Latin to Sundanese script conversion using Finite State automata algorithm

C Slamet\(^1\)*, Y A Gerhana\(^2\), D S Maylawati\(^2\), M A Ramdhani\(^1\) and N Z Silmi\(^1\)

\(^1\) Departement of Informatics, UIN Sunan Gunung Djati Bandung, Jalan A H Nasution 105 Bandung 40614, Indonesia
\(^2\) Departement of Informatics, Sekolah Tinggi Teknologi Garut, Jalan Mayor Syamsu No 1 Tarogong Kidul Kabupaten Garut 44151, Indonesia

*cepy_lucky@uinsgd.ac.id

Abstract. Indonesia is rich in local language. Several of them have their own script, such as Sundanese. Sundanese script is a culture legacy that deserves to be preserved. Sundanese script is syllabic that means each script illustrates the symbol of a syllable, so to change it needs spelling separation. This research proposed a system as solution to learn Sundanese script. The system accepted Latin input which would then be converted into Sundanese script. Latin script was splitted into syllable using the Finite State Automata algorithm with the Non-Deterministic Finite Automata (NDFA) model. Then, transition of syllable separation was represented by transition diagram and guided by the rules of writing and the pattern of Sundanese syllables. We did 6 test cases as a simple black box testing for making sure that the conversion process has done well. Then, we tested the system with 30 respondents using a questionnaire with 5 questions relating to the ease and accuracy of the information provided in the process of converting Latin script to the Sundanese script. With Likert technique for data processing of the questionnaire, obtained the result was around 79.8% of the respondent strongly agree that the information provided was precise, the system was easy to use and easy to understand the Sundanese script.

1. Introduction

Local language in every country has a unique grammar and syntax. Moreover, the local language appeared before official language. Several of them has an own scripts or symbols to reveal the intent in it before getting to know how to write. Sundanese is one of 583.

Indonesian local language that has unique scripts [1] [2]. With the script, the next generation can find important information about the history, the noble values of culture, and other cultural aspects that inherited by the predecessors. Therefore, script of local language should be maintained and empowered by the area concerned, one of the legacy that must be maintained is Sundanese script.

The Government of Indonesia has regulated in West Java Provincial Governor Regulation No. 5 of 2003 which expressly and specifically strives on the use, maintenance, and development of Sundanese language, literature, and script [3]. In support of government efforts in the era of information technology today, it is necessary to digitized Sundanese script to introduce and study it easily in the current conditions where many people, especially the people of West Java began to ignore it. In addition, many people who only know but do not know how the rules of writing Sundanese script due to lack of knowledge about it [4]–[6]. Sundanese script is syllabic, it means each script describes the symbol of a syllable. So, to convert it need the separation of word. In computerized science there is one branch of...
science that specifically learn about language, that is Theory of Automata [7]. The language discussed is written language which consists of symbols that make up a set that is called the alphabet. Automata theory deals with abstract machine theory, which is a sequential machine that receives input and outputs in discrete form [8], [9]. The model used in language and automata theory is a state machine model or state transition model that is inputted by an automated machine. Then, it will make a decision indicating whether or not the input is received.

Finite State Automata (FSA) is one of automata theory that can model syllable recognition [8]. FSA method had been used in several researches about Indonesian language and Indonesian local language, one of them was research done by Rieke Adrianti W. In this research, FSA was used to recognize and process language based on word phoneme which then becomes the basis as syllabic pattern formation in Bahasa Indonesia. The study compared the Aryanata method with the FSA method to decode in Bahasa Indonesia. The result showed that FSA method is capable of beating the word in Indonesian better than Aryanta method [10]. A similar study was conducted by Ana Mutmainnah using the FSA for converting Indonesian to Jawi script [11]. The results of this research obtained by the method of syllable recognition using FSA had a percentage of accuracy around 92.5%. Moreover, in the research conducted by Pyeit Rinekso Androyanto, FSA algorithm could applied for learning reading beginning Madura language and recognized the classification of Madurese syllables with 100% percentage of accuracy. Another research used FSA for Latin to Java language conversion [7] and recognized Indonesian syllable [12] were also succeed with a good result.

Based on previous research that FSA was good to recognise syllable pattern for language, so that we propose FSA for conversing Latin to Sundanese script. In the upcoming section, we describe about Sundanese Script in Section II, Finite State Automata (FSA) in Section III, and Latin to Sundanese text conversion process using FSA in Section IV. Then, for result of experiment and discussion are explained in Section V. Finally, conclusion of this research stated in Section VI.

2. Sundanese scripts

Scripts are used as a means of conveying speech in the form of symbols. Indonesia has 12 scripts of local language, such as: Balinese script, Batak, Bengkulu, Bima, Bugis, Java, Komering, Lampung, Makasar, Pasemah, Rejang, and also Sundanese [2]. Sundanese has 32 pieces of script consisting of 7 script as swara or independent vocal (describe in figure 1) and 23 script as ngalagena or consonant (describe in figure 2). Swara script symbolizes the sound of independent vowel phonemes as a syllable that can occupy the beginning, middle, and end of a word. While, ngalagena script symbolizes the sound of consonant phonemes as a word or syllable that can occupy the position of beginning, middle or end of a word. Thus, this Sundanese script is syllabic, it means writing that can represent a word and syllable.

![Figure 1. Swara or vowel of Sundanese script.](image)

Ngalagena scripts in ancient Sundanese script have an 18 symbols. However, currently ngalagena scripts have a 23 symbols because of absorption of foreign syllable. Those scripts are not a new symbol, but there are variant of the ancient Sunda script symbols that rarely used, among others the symbol of “ťa” and “ťa” are a variant of the pa script, the symbol of the “qa” and “xa” are a variant of “ka”, and the symbol of “za” is a variant of the symbol of “ja”. Then, there are two additional symbol among others “kha” and “sya”. In the Sundanese script, there are known vocalizations, namely rarangkén or sound markers that can change, add or eliminate vowel in the ngalagena script (explained in table 1). The symbol of vocalization has a 13 kinds which in its placement is divided into three groups. The first group, as many as 5 symbols placed above the basic script. The second group, as many as 3 symbols placed under basic script. The third group, as many as 5 symbols are placed parallel to the basic script, which is subdivided into: 1 symbol placed to the left of the basic script, 2 symbols placed on the right of the basic script, and as many as 2 symbols placed on the right with a little stretching down the basic script. In addition, also known symbols of numbers in the form of basic numbers (0-9), describe in figure 1.
3. The physical form of the Sundanese script embodied in its vocalisation is written on the slope position between 45°-75°. The comparison of physical size of the basic script, both vowel and consonant are generally written 4:4, except for /ra/ as the ngalagena script because it is 3:4; for /ba/ and /nya/ are 4:6; and for /i/ as the swara script is 4:3. As for the comparison of physical size the marks of vowelization are generally written 2:2, except for /+ng/ as panyecek is 1:1; /+r/ as panglayar is 2:3; /+ra/ as panyakra is 2:4; and /+ya/ as pamingkal is 2:4 for bottom and 3:2 for the right side. The comparison of the physical size of the base number is generally written 4:4, except for the numbers /4/ and /5/ are 4:3.

![Figure 2. Ngalagena or consonant of Sundanese script.](image)

![Figure 3. Number script of Sundanese script.](image)

Punctuations on Sundanese script are the same as Latin, including how to write a title, either academics or religious title. Furthermore, Sundanese script has been registered and standardized by Unicode [13]. Unicode is a character encoding system that designed to allow text and symbols of all writing systems in the world to be displayed and manipulated consistently by the computer [14]–[16].

| Name     | Symbol | Function                                | Example |
|----------|--------|-----------------------------------------|---------|
| Paghulu  | ⚫      | Changing the sound of vowal /a/ into /i/ | ![Example](image) |
| Pamepet  | ▲      | Changing the sound of vowal /a/ into /e/ | ![Example](image) |
| Paneuleung | ▲      | Changing the sound of vowal /a/ into /eu/ | ![Example](image) |
Table 1. Cont.

| Palaeography | Explanation | Example |
|--------------|-------------|---------|
| ioxid | Adding /+r/ in the end of basic script | [Symbol] |
| Panyarak | Adding the sound /+r/ to the middle of base script, and can be adjusted with the basic vowelization flags. | [Symbol] |
| oxid | Adding the sound /+r/ to the identical base script, and can be adjusted with the basic vowelization flags. | [Symbol] |
| Paneleng | Changing the sound of vowel /a/ into /é/ | [Symbol] |
| Panom | Changing the sound of vowel /a/ into /o/ | [Symbol] |
| Panyarak | Adding the sound /+ya/ to the identical base script, and can be adjusted with the basic vowelization flags. | [Symbol] |
| Panyarak | Adding the sound /+ya/ to the identical base script, and can be adjusted with the basic vowelization flags. | [Symbol] |
| Panyarak | Adding /+h/ in the end of basic script | [Symbol] |
| Paten or Pamaeh | Removing the sound of vowel. | “Ka” to be “k” |
3. Finite State Automata (FSA)

FSA is one of automata theory that aims to recognize and process the language based on word phonemes which became the basis for the formation of syllabic patterns [17], [18]. Automata machine can be used to recognize a language based on its grammar [19]. Automata theory and formal language will generate and recognize the sentences (strings). Then, those strings will be modelled by machine that has components among others input tape, head tape, Finite State Control (FSC), and memory. The recognition machine is deterministic when in every configuration, there is only one possibility that the machine can perform, otherwise the identification engine is non-deterministic FSA method applied for separate the words into syllables performed that determining the letters of input based on word phonemes. In this research, the FSA method used the Non-Deterministic Finite Automata (N DFA) model is expressed by \((Q, \Sigma, \delta, S_0, F)\) and transition diagrams [12], [20]. Where \(Q\) is set of state, \(S_0\) is initiation state or start state, and \(F\) is set of the last state. For example, figure 3 describes simple DFA machine with \(Q = \{q_0, q_1, q_2\}\), \(\Sigma = \{a, b\}\), \(S_0 = q_0\), and \(F = \{q_2\}\). Then figure 3 has transition function, among others \(d(q_0, a) = q_0\), \(d(q_0, b) = q_1\), \(d(q_1, a) = q_1\), \(d(q_1, b) = q_2\), \(d(q_2, a) = q_1\), and \(d(q_2, b) = q_2\) with DFA transition table is in table 2.

![Figure 4. Simple DFA Machine.](image)

### Table 2. Table of DFA transition of figure 4.

| \(\delta\) | \(A\) | \(B\) |
|---|---|---|
| \(q_0\) | \(q_0\) | \(q_1\) |
| \(q_1\) | \(q_1\) | \(q_2\) |
| \(q_2\) | \(q_1\) | \(q_2\) |

The process of DFA begin from DFA will be in start status (S), the head of the ribbon on the first symbol. Next, the head of ribbon will read the symbols of the ribbon and shift forward. For each symbol, DFA will switch the status according to the transition function \((\delta)\). Then, the process will end when the input symbol on the ribbon is exhausted, at the end of the process when the final state is reached then the input string is received (recognized as a string of the regular language), and then if not the input string is rejected (not recognized).

4. Latin to Sundanese text conversion using FSA

There are three process for doing Latin to Sundanese conversion using FSA. Begin from character recognizing, separating word, and separating syllable. In character recognizing process, the text of a latin script as input will be identified by the character type. The types of characters to be recognized at this stage consist of: vowels, consonants, numbers, punctuation, vocalization of the *rarangken*, and the final Sundanese. Final Sundanese serves to increase the vocalization of *rarangken* at the end of the script, if there are the letter of h, r, or ng after the last vowel. Next, words of sentences will be separated. Then, separating syllable of word will have special pattern according to the rules of writing the Sundanese script. In this process will be made into one level of transition diagram (Describe in figure 5), with \(M = (Q, \Sigma, \delta, S, F)\), \(Q = \{q_0, q_1, q_2, q_3, q_4, q_5, q_6, q_7, q_8, q_9\}\), \(\Sigma = \{\text{blank}, V, K, R\}\), \(S = q_0\), \(F = \{q_1, q_2, q_3, q_6, q_7, q_8, q_9\}\). There are six groupings of syllable pattern of Sundanese which are described in table 3, and from transition diagram is produced transition table of Finite State Diagram which described in table 4.
Table 3. Syllable Pattern of Sundanese script.

| Pattern | Description | Example |
|---------|-------------|---------|
| V       | There are seven vowels, among others a, i, eu, u, e, and ê | \( \overline{\varepsilon} - \text{u} | \varepsilon - \text{i} | \varepsilon - \text{u} |
| VK      | Consist of vowel and consonant series, among others k, q, g, ng, c, j, z, ny, t, d, n, p, f, v, b, m, y, l, w, s, x, h, kh, and sy. | \( \overline{\text{q}} = \text{ak} \ | \overline{\varepsilon} = \text{om} \ |
|         | Able to be combined with *rarangken* of h, r, and ng. | \( \overline{\text{R}} = \text{v} | \overline{\varepsilon} = \text{eng} \ |
| KV      | This pattern produces *ngalagena* with vowel *rarangken* | \( \overline{\text{v}} = \text{al} \ | \overline{\varepsilon} = \text{u} | \overline{\varepsilon} = \text{u} \ |
| KVK     | Consist with consonant-vowel-consonant series | \( \overline{\text{u}} = \text{ek} \ | \overline{\text{e}} = \text{eis} \ | \overline{\text{u}} = \text{ecn} \ |
|         | Able to be combined with *rarangken* of h, r, and ng | \( \overline{\text{v}} = \text{neh} \ | \overline{\varepsilon} = \text{ne} | \overline{\varepsilon} = \text{ne} \ |
| KRV     | On the principle of writing Sunda script, if a consonant meets a consonant letter that includes la, ya, or ra, it will form a Sundanese script with certain *rarangken* vocalizations | \( \text{\text{k} = ky} | \overline{\text{\text{k}}} = \text{kei} | \overline{\text{\text{k}}} = \text{kli} \ |
|         | Consist with two or three *rarangken* vocalization in one character. | \( \overline{\text{\text{k}}} = \text{gong} | \overline{\text{\text{k}}} = \text{gong} |
| KRVK    | Able to be combined with *rarangken* of h, r, and ng. | \( \overline{\text{\text{k}}} = \text{kly} | \overline{\text{\text{k}}} = \text{kly} | \overline{\text{\text{k}}} = \text{kly} \ |

Note:
V = Vowel
K = Consonant
R = Rarangken

5. Result and discussion
We developed web-based application for this research, with FSA inside it (the example of application was shown in figure 6 and 7). Our application was used in addition to preserving and facilitating the learning of Sundanese script, and also to prove and test FSA in recognizing the Sundanese script. We did several scenarios to test our research. First, we did black box testing for making sure that all of functionality of application had been run well without logical wrong. Second, we tested correctness of
Latin to Sundanese script conversion for each syllable pattern of Sundanese script. And the last, we did initial testing by interviewing 30 respondents with 5 questions to find out whether the application we build is good enough in converting Sundanese script and easy to use.

For first scenario we found that all of functionality of application were running well. The result of second scenario which was describe in table 4 showed that FSA could recognize all of syllable pattern of Sundanese scripts (described in table 5). It means that FSA succeed to convert Latin into Sundanese well. And for the last scenario (described in table 4), with Likert formula we found that in Question 1, 80% of respondent strongly agree that application give right Latin to Sundanese conversion. The result of Question 2 showed that 82% of respondents got complete information enough in Sundanese script, and the result of Question 3 showed that 76% of respondents strongly agree if the application was built in web based. The result of Question 4 showed that 79% of respondents strongly agree that the application was user friendly, and the last result of Question 5 showed that 82% of respondents strongly agree that the application was easy to use.

Application development in the conversion of these letters can then be developed into various platform, for example to android based devices. Android had proven had a good capability in data processing and information display [21] [22]. The advanced development of this algorithm can be a source of multi-media based learning, where multimedia-based information technology approach had a high effectiveness value in learning outcomes [23] [24] [25] [26] [27].
Table 4. The result of second testing scenario.

| No | Latin text, Vowel : a i u e o eu ê | Script Conversion | Conclusion |
|----|-----------------------------------|-------------------|------------|
| 1  | Latin text, Vowel : a i u e o eu ê |                   | Correct    |
| 2  | Latin text for Vowel-Consonant, example : 'am-is ar-ang' |                   | Correct    |
| 3  | Latin text for Consonant-Vowel, example:'da-di-du' |                   | Correct    |
| 4  | Latin text for Consonant-Vowel-Consonant, example : 'mas-mis-mus mah-mar-mang' |                   | Correct    |
| 5  | Latin text for Consonant-Rarangken-Vowel, example: 'dy-a-dl-dri' |                   | Correct    |
| 6  | Latin text for Consonant-Rarangken-Vowel-Consonant, example: 'syam-krab-kyar-syeh-kyong' |                   | Correct    |

Table 5. The result of third testing scenario.

| Question                                                                 | SS  | S  | BS | TS | STS | Percentage |
|-------------------------------------------------------------------------|-----|----|----|----|-----|------------|
| Is the application give correct information about Latin to Sundanese    | 10  | 11 | 9  | 0  | 0   | 80%        |
| Is the application give complete information about                       | 12  | 10 | 8  | 0  | 0   | 82%        |
| Do you agree that the application is web based?                         | 8   | 9  | 13 | 0  | 0   | 76%        |
| Is the application user friendly?                                        | 9   | 11 | 10 | 0  | 0   | 79%        |
| Is the application easy to use?                                          | 10  | 14 | 6  | 0  | 0   | 82%        |
| Average of Strongly Agree                                               |     |    |    |    | 79.8% |            |

Note:
SS : Strongly Agree
S  : Agree
BS : Ordinary
TS : Disagree
STS : Strongly Disagree

6. Conclusion

Local language was the nation's cultural heritage that need to be preserved. Sundanese was one of Indonesian local language that had its own script. With unique rules and grammar of Sundanese script, we succeed to recognized Sundanese script and converted from Latin into Sundanese. We used FSA algorithm for syllable recognizing of Sundanese script. The result showed that FSA was very good in recognized and convert Latin script into Sundanese Script. Moreover, based on interview with several
respondents, our research had high percentage enough for correctness of conversion process, completeness of output information of Sundanese script, user friendly, and also easy to use.

For future works, it is possible to improve our research to be interactive learning media about Sundanese script, with adding exercise function. Then, this research can be implement another method or algorithm beside FSA for recognizing the Sundanese scripts that make the process more efficient and effective.

Acknowledgement
We say many thanks for West Java Provincial Culture Office for its willingness to provide the data. And also, many thanks for Research and Publication Center of UIN Sunan Gunung Djati Bandung Indonesia for supporting publication of our research.

References
[1] Ayatrohaedi 2018 Kebanggan Bangsa Indonesia [Online]. Available: http://suarahatimifta.blogspot.co.id/2012/02/20-hal-yang-menjadi-kebanggan-indonesia.html.
[2] Baidillah I, Darsa U A, Abdurahman O, Permadi T, Gunardi G, Suherman A and Sutisna D 2008 Direktori Aksara Sunda untuk Unicode (Jawa Barat: Pemerintah Provinsi Jawa Barat)
[3] No. Peraturan Daerah Propinsi Jawa Barat Tahun 2003 Tentang Pemeliharaan Bahasa Sastra dan Aksara Daerah (5).
[4] G E Pribadi, U Syaripudin and W Uriawan, 2016 Aplikasi pembelajaran bahasa sunda dengan implementasi algoritma linear congruential generator dan fuzzy berbasis android J. Online Inform. 1(1) pp. 34–42
[5] H Purnama 2013 Pembangunan aplikasi pembelajaran aksara sunda berbasis android (Universitas Komputer Indonesia)
[6] M A F Andhasty 2017 Aplikasi pembelajaran aksara sunda berbasis android (Universitas Sebelas Maret Surakarta)
[7] C S Aji, E A Sarwoko and R Saputra 2012 Aplikasi Konversi Aksara Latin Ke Aksara Jawa Menggunakan Finite State Automata Dengan Visual Basic J. Informatics Technol 1(3) pp. 52–66
[8] G Abdurrahman 2011 Penerapan dan Implementasi Teori Bahasa dan Otomata 1st ed. (Yogyakarta: Universitas Yogyakarta)
[9] Hopcroft J E, Motwani R and Ullman J D 2001 Introduction to automata theory, languages, and computation Acm Sigact News 32(1) 60-65
[10] Tolle H and Setyawati O 2016 Pengembangan Aplikasi Text-To-Speech Bahasa Indonesia Menggunakan Metode Finite State Automata Berbasis Android Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI) 5(1)
[11] A Mutmainnah 2013 Pengenalan Suku Kata Bahasa Indonesia Menggunakan Finite State Automata untuk Alih Aksara Indonesia -aksara Jawi Universitas Tanjungpura
[12] T A Basuki 2000 Pengenalan Suku Kata Bahasa Indonesia Menggunakan Finite-State Automata Integral 5(2) pp. 7–74
[13] Unicode Unicode Aksara Sunda The Unicode Standard 9.0, Copyright © [Online] Available: http://www.kairaga.com/unicode-aksara-sunda.html.
[14] M Needleman 2000 The unicode standard Ser. Rev. 26(2) pp. 51–54
[15] M Davis and L Collins 1990 Unicode in 1990 IEEE International Conference on Systems, Man, and Cybernetics: conference proceedings (Los Angeles CA USA: Sheraton Universal Hotel) pp. 499–504.
[16] T U Consortium et al., The Unicode Standard, Version 8.0, vol.8, no. August. 2015.
[17] A Cleeremans, D Serva Schreiber and J L McClelland 1989 Finite State Automata and Simple Recurrent Networks Neural Comput. 1(3) pp. 372–381
[18] M Mukund 1996 Finite-state automata on infinite inputs Tcs 96 p. 2
[19] E Roche and Y Schabes 1998 Finite-state language processing *Computers & Mathematics with Applications* **35** p. 142

[20] T Kameda and P Weiner 1970 On the State Minimization of Nondeterministic Finite Automata *IEEE Trans. Comput.* **C-19**(7) pp. 617–627

[21] F S Irwansyah, Y M Yusuf, I Farida and M A Ramdhani 2018 Augmented Reality (AR) Technology on The Android Operating System in Chemistry Learning *IOP Conference Series: Materials Science and Engineering* **288**(2017) p.012068

[22] S Sari, R Anjani, I Farida and M A Ramdhani 2017 Using Android-Based Educational Game for Learning Colloid Material *Journal of Physics: Conference Series* **895**(1) p. 012012

[23] R Aisyah, I A Zakiyah, I Farida and M A Ramdhani 2017 Learning Crude Oil by Using Scientific Literacy Comics *Journal of Physics: Conference Series* **895**(1) p. 012011

[24] I Farida, I Helsy, I Fitriani and M A Ramdhani 2018 Learning Material of Chemistry in High School Using Multiple Representations *IOP Conference Series: Materials Science and Engineering* **228**(2017) p. 012078

[25] I Helsy, Maryamah, I Farida and M A Ramdhani 2017 Volta- Based Cells Materials Chemical Multiple Representation to Improve Ability of Student Representation *Journal of Physics: Conference Series* **895**(1) p. 012010

[26] F S Irwansyah, I Lubab, M A Ramdhani and I Farida 2017 Designing Interactive Electronic Module in Chemistry Lessons *Journal of Physics: Conference Series* **895**(1) p. 012009

[27] S Sari, D M Aryana, C Z Subarkah and M A Ramdhani 2018 Multimedia Based on Scientific Approach for Periodic System of Element *IOP Conference Series: Materials Science and Engineering* **288**(1) p. 012137