MY-AKKHARA: A Romanization-based Burmese (Myanmar) Input Method

Chenchen Ding, Masao Utiyama, and Eiichiro Sumita
Advanced Translation Technology Laboratory,
Advanced Speech Translation Research and Development Promotion Center,
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
{chenchen.ding, mutiyama, eiichiro.sumita}@nict.go.jp

Abstract

MY-AKKHARA is a method used to input Burmese texts encoded in the *Unicode* standard, based on commonly accepted Latin transcription. By using this method, arbitrary Burmese strings can be accurately inputted with 26 lowercase Latin letters. Meanwhile, the 26 uppercase Latin letters are designed as shortcuts of lowercase letter sequences. The frequency of Burmese characters is considered in MY-AKKHARA to realize an efficient keystroke distribution on a QWERTY keyboard. Given that the *Unicode* standard has not been extensively used in digitization of Burmese, we hope that MY-AKKHARA can contribute to the widespread use of *Unicode* in Myanmar and can provide a platform for smart input methods for Burmese in the future. An implementation of MY-AKKHARA running in Windows is released at http://www2.nict.go.jp/astrec-att/member/ding/my-akkhara.html

1 Introduction

Burmese (Myanmar) script is an abugida system, wherein basic characters can be modified using diacritics at all directions or can be combined vertically, rather than a simple left-to-right horizontal writing (Ding et al., 2016). Details of the Burmese language can be referred to in Okell and Allott (2001), Okell (2010a,b), and Okano (2007).

Although its use is encouraged in the government and universities, the use of *Unicode* for Burmese script[1] is not currently widespread. Traditional shape-based typefaces such as *Zawgyi*[2] are preferred for daily use. The issue can be regarded as *path dependence* due to traditional typewriters, wherein the input is exactly based on character shape, rather than the phonetic values of characters (Fig. 1). *Zawgyi* separately encodes all possible variants of characters and diacritics, and allows users to select correct variants manually. Hence, a redundant character set becomes incompatible with the *Unicode* standard, and extra effort is required for users to utilize typeface in detail. To provide a better interface and promote *Unicode* for Burmese digitization, we design a Burmese input method referred to as MY-AKKHARA by Romanization based on the *Unicode* standard. MY-AKKHARA is generally based on the mnemonics used in *Unicode* and the *Myanmar Language Committee Transcription System* (Department of the Myanmar Language Commission, 2014). The efficiency of the key distribution on the QWERTY keyboard layout is also considered in the design of MY-AKKHARA.

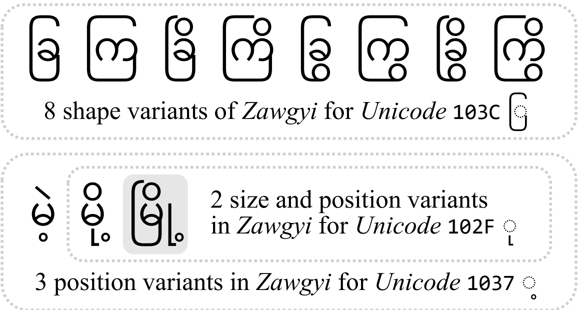
8 shape variants of *Zawgyi* for *Unicode* 103C ြ
2 size and position variants in *Zawgyi* for *Unicode* 102F ု
3 position variants in *Zawgyi* for *Unicode* 1037 ့

Figure 1: Shape, size, and position variants in *Zawgyi* for identical *Unicode* characters. The characters may affect each other: for those with gray background, the shape of 103C is determined by the inside combination, size and position of 102F by 103C, and the position of 1037 by 102F.

The implementation of MY-AKKHARA running in Windows has been released. In this paper, we first review the default layouts of Burmese provided in Windows (Win) and Macintosh (Mac) and subsequently provide detailed descriptions of MY-AKKHARA. In addition, the keystroke distribution of different methods is compared.

[1] https://www.unicode.org/charts/PDF/U1000.pdf

[2] https://code.google.com/archive/p/zawgyi/downloads

Proceedings of the 2019 EMNLP and the 9th IJCNLP (System Demonstrations), pages 157–162
Hong Kong, China, November 3 – 7, 2019. ©2019 Association for Computational Linguistics

2 Win and Mac Burmese Keyboards

The keyboard layouts used to input Burmese in *Unicode* have been provided in Win and Mac operating systems. Figure 2 illustrates the default layouts of the Burmese *Unicode* keyboard in these two mainstream operating systems. Both of the layouts are a simple mapping from characters to keys. Excluding special punctuation marks and native number digits, 63 *Unicode* characters are required to represent modern standard Burmese textual data, from *Unicode* 1000 to 104F.[3] Therefore, 26 keys with shift are not sufficient to cover the character set. In both of the layouts, extra punctuation (or digit), or alternative keys are necessary in typing.

The Win layout is adjusted from the traditional layout of a typewriter, by removing redundant character varieties and re-arranging the characters inputted using the Shift-key. Considering that this layout has a large portion of the traditional one but with a certain difference, many Win users are not interested in switching to this layout. Hence non-*Unicode* fonts are still inputted using the traditional keyboard in practice. Moreover, the Mac layout is completely redesigned based on Romanization manner, wherein the Burmese characters are arranged on the basis of their pronunciations as represented by the letters on a QWERTY keyboard. However, the design is inflexible, without considering the practical use of Burmese characters. Thus, the positioning of fingers when typing is tricky. The comparison of the keystroke distribution will be presented in Section 4.

3 MY-AKKHARA: Proposed Input Method

The proposed MY-AKKHARA is inspired by the Mac layout and deemed highly natural and efficient. Rather than a simple mapping between the *Unicode* characters and keys, we also facilitate character alternation processing by using the inputted Latin letters. Specifically, double keystrokes of e, f, h, i, j, r, u, v, w, and y, and the h- and g-keys at the middle of a QWERTY keyboard are used to alternate characters. This design naturally integrates the Romanization into the character alternation processing. The q-key is reserved to disambiguate in obscure cases through which the input method can precisely input any strings with the *Unicode* Burmese characters.[4] Lowercase a, o, x, and 26 uppercase Latin letters are assigned as optional shortcuts. Figure 3 shows an example on the technique of inputting a Burmese string with rare and stacked characters using the proposed method.

The instruction of the proposed input method can be printed by users on an A4 paper (Fig. 4). The proposed method can be formulated primarily through a finite-state automaton (Hopcroft et al., 2013), receiving strings comprising 23 lowercase Latin letters (excluding a, o, and x) and transiting among different states that represent Burmese characters. The **Appendix** provides the description of the automaton.

The shortcuts can be grouped in the following four categories:

- three lowercase letters for common combinations: a=qevq, o=qiuq, and x=qngfq;
- uppercase letters to save double keystrokes: E=qee, F=qff, H=qhh, I=qii, J=qjj, R=qrr, U=quu, V=qvv, W=qww, and Y=qyy;
- uppercase letters to save h/g: B=qbh, C=qch, D=qdh, G=qgh, K=qkh, L=qlg, M=qmg, P=qph, Q=qg, T=qth, and Z=qzh; and
- uppercase letters for other cases: A=qegg, N=qny, O=qsr, S=quug, and X=qng.

Lowercase letters a, o, and x can considerably save keystrokes. Note that the shortcuts have a preceding q in the implementation through which disambiguation can be realized. The recommended uppercase letters are Y, H, and Q, which can resolve almost all ambiguous cases when typing orthographically correct Burmese texts.

Two issues related to normalizing the encoding of the Burmese script in *Unicode* are addressed:

- 102B is a variant of 102C, exclusively used for narrow characters of 1001, 1002, 1004, 1012, 1015, and 101D. This alternation is executed automatically when typing v or a (i.e., shortcut for ev). However, qv and qvg can exactly input 102C and 102B, respectively.
- 1037 and 103A can appear successively; however, their order is not precisely identified. 103A 1037 will always be normalized in *Unicode* into the recommended order 1037 103A.

[3] Within this range, from 1040 to 104B are Burmese digits and punctuation marks; 1022, 1028, 1033, 1034, and 1035 are not used for standard Burmese.

[4] It is possible to intentionally input orthographically incorrect Burmese strings; however, orthographically correct strings can be inputted more naturally than incorrect ones.

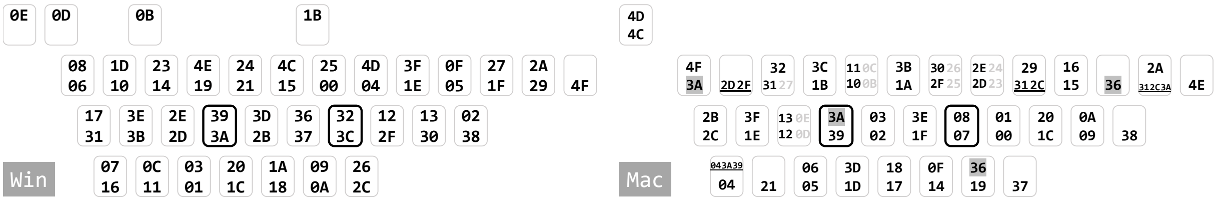
0E 0D 0B 1B
08 06 1D 10 23 14 4E 19 24 21 4C 15 25 00 4D 04 3F 1E 0F 05 27 1F 2A 29 4F
17 31 3E 3B 2E 2D 39 3A 3D 2B 36 37 32 3C 12 2F 13 30 02 38
Win
07 16 0C 11 03 01 20 1C 1A 18 09 0A 26 2C
4D 4C
4F 3A 2D2F 32 31 27 3C 1B 11 0C 10 0B 3B 1A 30 26 2F 25 2E 24 2D 23 29 312C 16 15 36 2A 312C3A 4E
2B 2C 3F 1E 13 0E 12 0D 3A 39 03 02 3E 1F 08 07 01 00 20 1C 0A 09 38
Mac
043A39 04 21 06 05 3D 1D 18 17 0F 14 36 19 37

Figure 2: Default Burmese layout in `Win` (left) and `Mac` (right). Only the final two digits of *Unicode* are shown for a compact presentation. The places of `f`- and `j`- keys on a `QWERTY` keyboard are marked by bold frame. For each key, the lower character is inputted using simple keystroke, whereas the upper character requires pressing the `Shift`-key. On the `Mac` keyboard, some character combinations are mapped on one key, which is underlined in the figure. Meanwhile several rare characters require `Alt`-key, which is in gray color. Note that `103A` and `1036` marked with gray background appear two times on the `Mac` keyboard.

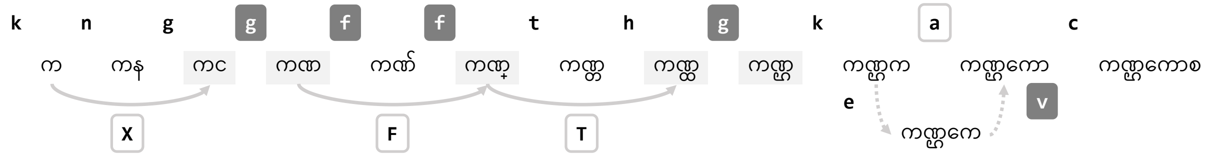

Figure 3: Example of the proposed input method. The top row is the typed Latin letters; the inputted Burmese string after each keystroke is presented in an increasing manner. Latin letters with frame are the shortcuts and those with dark background are special design that should be remembered by users. Burmese strings with gray background have a character alternation from their previous status. Although `g` and `h` are regarded as alternation operators, they are also part of the Romanization, i.e., the first `g` after `n` and `h` after `t`. The shortcuts mainly save the extra alternation by `g`, `h` and double keystroke (i.e., `X`, `T`, and `F`). Lowercase `a` is a shortcut for an extremely common character combination that can be inputted using `ev`.

4 Keystroke Distribution

The Burmese language has two different styles: literary and colloquial. For the literary style data, the publicly accessible Burmese dataset in the *Asian Language Treebank* (ALT) project (Riza et al., 2016) is used, containing approximately 20, 000 long sentences from news articles.[5] For the colloquial style data, we use an in-house translated Burmese version of the *Basic Travel Expression Corpus* (BTEC) (Kikui et al., 2003), comprising approximately 400, 000 daily expressions. Figures 5 and 6 show the comparison of the keystroke distribution in `Win` and `Mac` keyboards and by `MY-AKKHARA`, respectively.

The middle area of the `Mac` keyboard has not been efficiently used. Although the uppercase `F` can be used instead of lowercase `q`, the frequency of the `Shift`-key will increase considerably. The keystroke is more focused at the middle of the keyboard by `MY-AKKHARA` than that on the `Win` and `Mac` keyboards. The use of the `Shift`-key is optional in `MY-AKKHARA`, depending on the users' preference. When the `Shift`-key is completely applied, the frequency is less than two times that of used in `Win` keyboard, and it is approximately equal to the lower bound used in `Mac` keyboard. Generally, index fingers are mostly utilized and little fingers have fewer burdens in `MY-AKKHARA`.

5 Conclusion and Future Work

In this study, a Romanization-based Burmese input method called `MY-AKKHARA` is proposed to promote the *Unicode* standard for Burmese digitization. `MY-AKKHARA` can also be regarded as a Burmese-specified, lossless coding version of Ding et al. (2018), providing a platform to develop a further fuzzy and smart Burmese input method.

References

Department of the Myanmar Language Commission. 2014. *Myanmar-English dictionary (Myanma-anggalip abidan)*, 12 edition. Ministry of Education, the Republic of the Union of Myanmar.

Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. 2018. Simplified Abugidas. In *Proc. of ACL, Vol. 2*, pages 491–495.

[5] `http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/my-nova-170405.zip`

Consonants					
1000 က k	1001 ခ kh|K	1002 ဂ g|Q	1003 ဃ gh|G	1004 င ng|X	
1005 စ c	1006 ဆ ch|C	1007 ဇ z	1008 ဈ zh|Z	100A ည ny|N	1009 ဉ nyg|Ng
100B ဋ tg	100C ဌ thg|Tg	100D ဍ dg	100E ဎ dhg|Dg	100F ဏ ngg|Xg	
1010 တ t	1011 ထ th|T	1012 ဒ d	1013 ဓ dh|D	1014 န n	
1015 ပ p	1016 ဖ ph|P	1017 ဗ b	1018 ဘ bh|B	1019 မ m	
101A ယ yy|Y	101B ရ rr|R	101C လ l	101D ဝ ww|W	101E သ s	103F ဿ sg
	101F ဟ hh|H	1020 ဠ lg|L	1021 အ vv|V		

Independent Vowels							
1021 အ vv|V	1023 ဣ ig	1024 ဤ iig|Ig	1025 ဥ ug	1026 ဦ uug|Ug|S	1027 ဧ eg	1029 ဩ sr|O	102A ဪ srg|Og

Dependent Vowel Signs							
102B ◌ါ vg	102C ◌ာ v	102D ◌ိ i	102E ◌ီ ii|I	102F ◌ု u	1030 ◌ူ uu|U	1031 ေ◌ e	1032 ◌ဲ ee|E

Various Signs					Combined Vowel Signs		
1036 ◌ံ mg|M	1037 ◌့ jj|J	1038 ◌း j	1039 ◌္ ff|F	103A ◌် f	102D 102F ◌ို o	1031 102C ေ◌ာ a	1004 103A င် x

Dependent Consonant Signs				Various Signs			
103B ◌ျ y	103C ◌ြ r	103D ◌ွ w	103E ◌ှ h	104C ၌ nggg|Xgg	104D ၍ rrg|Rg	104E ၎ lgg|Lg	104F ၏ egg|A

Figure 4: Proposed input method. In each cell, the *Unicode*, the Burmese character, and the input manner are illustrated from top to bottom. For Burmese characters having more than one way to input, vertical bar is used to separate different manners.

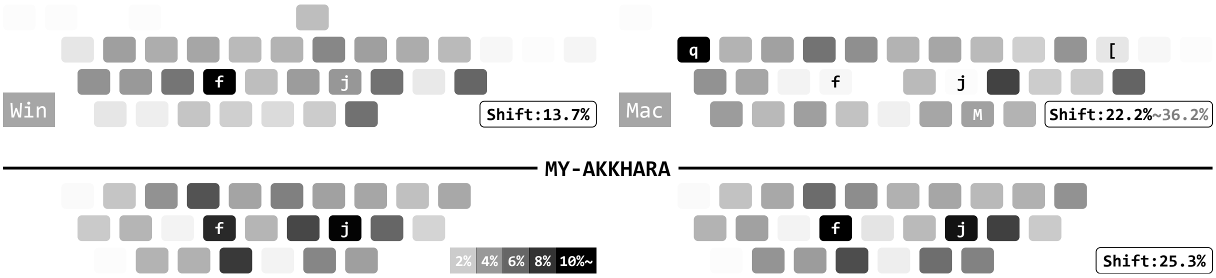

Figure 5: Keystroke distribution on the ALT literary data. The upper-left and upper-right diagrams are `Win` and `Mac` keyboards, respectively. The lower images are `MY-AKKHARA`, with `Shift` not used (left) and `Shift` completely used (right) manners, respectively. The usage frequency of the `Shift`-key is also presented. Note that `103A` and `1036` appear twice on the `Mac` keyboard. The two characters are counted by using `q` and `[` to input in the diagram, where the frequency of `Shift`-key is 22.2%. The two character can be also inputted by uppercase `F` and `M`. If they are always inputted using the `Shift`-key, then the frequency of `Shift`-key increases to 36.2%.

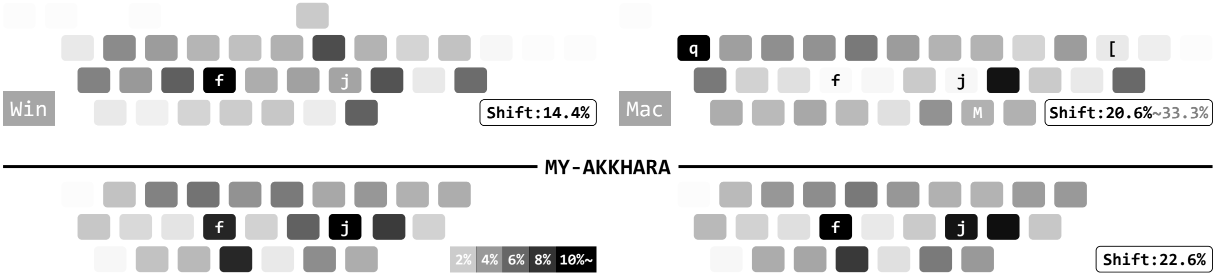

Figure 6: Keystroke distribution on the BTEC colloquial data. The configuration is the same as that of Fig. 5.

Chenchen Ding, Ye Kyaw Thu, Masao Utiyama, and Eiichiro Sumita. 2016. Word segmentation for Burmese (Myanmar). *ACM TALLIP*, 15(4):22.

John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. 2013. *Introduction to Automata Theory, Languages, and Computation*, 3 edition. Pearson.

Genichiro Kikui, Eiichiro Sumita, Toshiyuki Takezawa, and Seiichi Yamamoto. 2003. Creating corpora for speech-to-speech translation. In *Proc. of EUROSPEECH*, pages 381–384.

Kenji Okano. 2007. *Colloquial Burmese (Myanmar) Grammar*. Kokusai Gogakusha. (in Japanese).

John Okell. 2010a. *Burmese – An introduction to the Spoken Language, Book 1*. Northern Illinois University Press.

John Okell. 2010b. *Burmese – An introduction to the Spoken Language, Book 2*. Northern Illinois University Press.

John Okell and Anna Allott. 2001. *Burmese / Myanmar Dictionary of Grammatical Forms*. Routledge.

Hammam Riza, Michael Purwoadi, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Rapid Sun, Sethserey Sam, Sopheap Seng, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama, and Chenchen Ding. 2016. Introduction of the Asian language treebank. In *Proc. of O-COCOSDA*, pages 1–6.

Appendix

Figure 7 shows the overall configuration. Routes connecting the the initial (q_s) and final (q_s) states are listed in Figs. 8 – 16, where $q_n, (n \in \mathbf{N})$ are Burmese characters.[6] Although all q_n can be the final states, a separate q_e is used for clarity, and a `q` is marked explicitly on all the arcs to q_e.

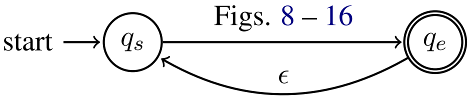

Figure 7: Overall configuration of the automaton.

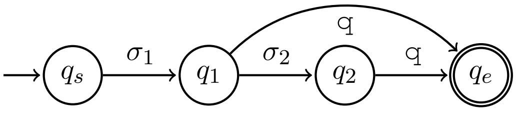

Figure 8: Simplest case. When σ_2 is `h`, (σ_1, q_1, q_2) can be (`k`, `00`, `01`), (`g`, `02`, `03`), (`c`, `05`, `06`), (`z`, `07`, `08`), (`p`, `15`, `16`), and (`b`, `17`, `18`). When σ_2 is `g`, (σ_1, q_1, q_2) is (`m`, `19`, `36`). When $\sigma_2 = \sigma_1$, (σ_1, q_1, q_2) can be (`y`, `3B`, `1A`), (`w`, `3D`, `1C`), and (`h`, `3E`, `1D`). All σ_1 are natural Romanization. When σ_2 is `h`, it is also a part of the Romanization.

[6] *Unicode* is referred to by the final two digits for brevity.

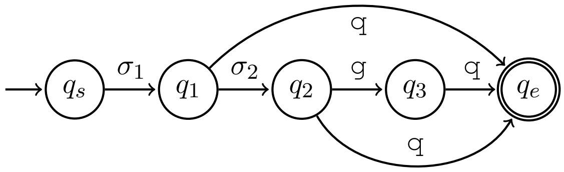

Figure 9: Two-step alternation. When (σ_1, σ_2) is (l, g), (q_1, q_2, q_3) is (1C, 20, 4E). Here, l is a natural Romanization for 101C and 1020, whereas 104E is a special abbreviated mark with l as onset. When (σ_1, σ_2) is (r, r), (q_1, q_2, q_3) is (3C, 1B, 4D), respectively. Here, r is a natural Romanization for 103C and 101B, whereas 104D is a special abbreviated mark with r as onset.

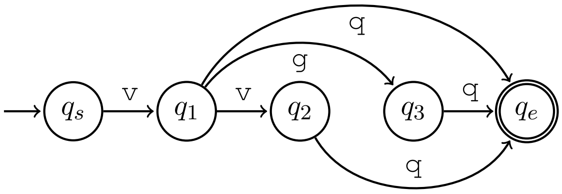

Figure 10: Alternation variant of Fig. 9. (q_1, q_2, q_3) is (2C, 21, 2B). Considering that 102C and 1021 are frequently used, the convenient v-key is assigned instead the natural Romanization by a.

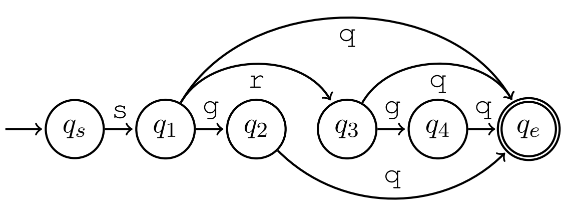

Figure 11: Alternation in Fig. 8 with an extra branch. (q_1, q_2, q_3, q_4) is (1E, 3F, 29, 2A). Here, s is a natural Romanization for 101E, whereas 103F, 1029, and 102A are extremely obscure.

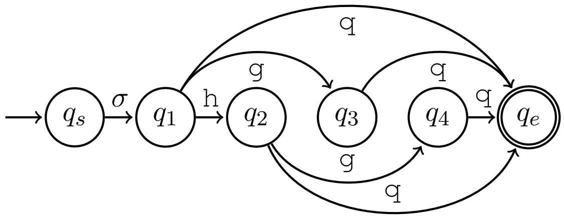

Figure 12: Alternation by h and g. $(\sigma, q_1, q_2, q_3, q_4)$ can be (t, 10, 11, 0B, 0C), and (d, 12, 13, 0D, 0E). Both t and d are the natural Romanization, and h is also a part of the Romanization.

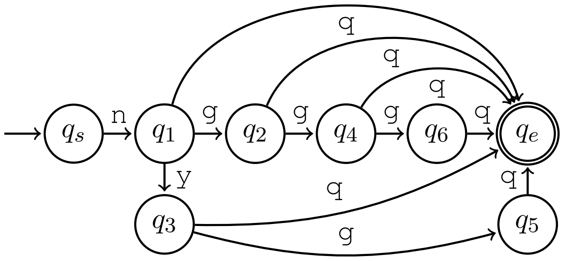

Figure 13: Most complex alternation. $(q_1, q_2, q_3, q_4, q_5, q_6)$ is (14, 04, 0A, 0F, 09, 4C). Here, n, ng and ny are the natural Romanization for 1014, 1004, and 100A, respectively. Other alternated characters are rare.

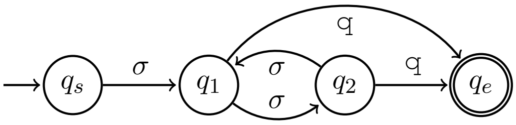

Figure 14: Doubled and looped alternation. (σ, q_1, q_2) can be (j, 38, 37), and (f, 3A, 39). Here, 1038 and 103A are remarkably frequent marks; hence convenient j- and f-keys are assigned, respectively.

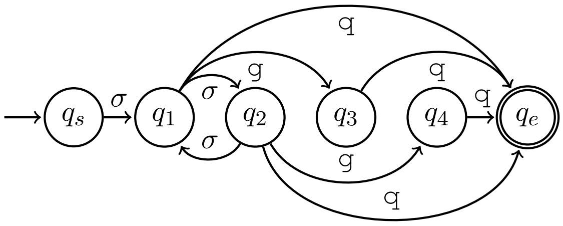

Figure 15: Combination of Figs. 12 and 14. $(\sigma, q_1, q_2, q_3, q_4)$ can be (i, 2D, 2E, 23, 24), and (u, 2F, 30, 25, 26). Here, i and u are the natural Romanization for the corresponding characters.

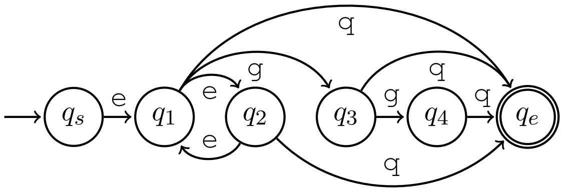

Figure 16: Alternation in Fig. 14 with an extra branch. (q_1, q_2, q_3, q_4) is (31, 32, 27, 4F). Here, e is a natural Romanization for 1031, 1032 and 1027, whereas 104F is an abbreviated mark derived from 1027.