Transliteration Balinese Script using Finite State Automata (FSA) Algorithm

I B T T Murti1,a, C Janis1 and I G A Sudhana1

1Master of Computer Science, Universitas Pendidikan Ganesha, Bali, Indonesia

E-mail : ateguh.teja@gmail.com

Abstract. Currently only few people are able to write with Balinese script. This research has main goal to transform the Latin alphabet in the language of Bali to Balinese script using Finite State Automata (FSA). FSA algorithm is using six syllables that exist in the Balinese language structure and generates by font Bali simbar-b. In this test, the use of the Application has advantages (a) the characters generated can be copied and pasted in another document (b) completeness of Balinese script can be written (c) applications run on Linux OS, Windows, Mac OS. The Application is developed in GTK with C# language. The application was tested in (1) Syllables base (2) Basic word (3) marker vocals word (4) Vocal independent (5) marker semi-vowel word (6) Characters Swaitha (7) Pengange Tengenan (8) other markers (9 ) syllable other (10) numbers (11) punctuation (12) variations in the use of Tedung (13) stands (14) transliteration paragraphs to result 94% accuracy. Some of the drawback in this transliteration using a local database technology are some rule from Balinese Script not applied in algorithms

1. Introduction
Bali Script is the alphabet written on lontar (palm leaves) by the Balinese using Balinese language. Which is coming from ancient Javanese language or Sanskrit. That script is derived from Brahmi script which has a lot commonality in South Asia and Southeast Asia. The use of Balinese script is reduced because it is not used every day. Balinese script is written to palm for family, religious, cultural records that is done at certain time so that only elders can write

Leastwise take of Balinese script raised fears from the threat of extinction that so the next generation will suffer will be the view of science, social and economic life of the Balinese. To avoid this threat then current generations can perform Balinese script related learning using technology approach.

The approach is using today's technology with a smartphone or websites. When using that media, it can be done the advantage of copy-pasting documents so as to facilitate learning by anyone to anyone. The next generation will be eased to know Script Bali and facilitate the use of literacy in daily life.

Resolve the problem above, this paper intends to make the process that makes translate Latin letters into Balinese script easily. The approach used is Finite State Automation. The approach was chosen because it was straightforward to change the characters with the function of chopping words into syllables. Basically, Balinese script represents 1 syllable containing 1-2 letters

2. Literature Review

2.1. Font Bali Simbar-B
There are several fonts as representation for Balinese Script in computerizing eg BSB, Dj Simbar Bali, Bali Simbar Galang and Google Noto Sans. BSB was created by Made Suatjana. Way to use BSB is almost the same as ordinary letters, for example the Latin letter "g" is changed to the "ג" Balinese character. The difficulty is when making Gantungan, Gempelan, and other special characters that require knowing the character unicode

2.2. Unicode
The standard is used to represent letters in BSB. The use of unicode letters displayed can run on Linux OS, Windows, or Mac OS. To see Unicode in BSB, it can be used native application OS Character Map. To view the letter 'g' that has U + 0067, it can be typed in the shortcut Ctrl + Shift + U + 0 + 0 + 6 + 7 after installing fonts BSB.

2.3. Syllable in Balinese Language
The letters in the alphabet Bali represents a syllable. To break a word into parts, then it is used algorithm FSA. There are 6 syllabify in Bali : V, VK (vowel + consonant), KV, KVK, KKV, KKVK [1]. There is a special Balinese alphabet representing the three letters the Latin alphabet that is -nya and -nga.
- V (Vowels): a syllable consisting only of a vowel eg: i-dup (live), a-dep (sell), ra-i (brother), ga-é (work)
- VK (Vowel + Consonant): a syllable consisting of a vowel and consonant eg. ar-ja (art of drama)
- KV (Consonant + Vowel): a syllable pattern consisting of consonants and vowels for example da-kén (shallow), di-ngéh (hear), kar-na (ear)
- KVK (Consonant + Vowel + Consonant): a syllable pattern consisting of consonant, vowel, consonant eg si-nah(appear), sam-pi (cow), an-dus (smoke), je-mak (take)
- KKV (Consonant + Consonant + Vowel): syllable pattern consisting of consonant, consonant, vocal eg pra-gat (done), blu-luk (sugar palm fruit), man-tra (mantra), tum-plu (corner)
- KKVK (Consonant + Consonant + Vowel + Consonant): syllable pattern consisting of consonants, consonants, vowels, consonants example klang-sah (woven coconut leaves), dres-ta (habit), cang-glak (grip), ang-klung (angklung), drum-pak (sat up)

2.4. Finite State Automata
The letters in the alphabet of Bali represents a syllable. To break a word into parts then it is used algoritam FSA. FSA is a mathematical model of computation based on the ideas of a system changing state due to inputs supplied to it and some of these states being acceptor or final states.
2.5. Balinese Script
Script Bali initially only totaled 18 units, namely: ha, na, ca, ra, ka, da, ta, sa, wa, la, ma, ga, ba, nga, pa, ja, ya, and nya [2]. Because a limited number aksaranya while the Balinese language continues to evolve, especially taking from the ancient Javanese language and Sanskrit, the Balinese script later evolved into more numerous.

2.6. Testing Balinese Alphabet
The Balinese Alphabet, which can be viewed on the babadbali.com website created by Ida Bagus Adi Sadewa, displays what characters are in BSB. The test must display the same character as the Balinese script. This test consists of:
- The Basic Syllables (Akśara Wreṣāstra) : There are eighteen basic syllables that are used for writing pure Balinese language.
- The Basic Word : the word that was tested using the Wreṣāstra script containing Gantungan and Gempelan.
- The Vowels (Akśara Suara) : There are twelve distinct vowel sounds in Balinese. They appear as either vowel signs attached to the syllables.
- Independent Vowels : if the word started with a vowel.
- Semi-Vowels (Arda Suara) : The Balinese script has four semi-vowels attached to syllable: ra, wa (ua), la, ya (ia).
- Akśara Šwalalita : In addition to the eighteen basic syllables, there are other syllables that are used for writing Kawi (Old Javanese) language.
- Sound Killers (Pangangge Tengenan) : To end the sound of a syllable, one may add one of the signs: ardha chandra, surang, cecek, bisah, or adegadeq.
- Miscellaneous Signs : The following two signs, if combined with some syllables, become part of the Akśara Modre, which are used to write words in Sanskrit, usually part of prayers.
- Miscellaneous Syllables : There are two more syllables that apparently borrowed from Javanese.
- Numerals : The Balinese use decimal system for numbers. There is a simple one-to-one mapping to the Arabic digits.
- Punctuations : These are some punctuations and symbols specific to the Balinese text.
- Some Variation of Usages : Although the most common used characters are already listed in the previous sections, some manuscripts were using non-standard variants of the characters.
- Abbreviations : Just like in other languages or scripts, long words can be abbreviated as well.
- Word Boundaries and Line Break Rules : In Balinese script, there are no spaces to separate words. In the old time of writing on palm leaves, spaces were scarce and the “page setup” for lontar was always a thin landscape.

3. Methodology
In the methodology, this paper is divided into 3 part pre-translation, transliteration and program testing. Pre-translation explains in detail and gives examples of the use syllables in the FSA. The transliteration process shows the use of the FSA in algorithms and application creation. Testing is the process of accurately verifying algorithms that are used in fit with the testing word

3.1. Pre-translation
Steps of Computation Model of Syllable into FSA are as follows. Existing syllable patterns which has same pattern are made more simple. For examples are V with VK to be (V)K, KV and KVK to be (KV)K, and KKV and KKVK to be (KKV)K so then there are 3 main state. Afterwards each state is connected so that makes FSA (for detail please look figure 2)
Example of FSA algorithm is like word breaking into characters then labeling for vowels and consonant. Once it is determined, it will automatically become its pattern. For example, there is the word 'idep' which is broken into the character i-d-e-p then determined the pattern of V and K becomes
V-K-V-K then from FSA will be determined into VK-VK but for first V(i) does not join with K(d) because the K(d) is not special consonant. So V(i) will stand alone, the first K(d) will join with V(e) however that syllable does not merge with K(p) because K(p) is not special. Outcome word “idep” for FSA is V-KV-K.

3.2. Translation
The translation process of Balinese script is divided into 3 parts, labeling V-K, merging character into syllables with FSA, then changing the syllables to Balinese script with BSB fonts. The initial part is done after inputting the word and breaking the word into character which is then labeled. Labeling is done if the letters a, i, u, e, o become V, numbers 1-9 become A, Punctuation (!, ? : ) becomes T, then the consonant becomes K (please look figure 3).

The labeling of K is quite complex, which is divided into several parts because in Balinese characters there are several Latin characters that can be represented by 1 Balinese characters. For example, the letters ny- and ng- are made into 1 Balinese character. So it becomes K1 and K2. Then in the letters r and h, it becomes special because it is part of Pangange Tenganan so if it meets with a vocal it will make a special character, so that r becomes K3, h becomes K4. Then there are additional syllables for dha, tha, gha, bha being K5 because this syllable represents 1 character of Balinese script. Others for s, p, this pha becomes K6 because this will become an Gempelan so that it gets additional characters.

After getting labeling, the next is combining character with the FSA. This merger is quite easy because the flow has been simplified. For example, the word idep is split into i-d-e-p which has the V-K-V-K label then based on the FSA to become V-KV-K or i-de-p. This merge was suitable for it did not have special K. Arja was split into a-r-j-a which had the V-K3-K-V label so that the merger became VK3-KV or ar-ja.

Start → Input: Word in Latin → Split Word into Character → Label Character with V, K, A, T, → Special Label for K1, K2, K3, K4, K5, K6 → Merge Character into Syllables with FSA → Syllables convert into BSB → Merge Character BSB into word → Output: Word in BSB → Finish

Figure 3. Flowchart FSA in Balinese Script

The further process is to change the syllables to the words in BSB Font. For example is the word "i-de-p" which will be changed from syllable i to i in Bali script. This basic dictionary has been created so that "i" becomes “hi” in Latin, but if converted into BSB it becomes “i” in Balinese script. Then the “de” syllable is readable in Balinese script, the letter “e” is written first and then “d” because “e” turns into a taleng and d becomes letter d so results in being ed. This p syllable itself at the end will be added with adeg-adeg, namely “/” in latin if all syllable combined in Latin to “hied /” but if changed with BSB it is read "idep”. The making of this basic dictionary is based on Wresastra script, Gantungan, Gempelan, Swara script and Tengenan. In addition to preparing a basic dictionary containing basic scripts, a special dictionary is also made which immediately converts certain words into the intended characters for example this OM word has its own holy script so it is entered into a dictionary.
3.3. Testing with Program

Testing is using Ida Bagus Ari Sudewa's Balinese Alphabet. This testing consists of 136 cases where each word is entered into the program and examined whether the word produced in Balinese script is similar to Balinese Alphabet. This test is in the form of basic letters, basic words, numbers, punctuation marks, abbreviations, word Boundaries and line break rules.

Testing in The Basic Syllables, Akśara Śwalalita, The Basic Word, The Vowels, Independent Vowels, Semi-Vowels, Numerals, Punctuation, Abbreviations, Word Boundaries, Line Break Rules and some Sound Killer and some Variation of Usages in this test are passed correctly because having a basic word dictionary and using the FSA. Abbreviation has prepared a dictionary for abbreviations so that it can be passed. For testing the word "karna" is wrong because the word gets “pasang sastra” so it is read “karnna”. The words Mang and Siddham are not appropriate because it is difficult to identify whether the word sacred or not. “samping” is not appropriate because the word in the test uses adeg-adeq in the middle of the word even though this is not allowed. Then the word “ksatria”, “smerti”, “utama”, “dharma” is not suitable because the Balinese script is read as “ksatriya”, “smrerti”, “utrama”, “dharmma”. The existence of certain rules for these words changes from Latin letters to Balinese characters that are not in Balinese Alphabet.

4. Results and discussion

In the previous test there were 8 discrepancies between word of program results and Balinese Alphabet testing. This discrepancy is caused by several rules that are not in the testing. So the percentage results from 136 test cases that match the 128 words produces 94.12%. This result can reach 100% if a dictionary is created for words that do not fit into a special word.

The use of the FSA in the transliteration process gives more value of 90 so that for the application of close to 100 is not impossible. To achieve an approach that is close to it, it can be done by knowing many rules that are outside testing. Knowing the other rules will make the FSA algorithm better. Besides knowing the rules, it can also be done by making a dictionary of special words. This special word dictionary contains, for example, the word dharma which is converted to dharmma. This particular word makes it easier for the conversion process. The adjustment of this algorithm further alleviates changes in the convert process from syllables to BSB.

The testing used have some disadvantages. The first is drawback of assessments for letters and words which is of equal value even though from transliteration the word is more difficult. Then the assessment for a sentence is only 1 piece. Ideally in making testing, the assessment for word is made for all letters plus changes, for example consonants and vowels. Then is the use the word for wresastra script, suara script, Arda Suara, Pengangge Tenganan, etc. Finally is trying to make a sentence or paragraph. The only advantage is that testing is currently available.

5. Conclusion

The FSA algorithm for speaking Latin letters to Balinese script using testing from Ida Bagus Ari Sudewa has a match of up to 94.12%. Although the results of this accuracy are high but there are problems, for example, some rules used in testing are not given a more detailed explanation and some special words that have their own rules. Particular word can be made a kind of dictionary directly translating the word.

References

[1] Purwiati IAM, Negari NPE, Sudiartha IW 2013 Pedoman umum ejaan bahasa bali dengan huruf latin Balai Bahasa Provinsi Bali 26
[2] Tinggen IN 1994 Celah-celah kunci pasang aksara bali Rhika Dewata Singaraja 4
[3] Simpen W 1995 Pasang aksara bali Upada Sastra 7