

FSA based Code Sequence Checking to Prevent Mal Use of Myanmar IDNs

Tin Htay Hlaing¹, and Yoshiki MIKAMI²

¹ Management and Information Systems Department, Nagaoka University of Technology
Nagaoka, JAPAN

² Management and Information Systems Department, Nagaoka University of Technology
Nagaoka, JAPAN

Abstract

With the development of new technologies, not only online contents but also domain names can be represented in local languages and this makes human society with better communication, better education and better business. Though domain names in some Asian languages such as Japanese, Indian languages, Sinhala and Urdu, are already implemented, not many works have been attempted on domain names in Myanmar language. Thus, the first part of this study aims to discuss fundamental issues for implementing Myanmar domain names by stating language-specific characteristics. Secondly, Myanmar language has different types of combining marks, similar-looking characters and homoglyphs which can open phishing attacks. Thus, possible spoofing attacks are addressed and finite state automata (FSA) based coded sequence checking method is proposed to prevent combining mark order spoofing in Myanmar language. Our tested results on Stuttgart FST tool (SFST) show that the proposed FSA could check code sequence order of Myanmar characters correctly. Though we use Myanmar script as an example, we expect that this approach will work properly for other Asian scripts.

Keywords: *Internationalized domain names, Myanmar domain names, finite state automata, homoglyph attack, spoofing.*

1. Introduction

With only availability of ASCII encoding, Latin script users could get the accessibility of the WWW and reach to the cyber community firstly. As compared to penetration of Internet by Latin script users, penetration of internet especially in Asian regions, as a percentage of population, is very low, i.e, 18.5% of Asian population as of year 2009. Among many possible reasons, language barrier to online access could be a major reason. However, global community has been working hard to represent non-Latin scripts in the computer systems to expand the cyber community. And, as a result, many scripts including Myanmar can be represented in the computer systems and it is able to post all web contents in our local languages

now-a-days. As of June 2012, penetration rate of Asia becomes 27.5%¹.

However, it still requires the knowledge of Latin script because the addresses on the Internet are still using Latin characters known as LDH, Letters a..z, Digits 0..9 and Hyphen. Thus, finally, Internationalized Domain Names (IDNs) are introduced for complete localization. IDNs mean web addresses represented by local language characters which enable more web users to navigate the Internet in their preferred scripts. For example, <http://www.parliament.lk>, the home page for the Parliament of Sri Lanka.

Among Asian countries, Myanmar having population of 54millions, has only 1% of penetration of Internet as of 2012 and this result shows more efforts and attempts are necessary to be carried out for the IT development. Implementation of domain names in Myanmar language has become a necessity in order to popularize the use of Internet among the rural masses of Myanmar because the majority of 50 million population of Myanmar count Myanmar language (Burmese) as their first language. Thus, as a first part in this study, linguistic issues to be considered for the implementation of Myanmar domain names are proposed.

The introduction of what are called internationalized Domain Names (IDNs) amplifies both the difficulty of putting names into identifiers and the confusion that exists between scripts and languages. The introduction of Unicode support in operating systems and applications has lead to a vastly increased number of available homoglyphs and a new threat arose from the use of characters which are visually indistinguishable from western characters but belong to a non western script (ICANN, 2005) [1]. Given the added complications of using a much broader range of characters than the original small ASCII subset, precautions are necessary in the deployment of IDNs in order to minimize confusion and fraud.² Considering that

¹ <http://www.internetworldstats.com/asia.htm>

² <http://www.ietf.org/rfc/rfc3743.txt>

a large number of users are not scholars of the language and hence can be easily cheated by homographs, and spoofing, and phishing will occur to a large extent in languages. This calls for great care and caution in supporting local languages and scripts in the domain names [2]. Unicode consortium also listed seven possible spoofings which are described as visual security issues namely

1. Mixed-script Spoofing
2. Single-script Spoofing
3. Whole-script Spoofing
4. Inadequate Rendering Support
5. Bidirectional Text Spoofing
6. Syntax Spoofing
7. Numeric Spoofs

and mitigating methods for IDNs are stated because Unicode contains such a large number of characters, and incorporates the varied writing systems of the world, incorrect usage can expose programs or systems to possible security attacks [3]. Likewise, Internet Corporation for Assigned Names and Numbers, ICANN, established the guidelines which are a list of general standards for IDN registration policies and practices that are designed to minimize the risk of cybersquatting and consumer confusion, and respect the interests of local languages and character sets.

Phishing attacks can also occur to Myanmar domain names as Myanmar script has similar looking characters and different ways of combination of characters (code sequences) within a syllable. Thus, the main purpose of this paper is to address potential threats for phishing with Myanmar IDNs and proposed finite state method for combining mark order spoofing to make Myanmar domain name strings as secure as possible. No published work or research has been found for mitigating method for mal use of Myanmar IDNs using the Unicode standard. Without such solution, Myanmar IDNs are highly vulnerable for implementation and applications of IDNs.

The rest of the paper is organized in five sections. Section 2 will discuss problem statement which mentioned possible spoofing for Myanmar language. Section 3 covers linguistic issues to implement Myanmar domain names. Proposed FSA based approach is explained in section 4 and experimental results, conclusion and some points for discussion are in section 5.

2. Problem statement

After internationalized domain names (IDNs) had been introduced in 2007, possible threats on IDNs have been discussed and also, guidelines and recommendations for IDN based attacks are developed as mitigating strategies by international organizations.

Though general guidelines and mitigating methods are stated to cover IDN implementation in all scripts, there are

some scripts like Myanmar, which need language specific mitigation methods. Thus in this section, Myanmar homoglyphs and possible spoofings for Myanmar domain names are described and types of spoofing which need language specific checking method for Myanmar is highlighted.

2.1 Myanmar Homoglyphs

Myanmar script or Burmese is a phonologically based script, adapted from Mon and ultimately based on an Indian (Brahmi) prototype. It is a syllabic script and thus words are composed of one or more syllables. Each syllable can be stand-alone syllable or composed of up to five sub-syllabic elements namely consonant, medial, vowel, asat and tone mark. Among them, consonant is a basic element and other four elements can be attached to it in different combinations.

Having look-alike characters in any script are prone to phishing attacks especially when they are displayed in a default address bar of browsers. A careful visual inspection of Myanmar letters shows that some letters open for visual spoofing because of their visual similarities. Our proposed Myanmar homoglyphs, look-alike consonants, and digits are listed in the following tables.

Table 1. Myanmar look-alike digits

No.	Glyph	Unicode Value	Name
1-a	၆	U+1046	Myanmar Digit Six
1-b	၉	U+1049	Myanmar Digit Nine

Table 2. Myanmar homoglyphs

No.	Glyph	Unicode Value	Name
2-a	၀	U+101D	Myanmar letter WA
2-b	၀	U+1040	Myanmar Digit Zero

Table 3. Myanmar look-alike consonants

No.	Glyph	Unicode Value	Name
3-a	က	U+1000	Myanmar letter KA
3-b	ခ	U+1018	Myanmar letter BA
3-c	ဂ	U+101A	Myanmar letter YA
3-d	ဃ	U+101E	Myanmar letter SA
3-e	င	U+101F	Myanmar letter HA
4-a	၇	U+101B	Myanmar letter YA
4-b	၇	U+1047	Myanmar Digit Seven
5-a	ဉ	U + 1025	Myanmar Vowel U
5-b	ဉ	U+1009	Myanmar letter NYA
6-a	တ	U + 1010	Myanmar letter TA
6-b	တ	U + 1011	Myanmar letter THA

2.2 Possible spoofings for Myanmar

Like other scripts, Myanmar script is highly vulnerable for different types of spoofing and some spoofing cannot be solved by using general guidelines and procedure as summarized in the table 4.

Table 4. Summary of types of spoofing, existing mitigation methods and threats for Myanmar IDNs

Types of Spoofing	Description	Existing Mitigating Methods	Possible Threats in Myanmar IDNs
Mixed-script spoofing	the existence of visually confusable characters across scripts	Mixed-script spoofing detection procedure in Unicode Security Mechanisms	“o” (U+03BF – Greek omicron), “o” (U+006F – Latin small letter o), “o” (U+043E - Cyrillic small letter o) “o” (U+101D - Myanmar letter WA) and “o” (U+1040 - Myanmar Digit Zero)
Single-script spoofing	Spoofing with characters entirely within one script, or using characters that are common across scripts	IDN implementation guidelines by ICANN	“o” (U+101D - Myanmar letter WA) and “o” (U+1040 - Myanmar Digit Zero) can lead to single-script spoofing.
Combining mark order spoofing	Spoofing by reordering of character	Not established yet as it is language specific issue	Highly vulnerable and need language specific checking method (please refer to table 5)
Inadequate rendering support	a font or rendering engine has inadequate support for characters or sequences of characters	it is mentioned in Unicode Security Considerations	Myanmar Unicode characters (U+1039 – Myanmar Sign Virama) is invisible, and it may affect the rendering of the characters around them.
Bidirectional text spoofing	visually confusable characters obtained by mixing inherent right-to-left and left-to-right writing directions	Unicode bidirectional Algorithm	Impossible
Syntax spoofing	Spoofing syntax characters eg. U+2044 (/) FRACTION SLASH can look like a regular ASCII '/' in many fonts	visual distinguishing is suggested in Unicode Security Consideration.	Impossible
Numeric spoofing	Individual digits may have the same shapes as digits from other scripts, even digits of different values.	IDN implementation guidelines by ICANN	“o” (U+1040 - Myanmar Digit Zero) has visual similarity with Basic Latin digit zero (U+0030).

In most cases, two sequences of accents that have the same visual appearance are put into a canonical order. This does not happen, however, for certain scripts in Southeast Asia, so reordering characters may be used for spoofs in those cases [3] as shown in Table 5.

Table 5. Combining Mark Order Spoofing in Myanmar

String	Unicode sequence	Punycode
ꠊꠎ	U+101C <u>U+102D</u> U+102F ꠊ ꠎ ꠏ	xn--gjd8ag.com
ꠊꠎ	U+101C U+102F <u>U+102D</u> ꠊ ꠏ ꠎ	xn--gjd8af.com

For Myanmar language combining diacritic marks shown in the above table, visual appearances of the given two strings are same though they have different underlying coded sequences and so as their punycode values. Current example is the dependent vowels combinations within a syllable and likewise, other Myanmar vowels and medial consonants can bring such kind of inconsistent ordering. Therefore, it is necessary to check underlying code sequence and propose an efficient code checking method to check such kind of language specific problem.

Moreover, complex and traditional writing styles of Myanmar language such as kinzi, consonant stacking, consonant repetition, and contractions could lead to combining mark order spoofing.

To summarize, if the registries and registrars follow the proposed guidelines by ICANN and Unicode which are discussed in section 2.1, potential threats based on homoglyphs can be greatly reduced. In other words, IDNs do not materially increase risks related to phishing. However, as there are still some attacks based on homographs, various mitigating methods are being developed for safe domain names. Nevertheless, detection mechanism for combining mark order spoofing for Myanmar has not yet been described in the previously mentioned standards and guidelines by Unicode and ICANN. Up to our knowledge, this kind of issue is language-specific issue and should be managed by respective language authorities. Therefore, the main purpose of this study is to express possible threats in Myanmar domain names including combining mark order spoofing and propose finite state automata (FSA) based coded sequence checking to prevent mal use of Myanmar domain names.

3. Myanmar Domain Names: Linguistic Issues

3.1 Encoding and Myanmar Character Set

Myanmar language, also known as Burmese is spoken by major ethnic group, Bamar as well as by other ethnic groups as their first language. And, Burmese or Myanmar language is the official language of Union of Myanmar. Both Burmese and other ethnic languages have their own scripts which have been encoded in Unicode. However, other ethnic scripts use some characters in Myanmar script. Myanmar script is abugida in Brahmi family used for writing Myanmar language and it is written from left to right without space between words [4].

Though different encodings are available for Myanmar script, the authors propose to use Unicode standard encoding for Myanmar domain names.

Myanmar character set has been standardized under Unicode with the range from U+1000 to U+109F. The Unicode table comprises different types of Myanmar letter namely (1) consonants (U+1000 to U+1020), (2) independent vowels (U+1021 to U+102A), (3) dependent vowel signs (U+102B to U+1035), (4) various signs (U+1036 to U+103A) (we refer to it as various sign group I), (5) dependent consonant signs (U+103B to U+103E), (6) digits (U+1040 to U+1049) and (7) various signs (U+104C to U+104F) (We refer to it as various sign group II). And some of these characters are shown in Table 1.

Table 6. Myanmar Unicode Character Set

No.	Type	Characters
1.	Consonants	က, ခ, ဂ, ဃ, င, စ, ဆ,...
2.	Independent Vowels/ Free standing Vowel Syllables	အ, ဣ, ဤ, ဥ, ဦ,...
3.	Dependent Vowels	ိ, ဝ, ဝိ, ဝီ, ဝု, ဝူ,...
4.	Various Signs I	့, ္,...
5.	Dependent Consonant Signs/ Medials	ျ, ြ, ျ, ြ
6.	Digits	၀, ၁, ၂, ၃, ၄,...
7.	Various Signs II	ံ, ဴ, ဵ, ံ

Though Myanmar letter A (U+1021) includes in the independent vowel group in the Unicode character set, it can be used as a glottal stop as consonant. It is listed 8 independent vowels in the Unicode character set, total 12 vowels are used in the language as some coded vowels can be combined to form the new vowels. Likewise, only 4 medial consonants are coded in the character set but total 11 medial consonants are used in the language. And details of such combinations can be seen in [5].

3.2 Allowed and Disallowed Characters

Unicode being a script based encoding standard groups all letters across all languages which use the same script. Thus, language specific conventions need to be given for controlling which characters may be allowed within and across scripts for a particular language. According to ICANN's PVALID characters in IDNA2008, except punctuation marks and various signs group II, all encoded characters for Myanmar script are allowed for IDN labels. These code points are summarized in the following table.

Table 7. ICANN's Myanmar character in IDNA2008

Unicode Code	Type	Description
U+1000 ~ U+1049	PVALID	Myanmar letter KA ~ Myanmar Digit Nine
U+104A ~ U+104F	DISALLOWED	Myanmar letter Sign Little Section ~ Myanmar Symbol Genitive
U+1050 ~ U+109D	PVALID	Letter Shan ~ Vowel Sign Aton AI
U+109E, U+109F	DISALLOWED	Myanmar Symbol Shan One, Shan Exclamation

IANA also maintains a collection of IDN tables which represent permitted code points allowed for Internationalised Domain Name registrations in particular registries [5]. For TLDs such as .COM, .NET, .NAME in Myanmar script supported by Verisign, allowable and disallowable Unicode points for Myanmar script are described the same.

In [6], character set selection using IDNAbis table and feedback from native speakers of 8 local language teams namely Bengali, Dzongkha, Khmer, Lao, Mongolian, Nepali, Pashto and Urdu, under PAN localization project are discussed because there are some characters which should be disallowed in respective domain names to reduce the risk of phishing but allowed in IDNAbis table.

Furthermore, allowing Myanmar sign Virama (U+1039) as PVALID character, complex words in Myanmar writing system can be used in Myanmar domain names. The following table shows the coded sequence in which how the Virama sign is used to represent words with complex forms such as consonant stacking, consonant repetition, kinzi and contraction.

Table 8. Myanmar Complex Writing Forms

Name	Example	Unicode Sequence	Code
Consonant repetition	တက္ကသိုလ် (University)	1010 1000 1039 1000 101E 102D 102F 101C 103A	
Consonant Stacking	ကုမ္ပဏီ (company)	1000 102F 1019 1039 1015 100F 102E	
Kinzi	သင်္ဘော (Ship)	101E 1039 1019 1018 1031 102C	
Contraction	သမီး (daughter)	101E 1039 1019 102E 1038	

Allowing all characters except punctuation marks and various signs is sufficient for successful implementation of

Myanmar domain names. One point to be considered is how to render such kind of complex writing form in the browsers` address bar.

3.3 Some Spelling Variants

Some words can be expressed in two different ways and this may cause serious IDN implementation problems. For example, the word “daughter, သမီး” as “သမီး”. Such kind of words should be listed and it is also necessary to establish a policy with regards to words with multiple forms.

3.4 gTLD and ccTLD in Myanmar language

The process namely gLTD translation process plays important role for implementation of Myanmar domain names. Myanmar language should have its own gTLD set and separate namespace but it is more likely to access existing namespace into Myanmar by using direct mapping or translation. The process should be done in the following steps:

1. Collecting Myanmar terms as possible as for English gTLD
2. Language authority and linguistic experts should consult for the best word or usage for naming gTLD
3. Suitable short forms or abbreviation for gTLD should be decided

We also suggest the possible words to use as Myanmar gTLD for direct mapping of some English gTLD as examples.

Table 9. Example of Myanmar gTLDs

English gTLD	Myanmar gTLD suggestions
.com	ကြော်ငြာ
.net	ကွန်ယက်
.info	သတင်း
.biz	စီးပွားရေး
.mob	မိုဘိုင်း
.museum	ပြတိုက်
.gov	အစိုးရ
.edu	ပညာရေး
.mil	စစ်ဘက်
.org	အဖွဲ့အစည်း

Like gTLD translation process, the most appropriate Myanmar words should be selected for Myanmar ccTLDs and these names should be mapped to existing ccTLDs at the client side.

3.5 Label Separators

In IDNA2003, three characters are listed to treat as label separators. In IDNAbis [6], only ASCII period is allowed to be used. If other characters are required for any language, they are expected to be mapped before using or storing the domain name. Some language specific delimiters identified in are as follows.

Language	Character	Unicode Value
Dzongkha	%	U+0F14
Urdu and Pashto	-	U+06D4

Language interface handling these particular languages will handle these delimiters in addition to FULL STOP U+002E. It is noted that all languages did not decide to use their sentence separator marks to be used as label separators. For example, in Nepali U+0964 is used as sentence marker but they decided to use FULL STOP U+002E as label separator in domain names [6]. Therefore, label separator for Myanmar domain names should be decided like other Asian scripts.

To sum up, the above mentioned five steps are crucial and the implementation of these steps could be accomplished by co-operation of technical experts, linguists and language authority.

4. Proposed FSA based Checking Method for Myanmar

In this section, Myanmar character combination sequences are discussed in detail and our proposed FSA based code sequence checking method is explained.

4.1 Combination of sub-syllabic elements within a syllable

Myanmar script is derived from Brahmi script of ancient India and there are other Indian-based scripts such as Sinhala, Bengali, Dzongka, Thai, Khmer and so on. Myanmar script is based on “a-vowel accompanying consonant syllabics”, i.e, this syllabary consists of consonant letters accompanied by an inherent vowel. Since, in this syllabary, a consonant associated with its inherent vowel can indicate a standalone syllable, it would be appropriate to call it a consonant syllable, but we simply call it a consonant letter for simplicity. Thus, only a consonant can be syllable breaking point in orthographic syllabification which means minimal syllable in orthographic view is C which stands for consonant. Basically, a Myanmar syllable can be described as $S = I | N | P | X$ where

S = Syllable
 | = logical operator OR

I = Free-standing vowel syllables/
 Independent Vowels
 N = Digits
 P = Abbreviated syllables
 X = a syllable formed by the combination of up to 5 sub-syllabic groups

An additional complexity to Myanmar syllable structure is that there are syllables (X) containing up to 5 sub-syllabic groups namely consonant (C), medial consonant/dependent consonant sign (M), dependent vowel (V), *Asat* or a vowel killer (K) and tones (D) (which are taken from the group of Various Signs) and these groups can appear in a syllable as one of the following combinations.

These combinations can be described as a regular expression as $X = C M? V? (C K)? D?$

where the symbol “ ? “ stands for 0 or 1 occurrence of the character. In this expression, the combination of consonant and *Asat* or vowel killer (CK) is called final consonant and the syllables end with this combination are known as closed syllables.

Further, as with the multiple-component vowels, the user reads the entire syllable as an entity. In Myanmar script, 3 vowels and 7 medials which are formed by the combination of other vowels or medials defined in the Unicode character chart. And, if we use this characteristics of vowels and medials in writing regular expression, the expression for syllable structure becomes like this $X = C M* V* (CK)? D?$ where the notation ? means 0 or 1 occurrence and * means 0 or more occurrence [5].

4.2 Myanmar combining marks : Vowels and Medials











Myanmar script encoded in Unicode has eight dependent vowels through code values U+102B to U+1032, however, new three vowels obtained by combining some of these individual vowels and Myanmar sign ASAT (U+1039). Two or more Myanmar attached vowels are combined and formed new three members { , ,  } in the vowel set.

Table 10. Vowel combining marks

Glyph	Unicode Values	Description
 + 	U+1031, U+102C	Vowel sign E , Vowel sign AA
 +  + 	U+1031, U+102C,U+103A	Vowel sign E, Vowel sign AA, ASAT
 + 	U+102D, U+102F	Vowel sign I, Vowel sign UU

Similarly, 4 basic Myanmar medials combine each other in some different ways and produce new set of medials { ချွဲ, ငြိ, ချွဲ, ငြိ, ချွဲ, ငြိ, ချွဲ, ငြိ }.

Table 11. Medial combining marks

Glyph	Unicode values	Description
ချွဲ + ငြိ	U+103B + U+ 103D	Consonant Sign Medial YA + WA
ငြိ + ငြိ	U+103C + U+103D	Consonant Sign Medial RA + WA
ချွဲ + ချွဲ	U+103B + U+103E	Consonant Sign Medial YA + HA
ငြိ + ချွဲ	U+103C + U+103E	Consonant Sign Medial RA + HA
ငြိ + ချွဲ	U+103D + U+ 103E	Consonant Sign Medial WA + HA
ချွဲ + ငြိ + ချွဲ	U+103B + U+103D + U+ 103E	Consonant Sign Medial YA+WA + HA
ငြိ + ငြိ + ချွဲ	U+103C + U+103D + U+103E	Consonant Sign Medial YA+WA + HA

4.3 FSA based code sequence checking for Myanmar domain names

Languages can be presented as entities generated by a computation. This is a very common situation in formal language theory: many language families are associated with computing machinery that generates them. The simplest computation device is Finite State Automata (FSA) which can be thought of a finite set of states, connected by a finite number of transitions.

Finite state automata are efficient computational devices for generating regular languages. An equivalent view would be to regard them as recognizing devices. Further, Most of the algorithms one would want to apply to finite-state automata take time proportional to the length of the word being processed, independently of the size of the automaton. In computational terminology, this is called *linear time complexity*, and is as good as things can get [7]. Therefore, many NLP applications use FSA and we also apply FSA for code sequence checking of Myanmar domain name strings to mitigate combining mark order spoofing.

For Myanmar language domain names, it is also necessary to follow the standard procedures established by Unicode, ICANN and IDN working teams. And by doing this, most of the potential attacks can be reduced. For single-script spoofing and mixed-script spoofing, script-level specifications are necessary to define allowable and disallowable characters, identifying similar-looking characters and homoglyphs between Myanmar scripts and other scripts. Further, it is also necessary to develop variant tables and the results can be achieved through discussions between technical and linguistic experts.

For combining mark order spoofing, finite state based approach is proposed here. Our method firstly analyses the orthographic rules for combining mark in Myanmar language and also Myanmar canonical ordering. Secondly, we develop FSA to check correct code sequence of input Myanmar characters according to Myanmar canonical order.

Then, the final step, we suggest to incorporate our FSA in the validation step of IDN registration process.

Finite state diagram for generalized Myanmar domain name is expressed as follows. In the following FSA, state 1 is the initial state and state 6 and 8 are the final state.

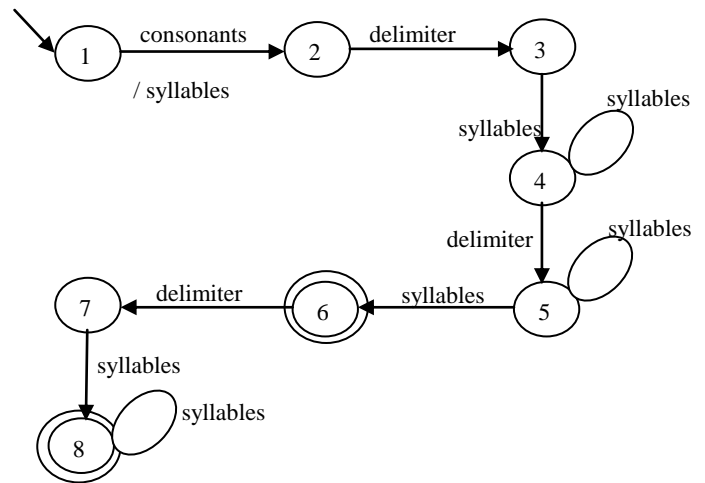


Fig 1. FSA for generalized Myanmar domain names

To check the coded sequence for syllables, we propose another FSA for correct combination of sub-syllabic elements within a syllable based on the regular expression $X = C M^* V^* (CK)? D?$

where X = syllable

C = consonant

M = medial consonant/dependent consonant sign

V = dependent vowel

K = asat or vowel killer

D = tone

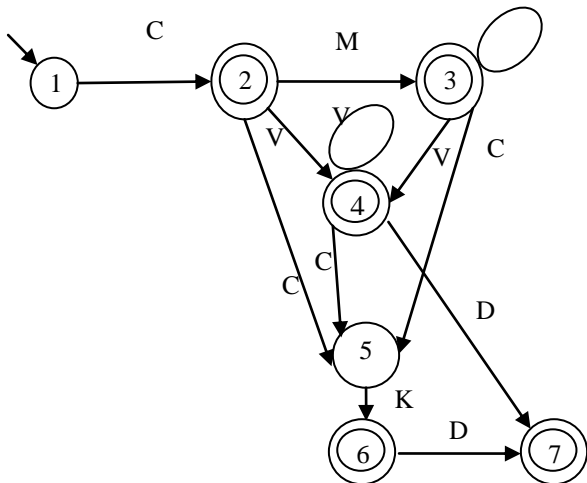


Fig 2. FSA for Myanmar syllable with sub-syllabic element in correct order

It is necessary to include FSA module somewhere in the registration process. There are two registration processes managed at the registrar and the registry. We propose to use our FSA module in the registration process at the registrar. We refer Verisign’s IDN registration process [12] in our example and a brief explanation about the process is given here. A registrant requests an IDN from a registrar that supports IDNs. The registrar converts the local language characters into a sequence of supported letters using an ASCII-compatible encoding (ACE). The registrar submits the ACE string to the Verisign® Shared Registration System (SRS) where it is validated. The IDN is added to the .com and .net TLD zone files and propagated across the Internet. This process is shown in the following figure.

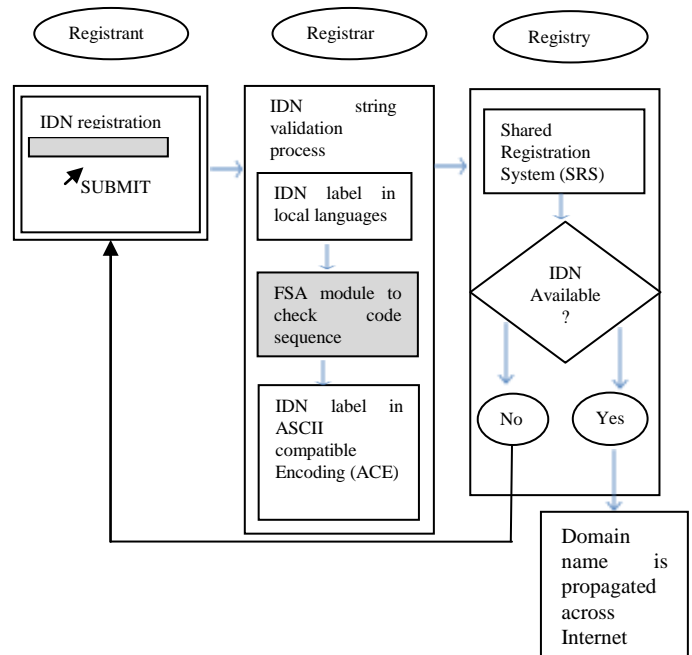


Fig 3. Registration process with proposed FSA module

5. Experimental Results and Discussion

5.1 Preliminary Experiment

To show proper working of our proposed FSA, we set up an experiment to check code sequence order in Myanmar domain names. Most of Myanmar domain names are in English names and so far, no Myanmar IDNs have been implemented. However, some romanized domain names are found, for example, <http://www.zawtika.com/> which means in Myanmar language www.ဇော်တိကာ.com. And, also there are some domain names which are the combination of Myanmar words and English words, for example, <http://www.hlagabarfurniture.com/> in which address, “hlagabar” is Myanmar word “လှူတံဆိပ်” and “furniture” is of course English word. Further, there is no complete web directory for Myanmar yet. For the above difficulties, we could set up a preliminary experiment using 22 active commercial websites with romanized domain names in our current study.

All collected domain names are in romanized form and we convert them into equivalent Myanmar words. Converted words are either in regular syllable structure or complex writing forms such as consonant stacking, consonant repetition, and kinzi. Based on our tested results, it is found that all 22 Myanmar domain names are correctly recognized by our proposed FSA.

5.2 Discussion

IDN is a societal issue as well as technical challenge and it calls for great care and caution in supporting local languages and scripts in domain names. IDN based phishing attacks are surveyed by the APWG as follows. According to the APWG IDN Phishing Report 1H2009, from January 1, 2007 to June 30, 2009 only 85 IDNs were used for phishing. The majority were .HK domain names apparently used by the Rock Phish gang early in 2008. Again in 2H2010 survey, it is stated that since January 2007, only one true homograph attack was found.

According to the global phishing survey (2011) by Anti-Phishing Working group (APWG), only 10 of the 42,624 domain names they studied were IDNs and only one was a homograph attack. In survey 2H2012, they stated that since January 2007, they have found only five homographic phishing attacks, and none since 2011. In July 2012, there were two interesting attacks. They were not homographic attacks, but were malicious IDN registrations [9].

Though a considerable amount of attacks has not been found yet for IDNs, we should prepare safety measures for respective scripts in advance as the number of IDNs are growing obviously and the attacks on IDNs are the threats which we have to face in near future.

In [10], the authors explore the various types of address spoofing attacks focusing on IDN, and presents a novel client-side web browser plug-in Quero to protect the user against visually undistinguishable address manipulations. Likewise, a client-side solution in the form of Firefox plug-in is developed [11].

Also, Unicode technical standard#39 known as Unicode Security Mechanisms [12] specifies mechanisms that can be used to detect possible security problems. For confusable detection, the tables in data files “confusables” provide a mechanism for determining when two strings are visually confusable. By following Unicode security mechanism, homograph attack could be greatly reduced.

For some Asian languages, domain names implementations in their languages have already documented by stating language specific characteristics, for example, Sinhala [13] and Urdu [14]. However, for Myanmar, any kind of attempt for such kind of work has not been documented yet so far. Further, Myanmar use different combining marks and it may bring an additional challenge to implementation of IDNs. And, there has not been provided any mitigating method for combining mark order spoofing which has high potential to happen in Asian scripts. Thus, this could be a risk to Asian script IDN based phishing so far no visible mal use of Asian IDN is reported.

Therefore, our study will cover this gap by introducing code sequences checking module so that cyber threats using IDNs could be reached to a tolerable level and thus

the Internet community would be safe. Secondly, it is reported that most of the syllabic writing systems can be described by using finite state automata (FSA) and we expect other scripts derived from Indic script can be checked by using our proposed method to prevent malicious registration of IDN labels.

Our current work is a preliminary stage and it requires collaboration between language authorities and technical experts for completion. It is expected that this work is just a proposal for Myanmar IDN work which has not been initiated yet.

References

- [1] Hannay, Peter, and Christopher Bolan. (2009) : Assessment of Internationalised Domain Name Homograph Attack Mitigation, In the *Proceedings of Australian Information Security Management Conference*. Perth.
- [2] Department of IT, Ministry of Communication and IT, Government of India. (2009) : Internationalized Domain Names in Indian Languages, A Draft policy document: Policy Framework and Implementation Plan. Available at <http://www.docstoc.com/docs/74240340/India--IDN--Policy> (Accessed June 2013)
- [3] Mark Davis and Michel Suignard. Unicode Security Considerations. Available at: <http://www.unicode.org/reports/tr36/> (accessed March 2013)
- [4] Peter T. Daniels, William B. *The World Writing Systems*, the second Edition. Oxford University Press.
- [5] Tin Htay Hlaing. (2010) : Manually Constructed Context-Free Grammar for Myanmar Syllable Structure. In the *proceeding of the European Chapter of the Association of the Computational Linguistics(EACL)*, Student Research Workshop, Avignon, France.
- [6] Sarmad Hussain , Nayyara Karamat. *Internationalized domain names: Feedback of PAN L10n project on IDNabis for Languages of Developing Asia*, Center for Research in Urdu Language Processing, Lahore, Pakistan.
- [7] Shuly Wintner. (2002) : “Formal language theory for natural language processing”. In the *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, pp. 71–76.
- [8] Verisign. The IDN registration Process. Available at: http://www.verisigninc.com/en_GB/products-and-services/domain-name-services/value-added-products/idn-domain-names/why-are-idns-important/index.xhtml?loc=en_GB
- [9] Antiphishing Working Group. Phishing Attack Trend Reports. Available at : <http://www.antiphishing.org/resources/apwg-reports/> (accessed May 2013)
- [10] Krammer, Viktor. (2006) : Phishing defense against IDN address spoofing attacks. In the *Proceedings of the 2006 International Conference on Privacy, Security and Trust*:

Bridge the Gap Between PST Technologies and Business Services. ACM.

- [11] Al Helou, Johnny, and Scott Tilley. (2010) : Multilingual web sites: Internationalized Domain Name homograph attacks, Web Systems Evolution (WSE), 2010 12th IEEE International Symposium on. IEEE.
- [12] Mark Davis and Michel Suignard. Unicode Security Mechanisms. Available at:
<http://www.unicode.org/reports/tr39/tr39-1.html>
(accessed June 20)
- [13] Wijayawardhana, Harsha, et al. (2008) : Implementation of Internet Domain Names in Sinhala, *International Symposium on Country Domain Governance*. Nagaoka, Japan, pp.20-23.
- [14] Hussain, Sarmad, and Nadir Durrani. (2006) : "Urdu Domain Names." In Multitopic Conference, 2006. INMIC'06. IEEE, pp. 299-304. IEEE.

Tin Htay Hlaing is a PhD candidate in Nagaoka University of Technology, JAPAN belonging to the department of Information Science and Control Engineering. She completed her Master degree in Computer Science (M.C.Sc) at University of Computer Studies, Yangon in 2003 and Master of Engineering (M.E) at Nagaoka University of Technology, JAPAN in 2011. Her research interests are Computational Linguistics and Formal Language Theory.

Yoshiki MIKAMI is the vice president and professor of Nagaoka University of Technology, JAPAN. He lead many language technology related research projects including Language Observatory Project, the Asian Language Resource Network Project, and Country Domain Governance Project. He also serves as a chairman of Joint Advisory Committee for ISO registry of character codes. He received a B.E in mathematical engineering from the University of Tokyo and a PhD from the Graduate School of Media and Governance at Keio University.