%	97%
30 words	96.10%	95%
50 words	94%	92.60%
Over 50 words	90%	88.60%

Table 2 and 3 also shows the erratum detection accuracy rate for type-face and handwritten characters. In this system, it is more prefer detects online handwritten character than type-face character. And then, it can detect two types of writing system (Actual writing and General writing), as shown in Figure 14.

7. CONCLUSION

This paper represented the erratum detection of Myanmar handwritten compound words applying MICR and Detection system for all compound words. The character recognition MICR method was successfully developed for Myanmar characters and was found to perform reasonable well with sufficient accuracy. Sometimes, system may misrecognize because of the similarity of Myanmar character (e.g., p and x). But, there is a minor error. Erratum detection system is a new contribution to research area without using other references. In this paper, the system can detect irregular form of all compound words in Myanmar writing system.

8. ACKNOWLEDGMENTS

The authors would like to thank all participants in MICR (Myanmar Intelligent Character Recognition) research field from the University of Computer Studies in Yangon.

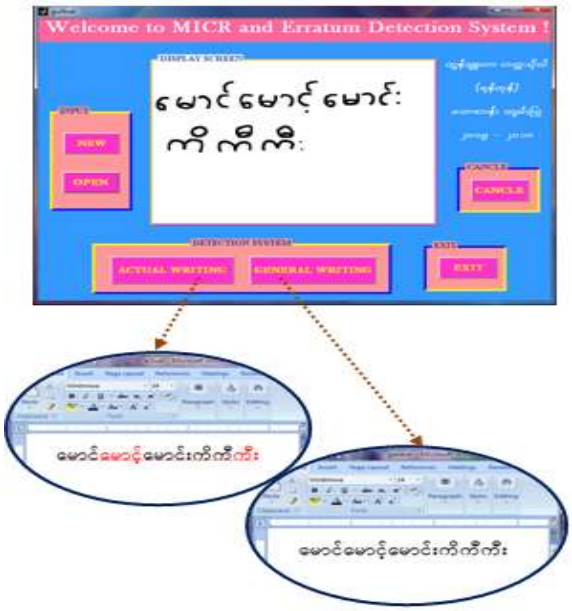

Figure 14. Editable text output with erratum detection for actual writing and general writing

9. REFERENCES

[1] E.E.Phyu, Z.C.Aye, E.P.Khaing, Y.Thein and M.M.Sein, "Recognition of Myanmar Handwritten Compound Words based on MICR", the 29th Asian Conference on Remote Sensing (ACRS), Colombo, Sri Lanka, 2008

[2] Ei Theingi, Ei Kay Khine, Thu Wai Kyaw Kyaw, Dr.Yadana Thein, "Enhancing the handwritten Myanmar characters recognition system for pali", 30th Asian Conference on Remote Sensing(ACRS),China, 2009

[3] Gerhard B. van Huyssteen, Menno M. van Zaanen, "Learning Compound Boundaries for Afrikaans Spelling Checking", North-West University (South Africa) & University of Tilburg (The Netherlands), Centre for Text Technology, North-West University, Potchefstroom, 2531, South Africa

[4] Naushad UzZaman, Mumit Khan, "A Comprehensive Bangle Spelling Checker" , Center for Research on Bangla BRAC University, Bangladesh

[5] Tun Thura Thet; Jin-Cheon Na, Wanna Ko Ko, "Word Segmentation of Myanmar Language", Journal of Information Science JIS, 2nd October,2007

[6] Yin Mon Aung, Ei Kay Khine, Khaing Wai Myo, Dr. Yadana Thein, "Writer Independent Online Myanmar Medial Hand-Printed into Machine Editable Text", 30th Asian Conference on Remote Sensing (ACRS), China, 2009

[7] Zaw HTUT (Mr.), "Features of Myanmar Language Document Styles", Executive Committee Member, MCSA Myanmar Computer Federation (MCF)

[8] Zar Chi Aye, Ei Ei Phyu, Yadana Thein and Myint Myint Sein, "Intelligent Character Recognition (MICR) and Myanmar Voice Mixer (MVM) System", the 29th Asian Conference on Remote Sensing (ACRS), Colombo, Sri Lanka, 2008.